EDV in Medizin und Biologie 16 (2), 41–46, ISSN 0300-8282

Datenbankaspekte bei statistischen Auswertungssystemen*)

R. Haux

Zusammenfassung

In dieser Arbeit soll verdeutlicht werden, daß die Datenverwaltung und -aufbereitung zu den wesentlichen Aufgaben eines statistischen Auswertungssystems gehört. Damit das statistische Auswertungssystem den Statistiker bei seiner Arbeit ausreichend unterstützen kann, benötigt es ein geeignetes Datenstruktur- und Datentypkonzept. Es soll außerdem gezeigt werden, daß es günstiger ist, die Datenverwaltungskomponente in einem statistischen Auswertungssystem integriert zu haben. Als letzten Aspekt möchte ich die Datenverwaltung und -aufbereitung durch statistische Expertensysteme behandeln. Es sollen die Möglichkeiten aufgezeigt werden, aber auch die Gefahren von Fehlanwendungen, die besonders bei mangelnder Kenntnis der statistischen Methodik des Expertensystemkonstrukteurs entstehen können.

Summary

The paper tries to point out the necessity of database management facilities within statistical analysis systems. In order to support a statistician efficiently, such a system should have appropriate data structure types and data types. Apart from these two points of view it is argued that database management has to be integrated in statistical analysis systems and that difficulties arise when database management is done within a database system and statistical data analysis is done separately in a statistical analysis system. The last part of the paper deals with database management of socalled statistical expert systems. Advantages of such expert systems for statistical data analysis are pointed out. However, the risks to abuse such systems are also demonstrated. This abuse is especially possible, if designers of statistical expert systems are not sufficiently familiar with statistical methodology.

1. Einleitung

DITTRICH et al. (1984) schreiben in ihrer Arbeit „Datenbankkonzepte für Ingenieuranwendungen":

„Datenbanksysteme stellen heute weitgehend ausgereifte Werkzeuge der Informatik zur Verwaltung größerer Datenbestände in vielerlei Anwendungen dar. ... Traditionell werden Datenbanksysteme im administrativ-betriebswirtschaftlichen Bereich eingesetzt. ... Mehr und mehr besteht jedoch auch seitens ingenieurwissenschaftlicher Anwender der Wunsch, sich die Vorteile von Datenbanksystemen zunutze zu machen." Sie schreiben weiter: „Der Versuch, auf dem Markt verfügbare Datenbanksysteme ..." (bei ingenieurwissenschaftlichen Anwendungen) „... einzusetzen, führte schnell zu einer Reihe von Problemen. Der Grund liegt darin, daß existierende Datenbanksysteme auf den traditionellen Einsatzbereich zugeschnitten sind und Ingenieuranwendungen zumindest für die Charakteristika ... Datenstrukturierung, Konsistenz ..." (gemeint ist hier die möglichst vollständige und widerspruchsfreie Repräsentation des interessierenden Wirklichkeitsausschnittes) „... teilweise drastisch andere Anforderungen stellen."

Ähnliches trifft für rechnergestützte statistische Auswertungen zu. Dies möchte ich im Folgenden versuchen zu belegen. Meine Thesen sind:

1. Eine adäquate Datenbankverwaltungskomponente ist für statistische Auswertungen größeren Umfangs wichtig.
2. Wir benötigen Konzepte, die der Statistik angepaßt sind, insbesondere geeignete Datenstruktur- und Datentypen.
3. Eine Datenbankverwaltungskomponente sollte in einem statistischen Auswertungssystem integriert sein.
4. Bei einer Datenbankverwaltungskomponente in einem Statistischen Expertensystem müssen wir ebenfalls in besonderem Maße auf die Anforderungen achten, die sich durch die statistische Methodik ergeben.

Die Thesen werden in den Abschnitten 3 bis 6 besprochen. Die Diskussion der Thesen kann hier nur in groben Zügen erfolgen; detailliertere Betrachtungen befinden sich in den angegebenen Literaturstellen. Die Ausführungen enthalten außerdem Überlegungen, die nicht notwendigerweise für derzeitige Anwendungen sinnvoll sind und die als Vorschläge für zukünftige Entwicklungen angesehen werden sollen.

2. Grundbegriffe

Zunächst müssen wir kurz beschreiben, was wir unter einem statistischen Auswertungssystem verstehen wollen. Eine ausführliche Definition befindet sich in HAUX (1983/84).

Ein statistisches Auswertungssystem besteht aus einer Datenbank, welche die auszuwertenden Daten enthält, und

*) Vortrag, gehalten auf dem 31. Biometrischen Kolloquium der Deutschen Region der Internationalen Biometrischen Gesellschaft, 12. bis 15. 3. 1985, Bad Nauheim.

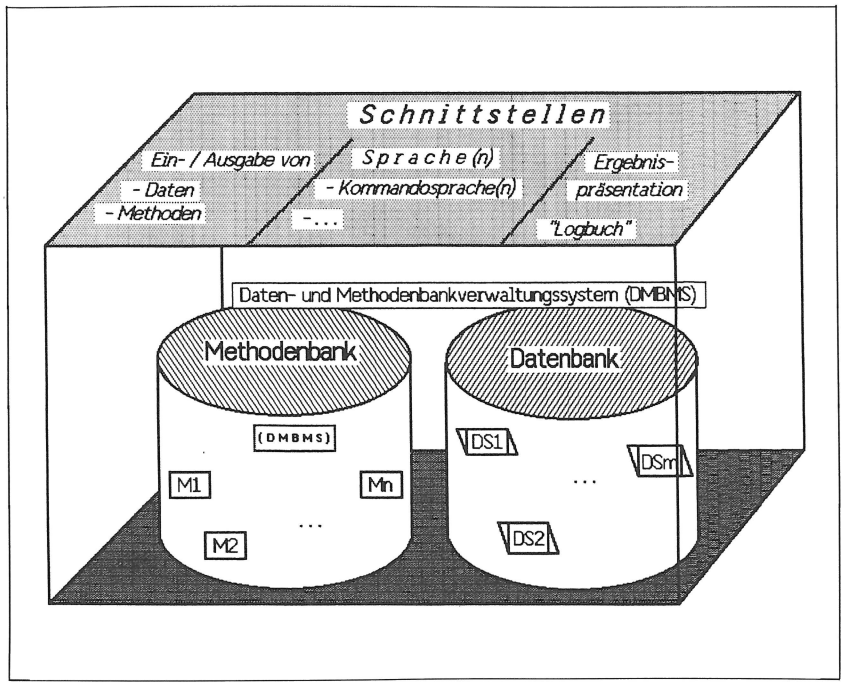


Abb. 1. Aufbau eines Statistischen Auswertungssystems (DS: Datenstruktur, M: Methode).

Studienplanung	– Entwurf von Datenstrukturen; Datenbankentwurf; Entwurf bzw. Auswahl von Programmen.
Studiendurchführung	– Datenerfassung; Integritätskontrollen.
Studienauswertung	– Aufruf der Programme und Übergabe der gespeicherten Daten an die Programme; Ausgabe der Ergebnisse; Ergebnispräsentation (Graphiken, Tabellen)

Abb. 2. Aufgaben bei den rechnergestützten statistischen Auswertungen in der Studienplanung, -durchführung und -auswertung.

aus einer Methodenbank (Abb. 1). In der Methodenbank stehen im wesentlichen Methoden (implementiert als Programme) zur statistischen Auswertung von Datenbeständen aus der Datenbank zur Verfügung. Die Datenbank und die Methodenbank werden verwaltet durch ein Daten- und Methodenbankverwaltungssystem.

Das Daten- und Methodenbankverwaltungssystem – das man als Programm bzw. als Menge von Programmen implementiert – kann Methoden aus der Methodenbank aktivieren und auf Datenstrukturen in der Datenbank zugreifen. Für einen Anwender sind die Schnittstellen zum statistischen Auswertungssystem wichtig. Eine dieser Schnittstellen ist die Kommandosprache, über die der Benutzer seine Datenbestände beschreibt und die Programme zur Auswertung der Datenbestände aktiviert.

Unter einer Datenbankkomponente wollen wir den Teil des Daten- und Methodenbankverwaltungssystems verstehen, der sich hauptsächlich mit der Verwaltung der Datenbank befaßt (also nicht mit der Verwaltung der Methodenbank). Diese Datenbankverwaltungskomponente gibt es auch als Datenbankverwaltungssystem in Datenbanksystemen.

3. Datenbankverwaltung

Rechnerunterstützung ist in vielen Fällen hilfreich und notwendig für den Statistiker. Bekannte Werkzeuge hierfür sind Datenbanksysteme, z. B. SIR, und statistische Auswertungssysteme, z. B. BMDP, SAS, SPSS (Literaturangaben zu diesen Systemen befinden sich in FRANCIS, 1981). Besonders bei Studien mit großen Stichprobenumfängen oder mit rechenintensiven Auswertungsmethoden erscheinen sie mittlerweile unerläßlich. Diese Systeme werden aber nicht nur zur statistischen Auswertung einer Studie verwendet. Sie kommen bereits bei der Studienplanung und -durchführung zum Einsatz. Die jeweiligen Aufgaben bei den rechnergestützten statistischen Auswertungen für die einzelnen Studienphasen sind in Abb. 2 dargestellt.

Bereits bei der Studienplanung haben wir es mit Datenbankanwendungen zu tun. Wir müssen uns dort Gedanken machen, wie wir die Daten strukturieren wollen, damit bei der späteren Studiendurchführungsphase die Datenerhebung, -erfassung und -speicherung möglichst einfach und möglichst fehlerfrei vonstatten gehen kann und damit wir Zwischen- und Endauswertungen ohne allzu großen Aufwand durchführen können. Wie ein solcher Datenbankentwurf, besonders bei klinischen Studien, aussehen kann (und aussehen sollte!), ist in LEIBBRAND (1984) und HOLLE und LEIBBRAND (1985) beschrieben. Sie zeigen die einfache und wirkungsvolle Anwendbarkeit des relationalen Datenmodells (CODD, 1970, 1979) für den Daten(bank)entwurf. HOLLE und LEIBBRAND weisen aber auch darauf hin, daß diese Entwurfstechniken bereits bei der Datenerhebung, beispielsweise für den Entwurf von Erhebungsbögen, anwendbar sind.

Ein solcher Datenbankentwurf, der dem Statistiker die Arbeit bei der Studiendurchführung und -auswertung erleichtern soll, läßt sich nur dann sinnvoll ausführen, wenn die uns zur Verfügung stehenden Werkzeuge eine Datenbankverwaltungskomponente enthalten, die eine Implementation dieses Entwurfes ermöglichen. Für die Studiendurchführung ist es auch wichtig, neben der bereits erwähnten Datenstrukturierung schon bei Beginn der Datenerfassungsphase Plausibilitätsbedingungen (semantische Integritätsbedingungen) zu for-

mulieren, um so rechtzeitig wie möglich auf Fehler, die während der Datenerfassungsphase auftreten, aufmerksam machen zu können. So läßt sich etwa der Anteil an fehlenden Werten (verursacht durch nicht durchgeführte Untersuchungen oder nicht im Erhebungsbogen eingetragene Untersuchungsergebnisse) gerade bei größeren, multizentrischen Studien in Grenzen halten, und so können wir versuchen, Verzerrungen bei der statistischen Auswertung der Daten zu vermeiden.

Eine Datenbankverwaltungskomponente in einem statistischen Auswertungssystem oder in einem Datenbanksystem, welche die oben erwähnten Aufgaben unterstützt, ist also wichtig: zum einen, um eine ausreichende Datenqualität zu gewährleisten, und zum anderen, um dem Statistiker die Arbeit zu erleichtern. Ein mangelhafter Datenbankentwurf kann, besonders bei großen Studien, dazu führen, daß nach der Datenerhebung erhebliche Inkonsistenzen in den Datenbeständen bestehen und daß der hohe Aufwand der Datenaufbereitung, vor allem für explorative und für deskriptive Auswertungen, die Gesamtstudienauswertung stark beeinträchtigt. Übrigens gilt bei dem Datenbankentwurf dieselbe Regel wie bei der Versuchsplanung (vgl. IMMICH, 1974, S. 1): Fehler kann (im allg.) nachträglich niemand mehr ausgleichen!

4. Datenstruktur- und Datentypen

Die derzeit vorhandenen statistischen Auswertungssysteme haben auf der Ebene der Kommandosprache fast alle nur den Datenstrukturtyp Datenmatrix (d. h. eine sequentielle Datei von gleichartig aufgebauten Verbundstrukturen (sog. „records"), vgl. Abb. 3) und den Datentyp „reelle Zahl", evtl. noch Zeichenkette, zur Verfügung. Mit diesen Datenstruktur- und Datentypen läßt sich der den Statistiker interessierende „Teil der Wirklichkeit" oft nur mangelhaft in die Datenbank eines statistischen Auswertungssystems oder eines Datenbanksystems abbilden. Somit ist die für einen Anwender wesentliche Konsistenzbedingung von DITTRICH et al. (1984) nicht erfüllt. Wir wollen dies anhand eines Beispiels betrachten. Eine ausführliche Darstellung des Problems findet man in HAUX (1984).

In der prospektiven, multizentrischen Langzeitstudie „Akute Virushepatitis" (KABOTH et al., 1980) wurden unter anderem die Merkmale

1. Patientenidentifikation,
2. Geschlecht,
3. HBsAg und
4. anti-HBs

erhoben. Die Patientenidentifikation soll den Patienten eindeutig identifizieren. HBsAg (Hepatitis-B-Surface-Antigen, ein im Serum nachweisbares Oberflächenpartikel des Hepatitis-B-Antigens) wird semiquantitativ gemessen: negativ, positiv aber nicht quantifizierbar (unter 0,5 Milligramm/Liter), als quantitativer Wert (größer oder gleich 0,5 Milligramm/Liter). Anti-HBs, der Antikörper gegen das HBsAg sei negativ, schwach positiv oder positiv erhoben.

Müssen wir solche Daten in einer Datenmatrix speichern, dann haben wir keine konsistente Beschreibung des Datenbestandes mehr gegeben.

Ein statistisches Auswertungssystem, das nur numerische Kodierungen, allenfalls noch Zeichenketten zuläßt, kann nicht mehr die semiquantitative Eigenschaft der HBsAg-Messung erkennen und berücksichtigen; es erkennt außerdem nicht das nominale bzw. ordinale Skalenniveau von Geschlecht bzw. anti-HBs. Die Eindeutigkeitseigenschaft der Patientenidentifikation ist in einem Datenstrukturtyp Datenmatrix ebenfalls nicht repräsentiert.

Identifikation	Behandlung	Kriterium
1	1	120
2	1	135
3	1	118
4	1	122
5	2	180
6	2	145
–	–	–
–	–	–
–	–	–

Abb. 3. Beispiel für eine Datenstruktur vom Typ Datenmatrix.

Sämtliche der hier aufgeführten Eigenschaften werden selbst in den zu Beginn erwähnten, bekannten Datenbank- und statistischen Auswertungssystemen nicht berücksichtigt. Bis auf SIR ist auch keines dieser Systeme in der Lage, semantische Integritätsbedingungen zur Datenstrukturspezifikation aufzunehmen. Beispielsweise ließe sich hier die Bedingung angeben: „Falls HBsAg positiv und anti-HBs positiv, dann soll das System eine Warnung ausgeben und den Datensatz kennzeichnen." Systeme, welche die in der Statistik vorkommenden Daten genauer, konsistenter speichern könnten, wären für den Statistiker ein weitaus geeigneteres, nützlicheres Werkzeug.

Statistische Auswertungssysteme und Datenbanksysteme sollten den Statistiker (oder den Fachwissenschaftler mit Statistikkenntnissen) bei der Erfassung, Aufbereitung und Auswertung von Daten unterstützen, indem sie

1. die genannten Aufgaben mit möglichst wenig Aufwand seitens des Statistikers sorgfältig durchführen und gleichzeitig – damit eine qualitativ hochwertige Auswertung gewährleistet ist –
2. die Tätigkeiten – zum Nutzen des Statistikers – konstruktiv kontrollieren.

Dazu benötigen wir eine konsistente Speicherung des für die statistische Auswertung relevanten Wirklichkeitsausschnittes. Um dies wenigstens annähernd zu erreichen, benötigen wir:

1. Datentypen, die – für statistische Auswertungen – ausreichend genaue Angaben über die Beschaffenheit der Daten enthalten, und
2. Datenstrukturtypen, bei denen sich die Beziehungen zwischen Variablen ebenfalls ausreichend genau spezifizieren lassen. Hierzu gehört auch die Angabe semantischer Integritätsbedingungen.

Beispielsweise wäre für Geschlecht der Datentyp „(ungeordneter) Aufzählungstyp", für anti-HBs der Datentyp „geordneter Aufzählungstyp" geeignet. HBsAg könnte man als Variable vom Typ „semiquantitativ" angeben. In einer Datenstruktur vom Typ Relation ließe sich die Eindeutigkeitseigenschaft der Variablen Patientenidentifikation auf einfache Weise spezifizieren. Systeme mit solchen Datenstruktur- und Datentypen könnten auch einen fehlerhaften Methodenaufruf – beispielsweise eines t-Test-Programmes, das als Kriterium die Variable HBsAg zugeordnet bekommt – erkennen und melden (HAUX, 1983; NELDER, 1977).

Ein neueres statistisches Auswertungssystem, welches zumindest einige der oben erwähnten Punkte berücksichtigt, ist das System S (BECKER und CHAMBERS, 1984). Im Bereich

der Informatik versucht man ebenfalls die erwähnten Mängel zu beseitigen; verwiesen sei hier etwa auf das System IDAMS (ERBE et al., 1980) und auf die Sprache PASCAL/R (SCHMIDT, 1977). Auch gibt es bereits existierende Systeme im kommerziellen Bereich, die eine bessere Datenhaltung ermöglichen als die in der Statistik eingesetzten Systeme. Aber ähnlich wie bei den ingenieurwissenschaftlichen Anwendungen werden auch dort die spezifisch statistischen Notwendigkeiten nicht berücksichtigt!

5. Datenhaltung in einem Datenbanksystem oder in einem statistischen Auswertungssystem?

Bisher haben wir eine Frage, die sich bei der Datenhaltung für statistische Auswertungen erhebt, noch nicht beantwortet: Soll man Daten in der Datenbank eines Datenbanksystems oder in der Datenbank eines statistischen Auswertungssystems speichern? Die Frage läßt sich nicht eindeutig beantworten und wird auch in der Literatur sehr unterschiedlich angegangen (vgl. z. B. RASSMANN, 1985; VICTOR, 1984 S. 113 und HAUX, 1983/84 S. 112). Wir können allerdings feststellen, daß die Datenverwaltung bei den meisten der derzeitigen statistischen Auswertungssystemen nur sehr mühselig (wenn überhaupt!) durchzuführen ist. So ist es bei größeren Studien zur Zeit oft nicht vermeidbar, den nicht integrierten Ansatz zu wählen (d. h. neben einem statistischen Auswertungssystem noch zusätzlich ein Datenbanksystem zu verwenden).

Es sprechen allerdings einige grundsätzliche Argumente für ein integriertes System, also für ein statistisches Auswertungssystem mit geeigneter Datenbankverwaltungskomponente:

1. Der Statistiker braucht nur eine Sprache (ein System) zu erlernen.
2. Wir müssen bei nicht integrierten Systemen bei der Auslagerung der auszuwertenden Daten aus der Datenbank des Datenbanksystems in die Datenbank des statistischen Auswertungssystems mit einem Informationsverlust rechnen. Dies besonders, wenn das statistische Auswertungssystem dem Statistiker nur Datenmatrizen mit ,,reellen" Zahlen zur Verfügung stellt.
3. Datenhaltung und Datenauswertung sind keine zeitlich getrennten, unabhängigen Vorgänge. Während der Datenerfassungsphase benötigt man bereits Verfahren der beschreibenden Statistik, um die Datenbestände zu überprüfen (Untersuchung auf Extremwerte, auf Mischverteilungen, ...) und um damit die Datenqualität zu erhöhen. Bei der Auswertung möchte man andererseits gerne Zwischenergebnisse in die Datenbank ablegen.

Die Integration einer ausreichenden Datenverwaltungskomponente in ein statistisches Auswertungssystem ermöglicht auch andere, der statistischen Methodik evtl. angepaßtere Programmentwicklungsverfahren (HAUX, 1983/84, S. 14–20). Nicht vergessen werden sollte auch WIRTHS Aussage:

,,Entscheidungen über die Strukturierung der Daten ..." (können) ,,... nicht ohne Kenntnis der auf die Daten anzuwendenden Algorithmen getroffen werden ... umgekehrt ..." (kann) ,,... die Struktur und Wahl der Algorithmen oft stark von der Struktur der zugrunde liegenden Daten ..." (abhängen). ,,... Kurz gesagt: Programmerstellung und Datenstrukturierung sind untrennbar ineinandergreifende Themen." (WIRTH, 1979, S. 7).

Dieses Ineinandergreifen unterbindet man, wenn man Datenstrukturen grundsätzlich vor Anwendung der statistischen Algorithmen in eine Datenmatrix umwandeln muß!

6. Statistische Expertensysteme

Expertensysteme sollen die Tätigkeiten von Experten unterstützen bzw. ganz oder teilweise übernehmen. Sie werden besonders dadurch charakterisiert, daß sie die Speicherung von Expertenwissen in einer Wissensbank ermöglichen (vgl. z. B. BUCHANAN und SHORTLIFFE, 1984, und CLANCEY und SHORTLIFFE, 1984).

Als statistische Expertensysteme wollen wir diejenigen Expertensysteme bezeichnen, die die Tätigkeiten des Statistikers unterstützen bzw. ganz oder teilweise übernehmen. Solche Systeme kann man als Erweiterungen von statistischen Auswertungssystemen ansehen. Sie enthalten nicht nur die (Auswertungs-)Programme in der Methodenbank, sondern auch sogenanntes heuristisches Expertenwissen in einer Wissensbank. Heuristisches Wissen bedeutet hier: Wissen, das man durch Erfahrung auf einem Anwendungsgebiet erlangt.

Statistische Expertensysteme sind für verschiedene Anwendungen denkbar, beispielsweise

1. als Beratungssystem zur automatischen Methodenauswahl im Bereich der explorativen Datenanalyse (HAJEK und IVANEK, 1982),
2. als Beratungssystem zur Entdeckung fehlerhafter Anwendungen und zur automatischen Methodenauswahl für bestimmte parametrische und nichtparametrische Methoden durch die Einbeziehung von struktureller Information (WITTKOWSKI, 1984);
3. als System zur Überwachung, Interpretation und Anleitung für Auswertungen mit Regressionsverfahren, u. a. durch Modellüberprüfung und Modellanpassung (PREGIBON und GALE, 1984);
4. als System, das Wissen, welches außerhalb der Erhebung oder des Experimentes vorliegt und das für die Fragestellung von Bedeutung ist, miteinbezieht (VICTOR, 1984, S. 115) und
5. als System zur automatischen, statistischen Auswertung von großen (klinischen) Datenbanken (BLUM, 1982).

Die absehbaren Tendenzen hin zur Entwicklung statistischer Expertensysteme wurden als Fortschritt im positiven Sinne (CHAMBERS, 1983) wie auch im negativen (ZELEN, 1984) angesehen. ZELEN beschreibt die Gefahren von Systemen zur automatischen Datenanalyse folgendermaßen:

,,The user will input a set of stylized questions dealing with various hypotheses or models. The system will choose one or more appropriate data analysis techniques and give the answer ... I do not welcome this future, as I believe it will stifle individual innovations on particular problems."

Da wir in dieser Arbeit Datenbankaspekte besprechen, müssen wir uns vor allem mit Punkt 5 befassen. Vorher soll aber noch folgendes festgehalten werden: Wollen wir eine Verbesserung des Werkzeuges Statistisches Auswertungssystem, dann dürfen wir uns dieser Entwicklung nicht grundsätzlich verschließen.

Besonders im Bereich des Meldens von Fehlern bzw. Warnungen bei Modellverletzungen arbeiten die existierenden statistischen Auswertungssysteme nur mangelhaft. Eine Verbesserung dieser Situation durch statistische Expertensysteme erscheint hier begrüßenswert. Wir sollten dies auch unter dem Aspekt sehen, daß viele Anwender mit relativ geringen Statistikkenntnissen Systeme wie SPSS anwenden, ohne einen Statistiker zu konsultieren. Hier könnten statistische Expertensysteme den Statistiker zwar nicht ersetzen, aber sie könnten eine, wenn auch rudimentäre statistische Beratung ermöglichen. Diese ,,Beratung" kann nur sinnvoll sein, wenn die statistische Methodologie – beispielsweise Grundsätze der

Versuchsplanung –, also das „statistische Expertenwissen" tatsächlich berücksichtigt wird. Solche Systeme werden jedoch eine negative Entwicklung einleiten, wenn sie von einem Anwender als Lehrbuchersatz angesehen werden bzw. als Systeme, mit denen man ohne große Mühe retrospektiv Datenbestände mit Verfahren der schließenden Statistik auswerten kann, ohne sich die Bedenken eines Statistikers anhören zu müssen. Um die Entwicklung statistischer Expertensysteme positiv zu beeinflussen, sind wir als Statistiker gefordert, da im Bereich der Informatik solches Statistik-Expertenwissen nicht in dem Maße zur Verfügung stehen kann.

Auch eine (teil-)automatisierte Methodenauswahl, besonders für Verfahren der schließenden Statistik, müssen wir vorsichtig und kritisch angehen. Wir dürfen nicht vergessen, daß vor der Auswertung zunächst ein Problem steht (bzw. sollte!), das man mit statistischen Verfahren lösen möchte. Daraufhin präzisiert man die Fragestellung, entwirft das statistische Modell (des Experimentes, ...) und formuliert das Test- oder Schätzproblem bzw. die Test- oder Schätzprobleme. Erst danach haben wir es mit Daten zu tun. Selbst wenn wir die in Abschnitt 5 geforderten Datenstruktur- und Datentypen zur Verfügung haben, läßt sich aus der strukturellen Information nicht mehr rückwirkend der Versuchsplan erstellen. Folglich lassen sich auch keine statistischen Modelle mit den dazugehörigen Schätz- bzw. Testproblemen mehr angeben. Abgesehen davon bleibt es immer noch schwierig, aufgrund der Beschaffenheit der Daten Aussagen, etwa über geeignete Hypothesen, zu machen. Nehmen wir noch einmal das HBsAg als Kriterium für einen Zweistichprobentest. Man kann relativ einfach zu dem Schluß kommen, daß ein linearer Rangtest, wie etwa der Wilcoxon-Mann-Whitney-Test (in verallgemeinerter Form beschrieben in HAJEK und SIDAK, 1967, S. 85 ff.) dem t-Test vorzuziehen ist. Aber ob man nun Ränge (HAJEK und SIDAK, 1967, S. 87) oder Exponential-Scores (HAJEK und SIDAK, 1967, S. 97) wählt, was zwei verschiedenen Testproblemen entspricht, läßt sich schon nicht mehr so leicht entscheiden.

Datenbankaspekte werden vor allem in Punkt 5 angesprochen. Der Ansatz von BLUM, der hier in knapper und etwas provokativer Form geschildert werden soll, tangiert aber auch die zuvor erwähnten Probleme. BLUM untersuchte die Möglichkeiten, einen Datenbestand von Rheumakranken mit Hilfe von automatisierten Verfahren statistisch zu analysieren. Er entwickelte einen „Statistikroboter" („robot statistician", S. 419), der bei gegebener Fragestellung u. a. automatisch ein statistisches Modell entwerfen („create the statistical model", S. 414), statistische Methoden auswählen („select statistical methods", S. 414) und die Datenbestände auswerten soll („interpret the results to determine significance", S. 414). Obwohl BLUM in den meisten Fällen von einer explorativen Analyse der Daten spricht, schreibt er aber auch auf S. 421 „Naturally, we are interested in knowing whether a given causal relationship is statistically significant ..." und bringt auch unter dem Abschnitt „medical results" auf Seite 423 eine Tabelle mit den 15 niedrigsten p-Werten, die bei einer Auswertung ermittelt wurden. Dort steht nicht, ob diese p-Werte „explorativer Natur" sind. BLUM erwähnt zwar, daß retrospektive Auswertungen problematisch sein können, er geht jedoch fast nicht auf die Möglichkeiten von Verzerrungen durch Selektionseffekte (vgl. z. B. JESDINSKY, 1977, oder RÜMKE, 1970) und gar nicht auf die Problematik des multiplen Testens (vgl. z. B. SONNEMANN, 1982) ein.

Obwohl dieses Beispiel bedenklich stimmt, dürfen wir dabei nicht vergessen, daß das sorgfältige Auswerten von Datenbeständen – auch von Beständen, die nicht prospektiv erhoben wurden – wichtig und notwendig ist. Gerade bei der explorativen Auswertung großer Studien kann ein Statistisches Expertensystem auch hier dem Statistiker eine wertvolle Unterstützung sein und ihm die Arbeit erleichtern. Aber wir müssen darauf achten, daß man die Anforderungen, die sich aus der statistischen Methodik ergeben, nicht vernachlässigt!

7. Schlußbemerkungen

Wir haben uns hier mit einigen Datenbankaspekten befaßt, die für statistische Auswertungen von Bedeutung sind. Für diese Aspekte war es wichtig, Methoden aus der Informatik und aus der Statistik zu berücksichtigen. Es bleibt zu hoffen, daß Datenbankanwendungen weiter in die Statistik integriert werden. Richtig angewandt, also auch unter Berücksichtigung spezifisch statistischer Erfordernisse, können sie eine wertvolle Hilfe für den Statistiker und für den Statistik-Anwender sein.

Literatur

BECKER und CHAMBERS (1984): Design of the S system for data analysis. Comm. ACM, **27,** 486–495.

BLUM, R. L. (1982): Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical data base: the RX project. Comp. and Biom. Research, **15,** 164–187, ebenfalls enthalten (und von dort zitiert) in CLANCEY und SHORTLIFFE (1984), 329–425.

BUCHANAN, B. G. und E. H. SHORTLIFFE (Hrsg.) (1984): Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Reading, Ma.: Addison-Wesley.

CHAMBERS, J. M. (1983): The new future of data analysis. Proc. 44th session of the ISI, Buch 1, 97–103, Madrid.

CLANCEY, W. J. und E. H. SHORTLIFFE (Hrsg.) (1984): Readings in medical artificial intelligence – the first decade. Reading, Ma.: Addison-Wesley.

CODD, E. F. (1970): A relational model for large shared data banks. Comm. ACM, **13,** 377–387.

CODD, E. F. (1979): Extending the database relational model to capture more meaning. ACM Trans. Database Syst., **4,** 397–434.

DITTRICH, K., A. KOTZ, J. MÜLLE und P. LOCKEMANN (1984): Datenbankkonzepte für Ingenieuranwendungen: eine Übersicht über den Stand der Entwicklung. EHRICH, H.-D. (Hrsg.): GI – 14. Jahrestagung, 175–192. Berlin: Springer.

ERBE, R., R. HARTWIG, H. LEHMANN, G. MÜLLER und U. SCHAUER (1980): Integrated data analysis and management for the problem solving environment. Inform. Systems, **5,** 273–285.

FRANCIS, I. (1981): Statistical software: a comparative review. New York: North Holland.

HAJEK, J. und Z. SIDAK (1967): Theory of rank tests. New York: Academic Press.

HAJEK, P. und J. IVANEK (1982): Artificial intelligence and data analysis. CAUSSINUS, H., P. ETTINGER und R. TOMASSONE (Hrsg.): COMPSTAT 1982, 54–60. Wien: Physika.

HAUX, R. (1983): How to detect and prevent errors in computer-supported statistical analysis: an example. Meth. Inform. Med., **22,** 87–92.

HAUX, R. (1983/84): Statistical analysis system – construction and aspects of method design. Stat. Software Newsl., **9,** 106–115 und 10, 14–27.

HAUX, R. (1984): The construction of statistical analysis systems on the basis of appropriate data types and data structure types. Manuskript, zur Veröffentlichung vorgesehen.

HOLLE, R. und D. LEIBBRAND (1985): Data design in clinical trials. Stat. Software Newsl., **11,** 9–14.

IMMICH, H. (1974): Medizinische Statistik. Stuttgart: Schattauer.

JESDINSKY, H. J. (1977): Statistische Auswertung großer Datenmengen – nur ein technisches Problem? Stat. Software Newsl., **3,** 68–75.

KABOTH et al. (DFG-Studiengruppe „Akute Virushepatitis") (1980): Kooperative prospektive Studie „Akute Virushepatitis". Verh. Dt. Ges. f. Innere Med., **86,** 749–756.

LEIBBRAND, D. (1984): Datenhaltung und Erfassung bei klinischen Studien. EDV in Med. u. Biol., **15,** 33–40.

NELDER, J. A. (1977): Intelligent programs, the next stage in statistical computing. BARRA, R. et al. (Hrsg.): Recent developments in statistics, 79–86. Amsterdam: North Holland.

PREGIBON, D. und W. A. GALE (1984): REX: an expert system for regression analysis. HAVRANEK, T., Z. SIDAK und M. NOVAK (Hrsg.): COMPSTAT 1984, 242–248. Wien: Physika.

RASSMANN, B. (1985): The use of database systems for clinical trials. Erscheint im Stat. Software Newsl., **11,** Heft 2.

RÜMKE, CH. L. (1970): Über die Gefahr falscher Schlußfolgerungen aus Krankenblattdaten (Berkson's Fallacy). Meth. Inform. Med., **9,** 249–254.

SCHMIDT, J. W. (1977): Some high level language constructs for data of type relation. ACM Trans. Database Syst., **2,** 247–261.

SONNEMANN, E. (1982): Allgemeine Lösungen multipler Testprobleme. EDV in Med. u. Biol., **13,** 120–128.

VICTOR, N. (1984): Computational statistics – tool or science? Stat. Software Newsl., **10,** 105–116.

WIRTH, N. (1979): Algorithmen und Datenstrukturen, zweite Auflage. Stuttgart: Teubner.

WITTKOWSKI, K. M. (1984): On the use of structural information for a statistical expert system in medical research. V. EIMEREN, W., R. ENGELBRECHT und CH. D. FLAGLE (Hrsg.): Third Int. Conf. on System Science in Health Care, 1140–1143. Berlin: Springer.

ZELEN, M. (1983): Biostatistical science: a look into the future. Biometrics, **39,** 827–830.

Eingegangen am 26. April 1985

Anschrift des Verfassers: Dr. Reinhold Haux, Institut für Medizinische Statistik und Dokumentation der RWTH Aachen, Pauwelsstraße, D-5100 Aachen.

EDV in Medizin und Biologie **16** (2), 47–48, ISSN 0300-8282

Statistical test of three-factor associations in the HLA system

J. Töwe, R. Wegener, G. Kundt and M. Kracht

Summary

Three-factor associations in the HLA system are frequently tested statistically by well-known procedures using $2 \times 2 \times 2$ contingency table analysis. The coupling index D_{ABC} given in many publications is suitable for none of these procedures. A more meaningful statistical test for the index D_{ABC} is described and illustrated by examples.

Zusammenfassung

Dreifaktorenassoziationen im HLA-System werden häufig mit bekannten Methoden aus der $2 \times 2 \times 2$-Kontingenztafelanalyse statistisch getestet. Der in vielen Publikationen angegebenen Kopplungsgröße D_{ABC} wird keine dieser Methoden gerecht. Ein adäquater statistischer Test für die Größe D_{ABC} wird angegeben und an Beispielen erläutert.

The extent of gene couplings must be taken into account during studies of the distribution of HLA characters in population samples (MATTIUZ et al. 1970).

The statistical tests for such coupling effects in cases of three-factor association are well-known methods for testing 2nd order interaction in $2 \times 2 \times 2$ contingency tables (BARTLETT 1935; PIAZZA 1975; HILL 1975). This procedure was also used for the population analyses performed at the VIII International HLA-Workshop (BAUR and DANILOVS 1980a); this contribution submits it to critical examination. Alleles must be regarded as factors. Let us consider the factors A, B and C. The corresponding probabilities are defined as follows:

$$P(A) = P_{1..};\ P(\overline{A}) = P_{2..} = 1 - P_{1..};\ P(B) = P_{.1.};$$
$$P(ABC) = P_{111};\ \ldots;\ P(\overline{ABC}) = P_{222} \qquad (1)$$

The above-mentioned authors give the following measure for a three-factor association:

$$D_{111} = P_{111} - P_{1..}P_{.1.}P_{..1} - P_{1..}D_{*11} - P_{.1.}D_{1*1} - P_{..1}D_{11*} \qquad (2)$$

where

$$D_{11*} = P_{11.} - P_{1..}P_{.1.};\ D_{1*1} = P_{1.1} - P_{1..}P_{..1};$$
$$D_{*11} = P_{.11} - P_{.1.}P_{..1}$$

The quantities D_{11*}, D_{1*1} and D_{*11} can be interpreted as measures of a corresponding two-factor association.

They are tested for deviation from zero by means of the four-field-χ^2 test.

The estimated quantity $\hat{D}_{111}$ is statistically tested as 2nd order interaction effect in the sense of the $2 \times 2 \times 2$-contingency table (Model I) (BARTLETT 1935; PIAZZA 1975; HILL 1975). The models used by above-mentioned authors are not fully consistent with equation (2). HILL (1975, 1976) uses the loglinear model, but this gives rise to contradictions. The HLA-A,B,C haplotype reference tables (BAUR and DANILOVS 1980b) contain examples for which the statistical test can yield positive results in spite of the absence of couplings ($D_{111} = 0$). On the other hand, haplotypes with high D-values but only small χ^2 values are included in the frequency tables. Two different concepts clearly clash with each other here.

TÖWE et al. (1982) have applied the concept defined by (2) to contingency tables. A definition of interaction given by LANCASTER (1969, S. 254) is totally consistent with this concept.

The cell probability in a three-dimensional contingency table can be written

$$P_{ijk} = P_{i..}P_{.j.}P_{..k} + P_{i..}D_{*jk} + P_{.j.}D_{i*k} + P_{..k}D_{ij*} + D_{ijk}$$

If $D_{.j*} = D_{i.*} = \ldots = D_{*j.} = 0$ and $D_{ij.} = D_{i.k} = D_{.jk} = 0$ the independent interaction parameters are reduced to D_{*11}, D_{1*1}, D_{11*} and D_{111} in the special case of the $2 \times 2 \times 2$ contingency table.

A $2 \times 2 \times 2$-table can be written in the following form:

$$\begin{aligned}
P_{111} &= P_{1..}P_{.1.}P_{..1} + P_{1..}D_{*11} + P_{.1.}D_{1*1} + P_{..1}D_{11*} + D_{111}\\
P_{112} &= P_{1..}P_{.1.}P_{..2} - P_{1..}D_{*11} - P_{.1.}D_{1*1} + P_{..2}D_{11*} - D_{111}\\
P_{121} &= P_{1..}P_{.2.}P_{..1} - P_{1..}D_{*11} + P_{.2.}D_{1*1} - P_{..1}D_{11*} - D_{111}\\
P_{122} &= P_{1..}P_{.2.}P_{..2} + P_{1..}D_{*11} - P_{.2.}D_{*11} - P_{..2}D_{11*} + D_{111}\\
P_{211} &= P_{2..}P_{.1.}P_{..1} + P_{2..}D_{*11} - P_{.1.}D_{1*1} - P_{..1}D_{11*} - D_{111}\\
P_{212} &= P_{2..}P_{.1.}P_{..2} - P_{2..}D_{*11} + P_{.1.}D_{1*1} - P_{..2}D_{11*} + D_{111}\\
P_{221} &= P_{2..}P_{.2.}P_{..1} - P_{2..}D_{*11} - P_{.2.}D_{1*1} + P_{..1}D_{11*} + D_{111}\\
P_{222} &= P_{2..}P_{.2.}P_{..2} + P_{2..}D_{*11} + P_{.2.}D_{1*1} + P_{..2}D_{11*} - D_{111}
\end{aligned} \qquad (3)$$

According to BARTLETT (1935)

$$P_{111}P_{122}P_{212}P_{221} = P_{112}P_{121}P_{211}P_{222}$$

if there is no 2nd order interaction effect.

The coupling quantity D_{111} does not measure the »BARTLETT interaction«, since the above condition cannot be met in the case of $D_{111} = 0$ either; e.g. in the case $D_{11*} \neq 0$ and $D_{1*1} \neq 0$.

$$P_{111}P_{122}P_{212}P_{221} - P_{112}P_{121}P_{211}P_{222} = [(P_{2..} - P_{1..})\,P_{.1.}P_{.2.}P_{..1}P_{..2} - (P_{..2} - P_{..1})\,P_{.1.}P_{.2.}D_{1*1} - (P_{.2.} - P_{.1.})\,P_{..1}P_{..2}D_{11*}]\,D_{11*}D_{1*1}$$

The 7 parameters $P_{1..}$, $P_{.1.}$, $P_{..1}$, D_{11*}, D_{1*1}, D_{*11} and D_{111} define a $2 \times 2 \times 2$ contingency table.

Maximum likelihood estimations (ML estimations) can be obtained from

Table 1. Examples (BAUR and DANILOVS 1980b; sample size n = 5294)
Frequencies and coupling parameters x 10^4

No.	Haplotype	P_A	P_B	P_C	D_{AB}	D_{AC}	D_{BC}	D_{ABC}	χ^2 (PIAZZA)	χ^2 new
1	A2, BW44, CW5	2492	1094	603	106	140	312	113	31.7	137.1
2	AW32, BW62, CW3	444	537	997	− 7	−25	370	− 1	6.3	0.1
3	A11, BW35, CW4	585	959	1212	92	83	699	62	0.0	70.4
4	A11, BW44, CW5	585	1094	603	−39	0	312	3	15.5	0.3
5	A25, B18, CW5	186	569	603	85	4	77	− 6	39.8	4.5

$$\hat{P}_{1..} = \frac{n_{1..}}{n_{...}}, \ldots, \hat{P}_{..1} = \frac{n_{..1}}{n_{...}}, \; D_{11*} = \frac{n_{11.}}{n_{...}} - \hat{P}_{1..}\hat{P}_{.1.},$$

$$\ldots, \hat{D}_{*11} = \frac{n_{.11}}{n_{...}} - \hat{P}_{.1.}\hat{P}_{..1}.$$

$\hat{D}_{111}$ is estimated as $\hat{P}_{111} = \frac{n_{111}}{n_{...}}$ in accordance with (2).

According to CRAMÉR (1951), the expression

$$\sum_{i,j,k} \frac{(n_{ijk} - n_{ijk}/H_o)^2}{n_{ijk}/H_o} \qquad (4)$$

is asymptotically χ^2-destributed with FG = 1 degree of freedom. The null hypothesis is formulated as $D_{111} = 0$. For the likelihood function L, we have

$$\ln L \sim \sum_{i,j,k} n_{ijk} \ln P_{ijk}$$

The probability P_{ijk} can be written as follows for $D_{ijk} = 0$

$$P_{ijk} = P_{i..}P_{.j.}P_{..k} + P_{i..}D_{*jk} + P_{.j.}D_{i*k} + P_{..k}D_{ij*} \qquad (5)$$

The ML estimations of the parameters $\hat{P}_{1..}$, $\hat{P}_{.1.}$, $\hat{P}_{..1}$, $\hat{D}_{*11}$, $\hat{D}_{1*1}$ and $\hat{D}_{11*}$ obtained from the simultaneous equations

$$\frac{\partial \ln L}{\partial P_{1..}} = 0, \ldots, \frac{\partial \ln L}{\partial D_{11*}} = 0$$

by the "regula falsi" procedure for simultaneous equations (KIESEWETTER and MAESS 1974).

The occupation numbers n_{ijk}/H_0 are yielded by

$n_{ijk}/H_0 = n_{...} \hat{P}_{ijk}/H_o$ and model (5).

The test statistic χ^2 can then be obtained from formula (4). A corresponding computer program has been written.

The following five examples have been taken from the HLA-A, B, C reference tables published in 1980 (BAUR and DANILOVS 1980b). The size of the sample is n = 5294. The 2 × 2 × 2 contigency tables were reconstructed according to (2) from the available data. First the ML estimations were ascertained on the assumption $D_{ABC} = 0$, whereafter the χ^2 value ("χ^2 new" in Table 1) was determined.

For example 3, the 2 × 2 × 2 contingency table (n = 5294) results from Table 1 (A11, BW35, CW4)

A11		
	CW4	$\overline{CW4}$
BW35	68	10
$\overline{BW35}$	13	218

A$\overline{11}$		
	CW4	$\overline{CW4}$
BW35	363	66
$\overline{BW35}$	197	4359

The ML estimations under null hypothesis ($D_{ABC} = 0$) for the paramters P_A, P_B, P_C, D_{AB}, D_{AC} and D_{BC} (multiplied by 10^4) are:

583,8; 965,4; 1216,0; 25,7; 17,2 and 696,0.

The corresponding numbers n_{ijk}/H_0 are

A11		
	CW4	$\overline{CW4}$
BW35	27,7	15,8
$\overline{BW35}$	19,0	246,6

A$\overline{11}$		
	CW4	$\overline{CW4}$
BW35	402,9	64,7
$\overline{BW35}$	194,1	4323,2

The results differ in some cases considerably from those obtained by BARTLETT's procedure (see Table 1). While the χ^2 values in PIAZZA's (1975) significance test often fail to correspond with the magnitude of the three-factor association, there is a meaningful connection between the indexes D_{ABC} and our test statistic "χ^2 new".

Literature

BARTLETT, M. S. (1935): Contingency table interactions. J. Royal. Statist. Soc., Suppl. **2,** 248–250.

BAUR, M. P. and J. A. DANILOVS(1980a): Histocompatibility testing 1980, ed. Teraski, P. I. , pp. 955–993, UCLA Tissue Typing Laboratory, Los Angeles

BAUR, M. P., and J. A. DANILOVS (1980b):Histocompatibility Testing 1980, ed. Terasaki, P. I., pp. 994–1210. UCLA Tissue Typing Laboratory, Los Angeles.

CRAMÉR, H. (1951): Mathematical Methods of Statistics. Princeton University Press, Princeton.

HILL, W. G. (1975): Tests for Association of Gene Frequencies at Several Loci in Random Mating Diploid Populations. Biometrics **31,** 881–888.

HILL, W. G. (1976): Non-Random Association of Neutral Linked Genes in Finite Populations. In: Populations Genetics and Ecology, ed. S. Karlin and E. Nevo, Academic Press, p. 339–358.

KIESEWETTER, H., and G. MAESS (1974): Elementare Methoden der numerischen Mathematik. Akademie-Verlag, Berlin 1974.

LANCASTER, H. O. (1969): The Chi-squared Distribution, Wiley & Sons, New York.

MATTIUZ, P. L., D. IHDE, A. PIAZZA, R. CEPPELLINI, and W. F. BODMER (1970): New approaches to the population genetic and segregation analysis of the HLA system. In: Histocompatibility Testing, ed. Terasaki, P. I., pp. 193–205. Munksgaard, Copenhagen.

PIAZZA, A., (1975): Haplotypes and linkage disequilibrium from the three-locus phenotypes. In: Histocompatibility Testing, ed. Kissmeyer-Nielsen, F., pp. 923–927. Munksgaard, Copenhagen.

TÖWE, J., J. BOCK, and G. KUNDT: Interactions in contingency table analysis. Biometrical Journal (in Druck).

Anschrift der Verfasser: Doz. Dr. sc. J. Töwe, Dr. G. Kundt und Dipl.-Math. M. Kracht, Abteilung für Medizinische Dokumentation und Statistik, Bereich Medizin der Wilhelm-Pieck-Universität Rostock, DDR-2500 Rostock 1, E.-Heydemann-Str. 16/17.
Doz. Dr. sc. med. R. Wegener, Institut für Gerichtliche Medizin, Bereich Medizin der Wilhelm-Pieck-Universität Rostock, DDR-2500 Rostock, Friedrich-Engels-Str. 108.

EDV in Medizin und Biologie **16** (2), 49–54, ISSN 0300-8282

Kovarianzanalyse mit heterogenen Regressionen: A. Externe Heterogenität

M. Beyerbach

Zusammenfassung

Bei der Interpretation der Ergebnisse einer Kovarianzanalyse treten Probleme auf, wenn der innerhalb der Prüfglieder realisierte Regressionskoeffizient nicht für die Korrektur der Prüfgliedmittel geeignet ist. Es wird gezeigt, daß eine Möglichkeit, dennoch zuverlässige Informationen zu gewinnen, darin besteht, den Schätzwert für den Prüfgliedeffekt zu zerlegen in einen Teil für die Heterogenität der Regressionskoeffizienten und einen anderen für die Abweichung der Prüfgliedmittel von deren Regressionsgerade. Anhand eines Beispiels zum Vergleich von Weizensorten mit Daten aus einer Erhebung wird gezeigt, daß sich diese Anteile getrennt voneinander interpretieren lassen.

Summary

Problems arise in the interpretation of the results from an analysis of covariance, if the regression coefficient, estimated within the treatments, is not available to adjust the treatment means.

It is demonstrated that one possibility to get reliable informations is to divide the treatment-effect estimator into one part for heterogeneity of regression-coefficients and another part of deviations of treatment means from their own regression-line.

An example of a comparison of wheat-varieties by observational data is given to demonstrate that these parts are separately interpretable.

1. Einleitung

Bei der Kovarianzanalyse wird der Regressionskoeffizient aus den Abweichungen von x und y innerhalb der Prüfglieder, also aus den Resten, geschätzt. Eine Schätzung des Koeffizienten aus den Prüfgliedmitteln oder der Gesamtstichprobe würde zu sehr ungenauen und eventuell auch verzerrten Werten führen, weil die Prüfgliedmittel von y (und in besonderen Fällen auch die von x) zusätzlich zum Beitrag der Regression auch von den Prüfgliedeffekten beeinflußt sind. Wird der aus den Resten geschätzte Regressionskoeffizient aber für die Korrektur der Mittel verwendet, muß vorausgesetzt werden, daß dieser auch für die Mittel gültig ist. Es muß folglich Homogenität der Parameter β_T (für die Mittel) und β_E (für die Reste) gefordert werden.

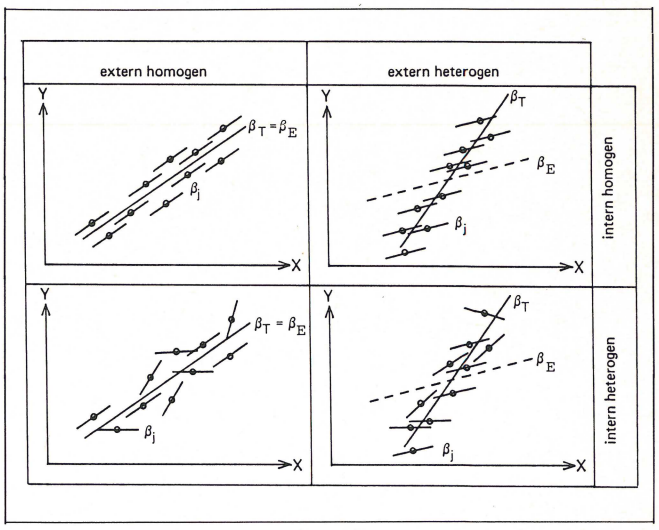
In der Literatur zur Kovarianzanalyse tauchen die Begriffe Homogenität/Heterogenität der Regressionen häufig auch in einer anderen Bedeutung auf, denn eine weitere Voraussetzung zur Kovarianzanalyse ist, daß auch die Regressionskoeffizienten innerhalb der einzelnen Prüfglieder homogen sind. Der Begriff Heterogenität sei deshalb wie folgt differenziert (vgl. Abb. 1):

„Externe" Heterogenität liegt vor, wenn sich der Koeffizient β_T der – wie SMITH (1957) sie bezeichnet – „external regression" von dem mittleren Koeffizienten innerhalb der Prüfglieder β_E der „internal regression" unterscheidet.

Unter „interner" Heterogenität sei die Verschiedenheit der Koeffizienten β_j innerhalb der J Prüfglieder verstanden. In Abbildung 1 sind die denkbaren Fälle in etwas idealisierter Form skizziert.

In zahlreichen häufig verwendeten Lehrbüchern (BORTZ, 1977; COX, 1958; FREUND und MINTON, 1979; KIRK, 1968; SNEDECOR und COCHRAN, 1956; STEEL und TORRIE, 1960; WEBER, 1972, und WINER, 1971) erstreckt sich die Diskussion um die Auswirkungen der externen Heterogenität allein auf den Hinweis, vorsichtig interpretieren zu müssen. Etwas näher gehen EVANS und ANASTASIO (1968), SPROTT (1970) und SMITH (1957) auf die Problematik ein, bieten aber auch keine allgemeingültige Lösung an. Es soll deshalb in dieser Arbeit der Versuch unternommen werden, zu demonstrieren, daß sich auch bei externer Heterogenität interpretierbare Informationen aus einer etwas modifizierten Kovarianzanalyse gewinnen lassen.

Abb. 1. Graphische Darstellung der externen und internen Heterogenität.



2. Modelle und Schätzwerte

Das der üblichen Form der Kovarianzanalyse für den i-ten Beobachtungswert im j-ten Prüffeld zugrundeliegende Modell lautet:

$$y_{ij} = \mu + \tau_j + \beta_E (x_{ij} - x..) + e_{ij} \quad (1)$$

Darin ist

μ = Gesamtmittel der Population der y, fix

τ_j = (korrigierter) fixer Effekt des j-ten Prüfgliedes, $j = 1,J$; $\sum_j n_j \tau_j = 0$;

n_j = Anzahl der Wertepaare im j-ten Prüfglied,

β_E = in allen Prüfgliedern gleicher Regressionskoeffizient, fix

$(x_{ij} - x..)$ = Abweichung des i-ten x-Wertes im j-ten Prüfglied vom Gesamtmittel der x

$x.. = \sum_j \sum_i x_{ij} / \sum_j n_j$; $i = 1, n_j$

Die Reste e_{ij} werden in allen Prüfgliedern als zufällig, unkorreliert und normalverteilt mit dem Mittel 0 und der Varianz σ_e^2 vorausgesetzt. Die x-Variable kann fix oder zufällig (WINER, 1971) sein, darf aber keine Meßfehler beinhalten. Wie im weiteren deutlich wird, muß auch vorausgesetzt werden, daß die Kovarianz zwischen den τ_i und den $x_{.j}$ gleich Null ist. Es ist

$$x_{.j} = \sum_i x_{ij}/n_j$$

entsprechend

$$y_{.j} = \sum_i y_{ij}/n_j$$

und

$$y.. = \sum_j \sum_i y_{ij} / \sum_j n_j.$$

Die Bezeichnung für die Summen der Quadrate und Produkte der Abweichungen für x und y sind der Tabelle 1 zu entnehmen. Die Berechnungsweise kann u. a. aus den in Abschnitt 1 aufgeführten Lehrbüchern ersehen werden. Als Beispiele seien gegeben:

$$T_{xx} = \sum_j n_j (x_{.j} - x..)^2$$

$$E_{Xy} = \sum_j \sum_i (x_{ij} - x_{.j})(y_{ij} - y_{.j})$$

Die Schätzwerte für die fixen Parameter μ, τ_j, β_E ergeben sich nach der Methode der kleinsten Abweichungsquadrate folgendermaßen:

$$\hat{\mu} = y.. \quad (2)$$

$$\hat{\beta}_E = E_{xy}/E_{xx} \quad (3)$$

$$\hat{\tau}_j = y_{.j} - y.. - \hat{\beta}_E (x_{.j} - x..) \quad (4)$$

Die Notwendigkeit der externen Homogenitätsvoraussetzung wird deutlicher, wenn in Modell (1)

$\beta_E (x_{ij} - x..) = \beta_A (x_{ij} - x_{.j}) + \beta_B (x_{.j} - x..)$

gesetzt wird: Geschätzt wird mittels Gleichung (3) nur β_A, in Gleichung (4) aber ein Schätzwert für β_B benötigt. Deshalb sind die Schätzwerte für die τ_j nur erwartungstreu, wenn $\beta_A = \beta_B$ ist.

Ein Modell zur Kovarianzanalyse mit externer Heterogenität enthält 2 Regressionskoeffizienten verschiedener Bedeutung:

$$y_{ij} = \mu + \beta_T (x_{.j} - x..) + \gamma_j + \beta_E (x_{ij} - x_{.j}) + e_{ij} \quad (5)$$

Dabei ist β_T der Regressionskoeffizient für die Prüfgliedmittel und γ_j die Abweichung der Mittel $y_{.j}$ von deren Regressionsgerade, mit $\sum_j n_j \gamma_j = 0$.

Wird für $\beta_T (x_{.j} - x..)$ der Ausdruck $(\beta_T - \beta_E)(x_{.j} - x..) + \beta_E (x_{.j} - x..)$ gesetzt, so läßt sich Gleichung (5) auch wie folgt schreiben:

$$y_{ij} = \mu + (\beta_T - \beta_E)(x_{.j} - x..) + \gamma_j + \beta_E (x_{ij} - x..) + e_{ij} \quad (6)$$

Ein Vergleich mit Modellgleichung (1) zeigt, daß bei externer Heterogenität der Prüfgliedeffekt τ_j durch die Summe $(\beta_T - \beta_E)(x_{.j} - x..) + \gamma_j$ ersetzt werden kann, wobei der erste, den Effekt der Heterogenität messende Summand im folgenden mit ω_j bezeichnet wird.

Somit ist bei Vorliegen externer Heterogenität der Parameter τ_j aus Modell (1) mit $x_{.j}$ korreliert und damit die bei Gleichung (1) gemachte Voraussetzung der Unabhängigkeit dieser 2 Größen verletzt. Schätzwerte für die im Modell (5) und (6) zusätzlich aufgenommenen Parameter ergeben sich wie folgt:

$$\hat{\beta}_T = T_{xy}/T_{xx} \quad (7)$$

$$\hat{\gamma}_j = y_{.j} - y.. - \hat{\beta}_T (x_{.j} - x..) = \hat{\tau}_j - \hat{\omega}_j \quad (8)$$

$\hat{\gamma}_j$ entspricht somit einem mit $\hat{\beta}_T$ anstelle von $\hat{\beta}_E$ korrigierten Prüfgliedeffekt.

Als Ergebnis des Vergleiches beider Modelle ist festzustellen: Es läßt sich der »übliche« Schätzwert für den Prüfgliedeffekt aus Modell (1) $\hat{\tau}_j$ zerlegen in die von $x_{.j}$ unabhängige Abweichung des Prüfgliedmittels von der Mittelwertsregression $\hat{\gamma}_j$ und in einen durch die externe Heterogenität erklärbaren und von $x_{.j}$ abhängigen Anteil $\hat{\omega}_j = (\hat{\beta}_T - \hat{\beta}_E)(x_{.j} - x..)$.

Der Ausdruck $\hat{\beta}_T - \hat{\beta}_E$ stellt, wie SMITH (1957) zeigen konnte, einen Schätzwert für den Regressionskoeffizienten der τ_j auf die $x_{.j}$ dar.

3. SQ-Werte

Für die weiteren Betrachtungen sind folgende den Schätzwerten zugehörigen Summen von Abweichungsquadraten von Bedeutung:

1. Der SQ-Wert für den Rest der Regression der Mittelwerte

$$T'_{yy} = \sum_j n_j \hat{\gamma}_j^2 = T_{yy} - (T_{xy}^2/T_{xx})$$

2. Der SQ-Wert der mittels (4) geschätzten Prüfgliedeffekte $\hat{\tau}_j$

$$T_{yyA} = \sum_j n_j \hat{\tau}_j^2 = T_{yy} - 2\hat{\beta}_E T_{xy} + \hat{\beta}_E^2 T_{xx}$$

3. Der »übliche« SQ-Wert für die korrigierten Prüfgliedmittel

$$T_{yyR} = T_{yy} - \frac{(T_{xy} + E_{xy})^2}{(T_{xx} + E_{xx})} + \frac{E_{xy}^2}{E_{xx}}$$

Trifft die Hypothese H_0: $\tau_1 = \tau_2 = \ldots = \tau_J$ zu, so ist bei Erfüllung der unter Modellgleichung (1) aufgeführten Voraussetzungen der Quotient

$$F = \frac{T_{yyR}/(J-1)}{\left(E_{yy} - \frac{E_{xy}^2}{E_{xx}}\right)/(N_t - J - 1)} \quad \text{F-verteilt.} \quad N_t = \sum_j n_j \cdot$$

Das Hauptinteresse gilt somit dem Wert T_{yyR}, dessen Beziehungen zu den anderen SQ-Werten wie folgt formulierbar sind (vergl. COCHRAN, 1957).

$$T_{yyR} = T'_{yy} + \left[(\hat{\beta}_T - \hat{\beta}_E)^2 T_{xx} E_{xx}/(T_{xx} + E_{xx})\right] \quad (9)$$

$$T_{yyA} = T_{yyR} + \left[(\hat{\beta}_T - \hat{\beta}_E)^2 T_{xx}^2/(T_{xx} + E_{xx})\right] = T'_{yy} + (\hat{\beta}_T - \hat{\beta}_E)^2 T_{xx} \quad (10)$$

Für $\hat{\beta}_T \neq \hat{\beta}_E$ gilt folglich:

$T_{yyA} > T_{yyR} > T'_{yy}$

Ob T_{yyR} dabei näher an T_{yyA} oder an T'_{yy} liegt, hängt vom Verhältnis $q = E_{xx}/T_{xx}$ ab:

$$T_{yyR} = T'_{yy} + (\hat{\beta}_T - \hat{\beta}_E)^2 T_{xx} \left(\frac{q}{1+q}\right) \qquad (11)$$

Da beide SQ-Werte völlig unterschiedliche Bedeutung haben, ist es im Einzelfall wichtig zu wissen, ob der unter Verwendung von T_{yyR} berechnete Testquotient eher ein Maß für die Größe der Abweichungen der Mittel von deren Regression (T'_{yy}) oder für die Differenzen zwischen den korrigierten Mitteln (T_{yyA}) ist.

Wenn q klein ist (Fall 1), also T_{xx} groß gegenüber E_{xx}, dann liegt T_{yyR} näher an T'_{yy}. Der F-Wert ist hier eher als Maß der Differenzen zwischen den $\hat{\gamma}_j$ zu bewerten und steht in einem deutlichen Mißverhältnis zu den Unterschieden zwischen den korrigierten Effekten $\hat{\tau}_j$, da er das Ausmaß dieser nur zum Teil anzeigt.

Liegt Fall 2 vor, nämlich, daß T_{xx} klein gegenüber E_{xx} ist, wird T_{yyR} näher an T_{yyA} liegen. Der F-Wert wird hier die Differenzen zwischen den korrigierten Mitteln deshalb eher vollständig anzeigen.

Im ersten Fall ist nur ein sehr kleiner Anteil der unter anderem durch die externe Heterogenität verursachten Varianz der $\hat{\tau}_j$ in T_{yyR} enthalten, im zweiten Fall hingegen beinhaltet T_{yyR} einen großen SQ-Anteil, der auf externe Heterogenität zurückzuführen ist.

Im Falle externer Heterogenität ist deshalb bei der Interpretation des F-Wertes auf die (relative) Größe der Unterschiede zwischen den x-Mitteln zu achten.

Die beiden von EVANS und ANASTASIO (1968) konstruierten Zahlenbeispiele liegen nahe dem ersten Grenzfall, somit verwundert es nicht, daß die Autoren zu dem Schluß kommen, daß bei externer Heterogenität der F-Wert kein Maß für die Differenzen der $\hat{\tau}_j$ ist. Die Beispiele sind für praktische Fälle, in denen sich die Verhältnisse etwas moderater gestalten, jedoch nicht repräsentativ. Dies zeigt auch das in Abschnitt 4 folgende Beispiel.

Liegt externe Heterogenität vor, ist generell mit dem vermehrten Auftreten signifikanter F-Werte in der Kovarianzanalyse zu rechnen, weil der rechte Teil der Gleichung (11) einen Term enthält, der bei externer Heterogenität T_{yyR} einen größeren Wert annehmen läßt, als es im entsprechenden Fall bei Homogenität zu erwarten wäre. Bei externer Heterogenität ergeben sich also selbst dann, wenn die Abweichungen der Prüfgliedmittel von deren Regressionsgerade $\hat{\gamma}_j$ nur zufällig von Null verschieden sind, vermehrt signifikante Unterschiede zwischen den $\hat{\tau}_j = \hat{\gamma}_j + \hat{\omega}_j$, die jedoch in diesem Fall nur durch den Anteil der externen Heterogenität

$\omega_j = (\beta_T - \beta_E)(x_{.j} - x_{..})$

erklärbar sind. (Bei externer Homogenität hat $\hat{\omega}_j$ den Erwartungswert Null.)

Tabelle 1. Bezeichnungen der SQ- und SP-Werte.

Ursache	SQ_x	SP_{xy}	SQ_y
Prüfglieder	T_{xx}	T_{xy}	T_{yy}
Rest	E_{xx}	E_{xy}	E_{yy}

4. Beispiel

Im südniedersächsischen Raum wird seit 1971 vom Beratungsring Ackerbau Südhannover unter Leitung von G. GOLISCH eine Erhebung der Umwelt- und Anbaufaktoren sowie der Ertragsgrößen zu Getreide und Rüben durchgeführt. Nähere Informationen zu dieser Erhebung und dem genannten Beispiel finden sich bei BEYERBACH (1983).

Bei der Auswertung sollten u. a. die verschiedenen Winterweizensorten miteinander verglichen werden, deren Leistungen wegen der naturräumlichen Heterogenität Südniedersachsens aber unter anderem um den Einfluß der Ackerzahl korrigiert werden mußten. Die dafür geeignet erscheinende Kovarianzanalyse erbrachte im Jahre 1977 die in Tabelle 2 zusammengestellten Ergebnisse. Anzumerken ist die Verschiedenheit der beiden Regressionskoeffizienten $\hat{\beta}_T$ und $\hat{\beta}_E$, und daß etwa die Hälfte des Betrages des SQ-Wertes für die korrigierten Prüfglieder auf die externe Heterogenität zurückzuführen ist. Andererseits ist dadurch T_{yyR} nur wenig kleiner als T_{yyA}. Dies wird durch den bei diesem Beispiel relativ großen q-Wert von 7,83 hervorgerufen. Auffällig ist ferner, daß in den Schätzwerten $\hat{\tau}_j = \hat{\gamma}_j + \hat{\omega}_j$ der Summand $\hat{\omega}_j$ relativ große Beträge besitzt, die $\hat{\tau}_j$ also recht verschieden von $\hat{\gamma}_j$ sind.

Aufgabe der KoVA wäre es, bei der Schätzung der Prüfgliedeffekte und deren Vergleich den störenden Einfluß

Tabelle 2. Mittelwerte, Regressionskoeffizienten und SQ-Werte für eine Erhebung bei Winterweizen. Südniedersachsen 1977.

Sorte	n	x-Mittel (Ackerzahl)	y-Mittel (Ertrag [dt/ha])	korr. y-Mittel (Ertrag [dt/ha])	$\hat{\tau}$ (korr. Effekt)	$\hat{\omega}$ (Heterog.-Anteil)	$\hat{\gamma}$ (Rest der Mittel-Regr.)
2	16	73,8	67,8	67,2	3,8	1,4	2,3
3	34	65,3	59,5	60,8	–2,6	–3,4	0,8
5	35	64,3	58,6	60,1	–3,3	–4,0	0,6
6	48	77,1	69,6	68,3	4,9	3,3	1,5
9	61	68,0	58,2	58,9	–4,5	–1,9	–2,6
10	36	74,8	67,0	66,2	2,7	1,9	0,8
11	51	73,4	65,1	64,6	1,1	1,1	0,0
12	12	61,7	62,0	64,1	0,6	–5,5	6,2
13	12	78,5	65,6	64,0	0,6	4,0	–3,5
15	39	75,2	67,6	66,7	3,2	2,1	1,1
17	15	71,1	56,1	56,2	–7,3	–0,1	–7,1
Total	359	71,3	63,4				

$\hat{\beta}_T = 0{,}79$ $\quad \hat{\beta}_E = 0{,}22$ $\quad \hat{\sigma}^2_\varepsilon = 51{,}67$ $\quad T_{xx} = 8323$ $\quad T_{xy} = 6603$ $\quad T_{yy} = 7328$

$T_{yyR} = 4505$ $\quad T_{yyA} = 4813$ $\quad T'_{yy} = 2090$ $\quad E_{xx} = 65\,156$ $\quad E_{xy} = 14\,420$ $\quad E_{yy} = 21\,121$

der Regression des Ertrages auf die nicht für alle Sorten gleich große Ackerzahl auszuschalten. Diese Forderung wird hier nur teilweise erfüllt, weil nur mit der (kleineren) Regression innerhalb der Sorten korrigiert wird, dabei aber der zusätzliche Effekt der (größeren) Regression der Sortenmittel unberücksichtigt bleibt. Die übliche Interpretation der korrigierten Sortenmittel würde hier lauten:

„Die korrigierten Sortenmittel entsprechen den Erträgen, die zu erwarten gewesen wären, wenn alle Sorten im Mittel auf Böden gleicher Güte angebaut worden wären." Diese Aussage ist jedoch kritisch zu hinterfragen, denn die Regression der Sortenmittel, die letztlich verglichen werden sollen, ist wesentlich höher als die der Erträge der Einzelschläge innerhalb der Sorten. Die wichtigste Frage lautet folglich, welches die geeignetere Regression ist, um abzuschätzen, welchen Ertrag eine Sorte erbracht hätte, wenn sie auf Böden mit anderer Ackerzahl gestanden hätte: die Regression der Sortenmittel oder die Regression innerhalb der Sorten?

Anhand der Tabelle 2 sollen dazu zwei spezielle Sortenvergleiche kritisch betrachtet werden:

1) Bei der üblichen Bewertung mittels der $\hat{\tau}_j$ würde sich zwischen den Sorten 6 und 9 ein signifikanter Unterschied von ca. 9,4 dt/ha ergeben und somit die Sorte 6 wesentlich besser beurteilt werden. In dieser Form ist der Vergleich jedoch fragwürdig, da er nicht ohne vollständige Berücksichtigung der Bodenverhältnisse gezogen wird, denn 5,2 dt/ha Differenz sind allein auf die externe Heterogenität zurückzuführen.

2) Die übliche Korrektur ($\hat{\tau}_j$-Vergleich) weist die Sorten 12 und 13 als gleichwertig aus. Wird entsprechend der Regression der Mittel korrigiert ($\hat{\gamma}_j$-Vergleich), zeigt sich Sorte 12 gegenüber der Sorte 13 ertragreicher.

Für die übliche Korrektur mit $\hat{\beta}_E$ spricht, daß dieser Koeffizient für die Einzelschläge innerhalb der Sorten als gültig beobachtet wurde und deshalb auch für die Korrektur der Mittel geeignet sein sollte. Dies trifft insbesondere dann zu, wenn das Gesamtmittel für x innerhalb des beobachteten Bereiches für jede einzelne Sorte liegt.

Die Korrektur mit $\hat{\beta}_E$ führt jedoch nicht zu plausiblen Ergebnissen: Wird mittels $\hat{\beta}_E$ geschätzt, wie hoch der Ertrag der auf guten Böden angebauten Sorten wäre, wenn sie auf weniger fruchtbaren Böden stünden, dann wären sie den dort anbauüblichen Sorten noch weit überlegen. Diese Schätzung ist jedoch anzuzweifeln, denn wenn dies zuträfe, also $\hat{\beta}_E$ ein für die Korrektur geeigneter Koeffizient wäre, dann wären diese Sorten auf den schlechteren Böden auch stärker vertreten. Entsprechendes gilt für die Korrektur mit $\hat{\beta}_T$: Wäre dieser Koeffizient der für die Korrektur geeignete, würden einige der auf minderwertigeren Böden angebauten Sorten – wie etwa Sorte 12 – auf gutem Boden sehr hohe Erträge versprechen und hier häufiger im Anbau stehen, was jedoch auch nicht der Fall ist. Offenbar ist somit keiner der beiden Koeffizienten für die Korrektur verwendbar.

Die Schätzung der γ_j liefert aber in jedem Fall eine wertvolle Zusatzinformation: Da $\hat{\gamma}_{12}$ hier positiv ist, der (unkorrigierte) Mittelwert für die Sorte 12 also oberhalb der Regressionsgeraden für die Mittelwerte liegt, läßt sich aus dieser Betrachtungsweise heraus sagen, daß Sorte 12 eine „relativ gute" Sorte ist, da sie auf schlechten Böden bessere Erträge verspricht als die anderen Sorten auf vergleichbaren Böden. Diese Aussage gilt aber nur für die durch die mittlere Ackerzahl $x_{.12}$ gegebene Umweltbedingung und ist nicht unbedingt auf andere Bodenverhältnisse übertragbar.

Die Varianz der „mit $\hat{\beta}_T$ korrigierten Effekte" $\hat{\gamma}_j$ läßt sich aus der Streuung der Sortenmittel um deren eigene Regressionsgerade schätzen und mit einer geeigneten Restvarianz, etwa dem Rest innerhalb der Gruppen $\hat{\sigma}_e^2$, vergleichen. Dieser Test entspräche dann dem von WINER (1971) aufgeführten Test mit dem Quotienten F_3 und ist im Falle externer Heterogenität als ein Test der Hypothese $\gamma_j = 0$, alle j anzusehen. Große Testquotienten deuten darauf hin, daß die Sortenmittel recht stark um die Regressionsgerade im Mittel streuen, daß es sich also lohnt, nach „relativ" schlechten ($\hat{\gamma}$ negativ) und relativ guten Sorten zu suchen. Für das Beispiel ist F_3 hochsignifikant.

Da eine Betrachtung der $\hat{\gamma}_j$ recht informativ ist, wurde aus allen Erhebungsdaten der Jahre 1973 bis 1979 für jedes Jahr eine KoVA mit den Sorten, für die $n_j \geq 10$ galt, in der gleichen Weise berechnet. Die Ergebnisse sind in den Tabellen 3 bis 6 aufgeführt. Tabelle 6 zeigt deutlich, daß die im Beispiel angesprochene externe Heterogenität nicht zufällig auftrat, sondern in allen Jahren zum Ausdruck kommt.

Ausgehend von den korrigierten Effekten $\hat{\tau}_j$ würden nun die Sorten Benno und Saturn als recht gut, Diplomat und Topfit hingegen als weniger geeignet beurteilt werden. Dies liegt jedoch, wie aus Tabelle 4 ersichtlich ist, größtenteils an den Heterogenitätseffekten $\hat{\omega}_j$ und somit also an der Korrektur mit der mittleren Regression. Die Sorte Topfit wird nach ihrem $\hat{\tau}_j$ deshalb so schlecht beurteilt, weil deren $\hat{\omega}_j$ stets negativ ist, Topfit also häufiger auf schlechteren Böden angebaut wird. Die $\hat{\gamma}_j$ in Tabelle 5 zeigen, daß Topfit keinesfalls unterbewertet werden sollte, da sie auf Böden niedrigerer Ackerzahl eher eine „relativ gute" Sorte darstellt.

Deutliche Aussagen über die Eigenschaften aller Sorten sind jedoch nicht möglich. Sowohl die $\hat{\tau}_j$ als auch die $\hat{\gamma}_j$ ändern sich für viele Sorten von Jahr zu Jahr recht stark. Typische gut bzw. relativ gut zu bewertende Sorten sind nicht erkennbar. Dies hat u. a. die Ursache, daß die Regression der Mittelwerte mit der jeweiligen Stichprobengröße der einzelnen Sorten gewichtet ist und deshalb die Gerade etwas zu den stärker vertretenen Sorten „hingezogen" wird, diese also kleinere Beträge von $\hat{\gamma}_j$ aufweisen. Ferner wirken sich der starke Sortenwechsel sowie die Interaktionen Jahre * Sorten, Jahre * Orte und vor allem Sorten * Orte negativ auf die γ-Analyse aus. Die Daten dieser Erhebung sind deshalb für eine Aussage über die relative Güte einer Sorte nur beschränkt geeignet.

5. Diskussion

In der Kovarianzanalyse können zwei verschiedene Regressionskoeffizienten berechnet werden: $\hat{\beta}_T$ für die Prüfgliedmittel und $\hat{\beta}_E$ für den Rest. Die Korrektur mit $\hat{\beta}_T$ gibt einen Hinweis auf die „relative Güte" des Prüfgliedes bei der mittleren Ackerzahl $x_{.j}$. Die übliche Korrektur mit $\hat{\beta}_E$ führt zu einem Vergleich, der nur dann gerechtfertigt und richtig ist, wenn sich die Sorten in den ihnen zugehörigen Ackerzahl-Mitteln nur wenig unterscheiden, wenn also der Einfluß der externen Heterogenität gering ist. Ein Vergleich der mit $\hat{\beta}_E$ korrigierten Mittel zweier Sorten, die auf Böden recht unterschiedlicher Güte angebaut wurden, ist aus folgenden Gründen wenig sinnvoll:

1. Die externe Heterogenität verzerrt den Vergleich, da die in der Differenz der korrigierten Mittel enthaltene Differenz der $\hat{\omega}_j$ in diesem Fall recht groß werden kann.
2. Die korrigierten Mittelwerte sind dann mit Extrapolationsfehlern versehen, wenn das Gesamtmittel für die Ackerzahl nicht innerhalb der Variationsbreite jeder einzelnen Sorte liegt.
3. Eventuell vorhandene sortenspezifische Regressionen sind nicht berücksichtigt.

4. Bei großen Unterschieden zwischen den x-Mitteln wird möglicherweise eine Korrigierbarkeit auf einen mittleren Wert für x unterstellt, der biologisch gar nicht sinnvoll oder möglich ist.

Die sich für den Getreidebauer ergebende Konsequenz aus dem Auftreten externer Heterogenität ist, daß bei einem Bodenwechsel für Winterweizen auch die Sortenfrage neu überdacht werden muß. Steht dem Landwirt durch Zukauf oder Pacht Boden von angenommen 10 Bodenpunkten höherer Qualität zur Verfügung, so ließe sich bei Anbau der alten Sorte im Mittel der Jahre 1973–1979 eine Ertragssteigerung von ca. 2 dt/ha erwarten, ein geeigneter Sortenwechsel hingegen verspricht einen Mehrertrag von ca. 5,5 dt/ha (vgl. Tab. 6). Aber auch eine solche Kalkulation ist kritisch zu beurteilen, weil sie durch die Wechselwirkungen der Faktoren Sorten, Orte und Jahre in Frage gestellt wird.

Abschließend soll die Frage diskutiert werden, wodurch die externe Heterogenität zustande kommt:

Es ist zu bedenken, daß sich hinter den verschiedenen Sortenmitteln für die Ackerzahl jeweils andere Orte bzw. Regionen verbergen, die sich auch in den klimatischen Daten, den Anbaumaßnahmen und anderen Merkmalen unterscheiden. In die hier dargestellten Sortenertragsmittel gehen also auch die Ortseffekte und die Interaktionen beider Faktoren mit ein. Die externe Heterogenität erklärt sich also formell durch die bei höheren Ackerzahlen ebenfalls höheren Mischeffekte, die sich mittels der vorhandenen Daten und des Modells aber nicht in Orts-, Sorten- und Interaktionseffekt aufgliedern lassen. Ferner hat SMITH (1957) darauf hingewiesen, daß die Regression von Genotypenmitteln nicht denselben Koeffizienten wie die Regression der Individuen innerhalb der Genotypen erwarten läßt, weil die Varianzen und Kovarianzen zumeist eine Komponente für „Zwischen-Genotypen" enthalten. KAHNEMANN (1965) zeigt auf, daß externe Heterogenität auch durch den Schätzfehler eines der beiden Regressionskoeffizienten hervorgerufen werden kann. Diese Gefahr ist besonders groß, wenn die x-Variable kein sehr zuverlässiger Vertreter für das Merkmal ist, das sie messen sollte, denn dann wird insbesondere der Koeffizient innerhalb der Sorten zu klein geschätzt. Diese Möglichkeit könnte auch bei dieser Erhebung gegeben sein. Ein anderer Fall liegt vor, wenn der sortenspezifische Regressionskoeffizient der „Intensivsorten" höher ist als der auf den minderwertigen Böden angebauten Sorten (vgl. WRICKE und WEBER, 1980). So könnten die Anbauer auf guten Böden die Sorten mit hohem spezifischem Regressionskoeffizienten bevorzugen, auf schlechteren Böden hingegen die Sorten, die auf eine Standortverschlechterung wenig reagieren. Dies würde bedeuten, daß es sich bei der Regression der Sortenmittel um eine durch die Sortenwahl beeinflußte Beziehung handelt, während dieser Effekt bei der mittleren Regression innerhalb der Sorten in dieser Form nicht auftritt. Somit kann die externe Heterogenität auch allein als Folge der Wechselwirkung Sorten * Orte verstanden werden, da der Landwirt bei gegebener Bodengüte (Ort) bewußt eine spezielle Sorte, die sich unter den gegebenen Verhältnissen bewährt hat, auswählen wird. Für die Erklärung der externen Heterogenität sind also mehrere Ansätze denkbar. Bei geplanten Versuchen hat der Ansteller die Möglichkeit, das Auftreten externer Heterogenität zu vermeiden, indem alle Sorten unter gleichen Bedingungen geprüft werden. Nicht gewährleistet ist die Voraussetzung der Unabhängigkeit von Prüfglied und Kovariabler, wenn die Größe der x-Variablen eine Eigenschaft des Prüfgliedes ist, oder das Prüfglied, wie etwa eine Behandlung, auf die x-Variable einwirkt. Unter solchen Gegebenheiten kann das Vorliegen externer Heterogenität nicht ausgeschlossen werden.

Tabelle 3. Korrigierte Sorteneffekte $\hat{\tau}_j$ (dt/ha).

Sorte	1973	1974	1975	1976	1977	1978	1979
Benno		1,5	3,6	0,8			
Saturn			2,9	–0,8	3,8	3,8	0,2
Caribo	–0,4	–1,2	0,5	–2,6	–2,6	1,9	0,3
Diplomat	–1,3	0,5	–2,5	–1,0	–3,3	–6,4	
Joss		–3,0					
Jubilar	–1,5	–4,2	–3,5				
Kormoran		–1,3	2,0	1,2	–4,5	1,6	–3,8
Kranich	3,4	2,5	0,5	2,4	2,7	–2,6	–4,5
M. Huntsman				2,8	1,1	–1,9	–4,7
Topfit	0,8	0,3	–2,7	–2,2	0,6	–5,9	
Carimulti					0,6	5,9	2,9
Vuka					3,2	2,1	–0,6
Kolibri				–4,1	–7,3	–11,2	
Sonstige	–2,8	–1,9		0,1		0,3	1,4

Tabelle 4. Heterogenitätsanteil $\hat{\omega}_j$ von $\hat{\tau}_j$ (dt/ha).

Sorte	1973	1974	1975	1976	1977	1978	1979
Benno		1,6	4,3	0,2			
Saturn			2,4	1,1	1,4	1,4	0,5
Caribo	0,4	–0,2	–0,7	–0,3	–3,4	–2,1	–0,8
Diplomat	–0,8	–0,8	–1,6	–0,1	–4,0	–1,4	
Joss		–0,4					
Jubilar	–1,2	–1,0	–0,1				
Kormoran		1,9	0	0,1	–1,9	–2,3	–1,0
Kranich	2,6	0,5	0,7	–0,1	1,9	0,3	–0,6
M. Huntsman				0,5	1,1	1,7	0,9
Topfit	–2,5	–2,2	–2,7	–2,1	–5,5	–0,9	
Carimulti					4,0	0,9	–0,1
Vuka					2,1	1,2	0
Kolibri				0,4	–0,1	–1,8	
Sonstige	–3,9	0,7		1,2		0,5	0,4

Tabelle 5. Abweichung $\hat{\gamma}_j$ der Sortenmittel von der Mittelwertregression (dt/ha).

Sorte	1973	1974	1975	1976	1977	1978	1979
Benno		–0,1	–0,7	0,6			
Saturn			0,4	–1,9	2,3	2,3	–0,3
Caribo	–0,8	–1,0	1,2	–2,3	0,8	3,9	1,1
Diplomat	–0,6	1,3	–0,9	–0,9	0,6	–4,9	
Joss		–2,7					
Jubilar	–3,3	–3,2	–3,4				
Kormoran		–3,2	1,9	1,1	–2,6	3,9	–2,7
Kranich	0,8	1,9	–0,1	2,5	0,8	–2,8	–3,9
M. Huntsman				2,3	0	–3,6	–5,7
Topfit	3,4	2,6	0	–0,1	6,2	–4,9	
Carimulti					–3,5	5,1	3,1
Vuka					1,1	0,8	–0,6
Kolibri				–4,5	–7,1	–9,4	
Sonstige	1,1	–2,6		–1,1		–0,2	1,0

Tabelle 6. Koeffizienten der Regression des Ertrages auf die Ackerzahl in (dt/ha)/Bodenpunkt.

	1973	1974	1975	1976	1977	1978	1979
$\hat{\beta}_E$	0,19	0,12	0,20	0,37	0,22	0,09	0,17
$\hat{\beta}_T$	0,86	0,41	0,58	0,57	0,79	0,38	0,31

Danksagung

Für die Durchsicht des Manuskriptes und wertvolle Diskussion danke ich Herrn Prof. Dr. H. Rundfeldt und Herrn Dr. W. E. Weber.

Literatur

BEYERBACH, M. (1983): Einsatzmöglichkeiten kovarianzanalytischer Verfahren bei der Auswertung von Erhebungen. Dissertation. Fachbereich Gartenbau der Universität Hannover.

BORTZ, J. (1977): Lehrbuch der Statistik für Sozialwissenschaftler. Springer Verlag, Berlin, Heidelberg, New York.

COCHRAN, W. G. (1957): Analysis of Covariance: Its Nature and Uses. Biometrics **13,** 261–281.

COX, D. R. (1958): Planning of Experiments. John Wiley & Sons, Inc., New York, London, Sydney.

EVANS, H. S., and E. J. ANASTASIO (1968): Misuse of Analysis of Covariance when Treatment Effect and Covariable are confounded. Psychological Bulletin, Vol. 69, No. 4, 225–234.

FREUND, R. J., and P. D. MINTON (1979): Regression Methods. Marcel Dekker Inc., New York, Basel.

KAHNEMANN, D. (1965): Control of spurious association and the reliability of the controlled variable. Psychological Bulletin **64,** 326–329.

KIRK, R. E. (1968): Experimental Design: Procedures for the Behaviorial Sciences. Brooks/Cole Publishing Company, Belmont, California.

SMITH, H. F. (1957): Interpretation of adjusted treatment means and regressions in analysis of covariance. Biometrics **13,** 282–308.

SNEDECOR, G. W., and W. G. COCHRAN (1956): Statistical Methods. Fifth edition, Iowa State University Press, Ames, Iowa, USA.

SPROTT, D. A. (1970): Note on Evans and Anastasio on the Analysis of Covariance. Psychological Bulletin, Vol. 73, No. 4, 303–306.

STEEL, R. G. D., and J. H. TORRIE (1960): Principles and Procedures of Statistics. McGraw Hill Book Company, Inc., New York.

WEBER, E. (1972): Grundriß der biologischen Statistik. 7. Aufl. Gustav Fischer Verlag, Stuttgart.

WINER, B. J. (1971): Statistical Principles in Experimental Design. Second Edition, McGraw Hill, Kogakusha, Ltd.

WRICKE, G. und W. E. WEBER (1980): Erweiterte Analyse von Wechselwirkungen in Versuchsserien. Medizinische Informatik und Statistik **17,** 87–95. Springer-Verlag, Berlin, Heidelberg, New York.

Anschrift des Verfassers: Dr. M. Beyerbach, Inst. f. Statistik und Biometrie der Tierärztlichen Hochschule Hannover, Bischofsholer Damm 15, 3000 Hannover 1

EDV in Medizin und Biologie **16** (2), 54–60, ISSN 0300-8282

Kovarianzanalyse mit heterogenen Regressionen B. Interne Heterogenität

M. Beyerbach

Zusammenfassung

Bei der Kovarianzanalyse ergeben sich bei der Schätzung und dem Testen von Parametern Probleme, wenn die Regressionskoeffizienten innerhalb der Prüfglieder heterogen sind. Diese Schwierigkeiten und alternative Methoden der Korrektur und des Vergleichs von Mitteln werden diskutiert.

Summary

In the analysis of covariance some problems of estimation and testing the parameters arise, when the regression coefficients within the groups are heterogeneous. These problems and alternative methods of adjusting and comparing the means were under discussion.

1. Einleitung

Eine wichtige Voraussetzung für die erwartungstreue Schätzung korrigierter Prüfgliedeffekte bei der Kovarianzanalyse ist die Homogenität der Regressionen innerhalb der Prüfglieder. Sind die Regressionskoeffizienten dagegen ungleich, liegt eine Heterogenität vor, die hier präziser als ,,interne" Heterogenität bezeichnet werden soll. Sie unterscheidet sich damit von der ,,externen" Heterogenität, die dann gegeben ist, wenn bei interner Homogenität die Prüfgliedmittel einem anderen Regressionskoeffizienten als die Einzelwerte innerhalb der Prüfglieder folgen (vgl. BEYERBACH, 1985). Bei interner Heterogenität sind, wie im folgenden deutlich wird, einige F-Testquotienten verzerrt, ferner sind Schätzung und Vergleich der korrigierten Prüfgliedeffekte erschwert. In dieser Arbeit werden dazu von Ergebnissen aus Simulationsstudien unterstützte Vorschläge gemacht.

2. Modelle, Schätzwerte und Tests

Das der üblichen Form der Kovarianzanalyse für den i-ten Beobachtungswert im j-ten Prüfglied zugrundeliegende Modell lautet:

$$y_{ij} = \mu + \tau_j + \beta_E(x_{ij} - x..) + e_{ij} \quad (1)$$

Darin ist

μ = Gesamtmittel der Population der y, fix

τ_j = (korrigierter) fixer Effekt des j-ten Prüfgliedes, $j = 1, J$; $\sum_j n_j\tau_j = 0$;

n_j = Anzahl der Wertepaare im j-ten Prüfglied,

β_E = in allen Prüfgliedern gleicher Regressionskoeffizient, fix

$(x_{ij} - x..)$ = Abweichung des i-ten x-Wertes im j-ten Prüfglied vom Gesamtmittel der x,

$x.. = \sum_j \sum_i x_{ij} / \sum_j n_j$; $i = 1, n_j$

Die Reste e_{ij} werden in allen Prüfgliedern als zufällig, unkorreliert und normalverteilt mit dem Mittel 0 und der Varianz σ_e^2 vorausgesetzt. Die x-Variable kann fix oder zufällig (WINER, 1971) sein, darf aber keine Meßfehler beinhalten. Vorausgesetzt werden muß ferner, daß die Kovarianz zwischen den τ_j und den $x_{\cdot j}$ gleich Null ist. Es ist

$$x_{\cdot j} = \sum_i x_{ij}/n_j,$$

entsprechend

$$y_{\cdot j} = \sum_i y_{ij}/n_j$$

und

$$y.. = \sum_j \sum_i y_{ij} / \sum_j n_j.$$

Als Bezeichnungen der Summen der Quadrate und Produkte der Abweichungen von x und y werden im folgenden verwendet:

$$E_{xy_j} = \sum_i (x_{ij} - x_{\cdot j})(y_{ij} - y_{\cdot j}), \qquad E_{xy} = \sum_j E_{xy_j},$$

$$E_{xx_j} = \sum_i (x_{ij} - x_{\cdot j})^2, \qquad E_{xx} = \sum_j E_{xx_j},$$

E_{yy_j} und E_{yy}

entsprechend.

Die Schätzwerte für die fixen Paramter μ, β_E, τ_j ergeben sich nach der Methode der kleinsten Abweichungsquadrate folgendermaßen:

$$\hat{\mu} = y.. \quad (2)$$

$$\hat{\beta}_E = E_{xy}/E_{xx} \quad (3)$$

$$\hat{\tau}_j = y_{\cdot j} - y.. - \hat{\beta}_E(x_{\cdot j} - x..) \quad (4)$$

Liegt interne Heterogenität vor, ist das Modell (1) unzutreffend. Geeigneter ist die Form:

$$y_{ij} = \mu + \beta_j(x_{ij} - x..) + \tau_j^* + f_{ij} \quad (5)$$

Bei interner Heterogenität muß die Definition des Prüfgliedeffektes auf einen bestimmten x-Wert bezogen werden: τ_j^* sei der Effekt des j-ten Prüfgliedes an der Stelle x..! Es ist β_j der prüfgliedspezifische Regressionskoeffizient und f_{ij} die Restabweichung von der prüfgliedspezifischen Regression. Ein Vergleich mit Modell (1) zeigt, daß

$$f_{ij} = e_{ij} - (\beta_j - \beta_E)(x_{ij} - x_{\cdot j})$$

ist, wobei die f_{ij} wieder als unkorreliert und normalverteilt mit dem Mittel Null und der Varianz σ_f^2 vorausgesetzt werden. Aus der Gegenüberstellung wird deutlich, daß $\hat{e}_{ij}$ bei interner Heterogenität einen systematischen Fehler vom Betrage

$$(\hat{\beta}_j - \hat{\beta}_E)(x_{ij} - x_{\cdot j})$$

aufweist. Einen Schätzwert für σ_f^2 liefert MS_f' in Tabelle 1. Liegt Heterogenität vor, ist $MS_f' < MS_e'$.
Unter Berücksichtigung von

$$\hat{\beta}_j = E_{xy_j}/E_{xx_j} \quad (6)$$

wird bei interner Heterogenität der Prüfgliedeffekt bei x.. wie folgt geschätzt:

$$\tau_j^* = y_{\cdot j} - y.. - \hat{\beta}_j(x_{\cdot j} - x..) \quad (7)$$

(vergl. HENDERSON, 1982).
Ein Vergleich der Schätzformeln (4) und (7) führt zu folgender Feststellung: Bei Vernachlässigung der internen Heterogenität ergibt sich anstelle eines Schätzwertes für den Prüfgliedeffekt τ_j^* der Wert

$$\hat{\tau}_j = \hat{\tau}_j^* + (\hat{\beta}_j - \hat{\beta}_E)(x_{\cdot j} - x..) \quad (8)$$

Die übliche Zerlegung des SQ-Wertes für y innerhalb der Prüfglieder zum Testen der internen Heterogenität ist in Tabelle 1 aufgeführt (vergl. WINER, 1971). Primäres Interesse

Tabelle 1. Zerlegung von E_{yy} und F-Tests der Regressionen $\left(N_t = \sum_j n_j\right)$.

Variationsursache	FG	SQ	MQ	F
Mittlere Regression innerhalb der Prüfglieder β_E	1	$\hat{\beta}_E \cdot E_{xy}$		$F_E = \hat{\beta}_E \cdot E_{xy}/MS_e'$
Abweichungen von der mittleren Regression e_{ij}	$N_t - J - 1$	$E_{yy}' = E_{yy} - \hat{\beta}_E \cdot E_{xy}$	MS_e'	
→ Interne Heterogenität $\beta_j - \beta_E$	$J - 1$	$S_2 = E_{yy}' - S_1$		$F_2 = \frac{S_2}{J-1}/MS_f'$
→ Summe der Abw. von den spez. Regressionen f_{ij}	$N_t - 2J$	$S_1 = \sum_1^J \left(E_{yy_j} - \hat{\beta}_j \cdot E_{xy_j}\right)$	MS_f'	
Gesamt innerhalb der Prüfglieder	$N_t - J$	$E_{yy} = \hat{\beta}_E \cdot E_{xy} + S_1 + S_2$		

bei der Kovarianzanalyse gilt zumeist dem F-Wert F_k für die korrigierten Prüfgliedeffekte. Seine Berechnung ist in zahlreichen Lehrbüchern, u. a. bei STEEL und TORRIE (1960, S. 315) und WINER (1971, S. 770) beschrieben. F_k dient zum Überprüfen der Hypothese der Gleichheit aller τ_j. Ein weiterer von WINER (1971) für den Vergleich von nur je zwei korrigierten Effekten $\hat{\tau}_j$ angegebener F-Testquotient zur Hypothese $H_0: \tau_1 = \tau_2$ hat die Form

$$F_w = (\hat{\tau}_1 - \hat{\tau}_2)^2 / \left(MS'_e \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(x_{.1} - x_{.2})^2}{E_{xx}} \right) \right) \qquad (9)$$

$$FG = 1; N_t - J - 1 \text{ mit } N_t = \sum_j n_j$$

Bei interner Heterogenität sind die beiden Testquotienten F_k und F_w verzerrt, weil der Schätzwert für den Rest, MS'_e, zu groß ausfällt. Dies ist dadurch zu erklären, daß die Koeffizienten β_j innerhalb der J Prüfglieder unterschiedlich sind und deshalb in dem SQ-Wert für die Abweichungen von der mittleren Regression E'_{yy} auch ein recht großer Anteil S_2 für die interne Heterogenität enthalten ist (vgl. Tabelle 1). Da diese Restvarianz MS'_e jedoch für diese F-Tests als Nenner genutzt wird, führen die Tests zu konservativen Entscheidungen. Bei Vorliegen interner Heterogenität liegen die Testquotienten also seltener über den tabellierten Werten, als es bei rein zufälliger Versuchsanordnung und Regressionshomogenität zu erwarten wäre. Die Testquotienten fallen um so kleiner aus, je größer der Anteil der Heterogenität an der Restvarianz innerhalb der Gruppe ist. Für die Praxis bedeutet dies, daß Differenzen in den Prüfgliedeffekten, die bei interner Homogenität eventuell entdeckt worden wären, bei Heterogenität übersehen werden. Der F-Wert für die mittlere Regression F_E (vgl. Tabelle 1) fällt bei interner Heterogenität (bei einem mittleren Koeffizienten $\beta = 0$) zu oft in den signifikanten Bereich, es ergeben sich also häufiger »gesicherte« Werte, als bei Homogenität zu erwarten wäre.

Nähere Angaben zu den hier gemachten Aussagen und Informationen zum Einfluß der internen Heterogenität auf weitere hier nicht behandelte Testquotienten zu Themenkreis Kovarianzanalyse sind bei BEYERBACH (1983) aufgeführt.

3. Korrektur und Vergleich von Mittelwerten

Sind die Regressionskoeffizienten innerhalb der Prüfglieder heterogen, entsprechen die – mit der mittleren Regression – korrigierten Mittelwerte nicht den Mittelwerten, die für die Prüfglieder bei x.. erwartet werden. Es liegt deshalb nahe, bei interner Heterogenität die Mittel jeweils mit ihrer eigenen prüfgliedspezifischen Regression zu korrigieren. Diese Methode hat den Vorteil, dem eigentlichen Anliegen der Kovarianzanalyse, nämlich der Schätzung von y-Mitteln und deren Vergleich bei vorgegebenen gleichen x-Mitteln, näherzukommen. Ein Nachteil dieses Verfahrens besteht darin, daß die prüfgliedspezifischen Regressionskoeffizienten nicht so genau geschätzt werden können wie eine mittlere Regression bei interner Homogenität. Für den Fall, daß sich die Streubreiten der einzelnen Prüfglieder bezüglich der x-Variablen deutlich unterscheiden, ist es auch fraglich, ob diese spezifischen Regressionen für den Gesamtbereich von x gültig sind, ob also zum Beispiel die Linearität der Regressionen außerhalb der prüfgliedspezifischen Streubreite von x noch zutrifft und somit extrapoliert werden darf.

Bei prüfgliedspezifischer Korrektur tritt der neue Aspekt auf, daß die Differenz zweier derart korrigierter Mittel von x abhängig ist, weshalb ein biologisch sinnvolles Vergleichsniveau gegeben sein muß. Allgemeingültige Regeln zur Festlegung eines solchen x-Wertes lassen sich nicht angeben, da sich die Wahl des Vergleichsniveaus an der biologischen Fragestellung orientieren sollte.

Die Probleme sind bei einem Vergleich von nur 2 Prüfgliedmitteln am leichtesten zu übersehen. Deshalb sind in Abbildung 1, welche die Verhältnisse verdeutlichen möge, auch nur 2 von ja zumeist insgesamt mehr als 2 Prüfgliedern berücksichtigt.

Für einen auf ein gemeinsames x-Mittel bezogenen Vergleich kommt zunächst die in der üblichen Form korrigierte Differenz DA in Frage, die jedoch, wie u. a. aus Formel (8) hervorgeht, im Falle interner Heterogenität nicht geeignet ist, da sie dann den Erwartungswert

$$(\tau_1^* - \tau_2^*) + (f_{.1} - f_{.2}) + (\beta_1 - \beta_E)(x_{.1} - x..) - (\beta_2 - \beta_E)(x_{.2} - x..)$$

besitzt, welcher auch, wenn τ_1^* bei x.. gleich τ_2^* ist, ungleich Null wird. Die $f_{.j}$ stellen die Mittel der f_{ij} im j-ten Prüfglied (mit dem Erwartungswert Null) dar.

Eventuell sinnvollere Vergleiche erlauben die Mittel MP_j. Hier ist MP1 der Mittelwert von Prüfglied 1, der zu erwarten wäre, wenn er den x-Wert von Prüfglied 2 aufgewiesen hätte. Entsprechend stellt MP2 den unter den Bedingungen von Prüfglied 1 bezüglich der x-Variablen zu erwartenden Mittelwert von Prüfglied 2 dar (Interpretationsmöglichkeiten werden an späterer Stelle gegeben). Angemerkt sei, daß die Differenz

$$D1 = |y(M1) - y(MP2)| \text{ kleiner ist als die Differenz}$$
$$D2 = |y(M2) - y(MP1)|,$$

weil die Vergleiche bei verschiedenen x-Werten gezogen werden.
Die Schätzgleichung für D1 lautet:

$$D1 = y_{.1} - \left(y_{.2} - \hat{\beta}_2 (x_{.2} - x_{.1}) \right) \qquad (10)$$

Die Varianz von D1 läßt sich folgendermaßen herleiten: Die Varianz der Restabweichungen von den Regressionen (Streuung um die Gerade) wird bei interner Heterogenität aus MS'_f geschätzt, damit die Verschiedenheit der Regressionen nicht in den Schätzwert für die Restvarianz eingeht (vergl. Tab. 1). Die Varianz von M1 wird durch MS'_f/n_1 geschätzt (STEEL und TORRIE, 1960). Die Varianz von MP2 entspricht der eines korrigierten Mittelwertes, ergibt sich also aus:

$$s^2_{MP2} = MS'_f \left(\frac{1}{n_2} + \frac{(x_{.2} - x_{.1})^2}{E_{xx_2}} \right) \qquad (11)$$

(vergl. STEEL und TORRIE, 1960). Die Varianz der Differenz (hier D1) eines Mittels M_j und eines »paarweise prüfgliedspezifisch korrigierten Mittels« MP_k ergibt sich somit zu:

$$s^2_{D_j} = MS'_f \left(\frac{1}{n_j} + \frac{1}{n_k} + \frac{(x_{.k} - x_{.j})^2}{E_{xx_k}} \right) \qquad (12)$$

Einen F-Test für die Differenz D_j mit der Hypothese $D_j = 0$ kann aber der Quotient

$$F_{D_j} = D_j^2 / s_{D_j}^2 \quad ; \quad FG = 1, N_t - 2J \qquad (13)$$

nicht liefern, weil beispielsweise der Erwartungswert für D1 gleich

$$(\beta_1 - \beta_2)(x_{1.} - x..) + (\tau_1^* - \tau_2^*) + (f_{1.} - f_{2.})$$

ist, und somit bei interner Heterogenität auch bei Gültigkeit der Nullhypothese ($\tau_1^* = \tau_2^*$ bei x..) einen von Null verschiedenen Summanden enthält.

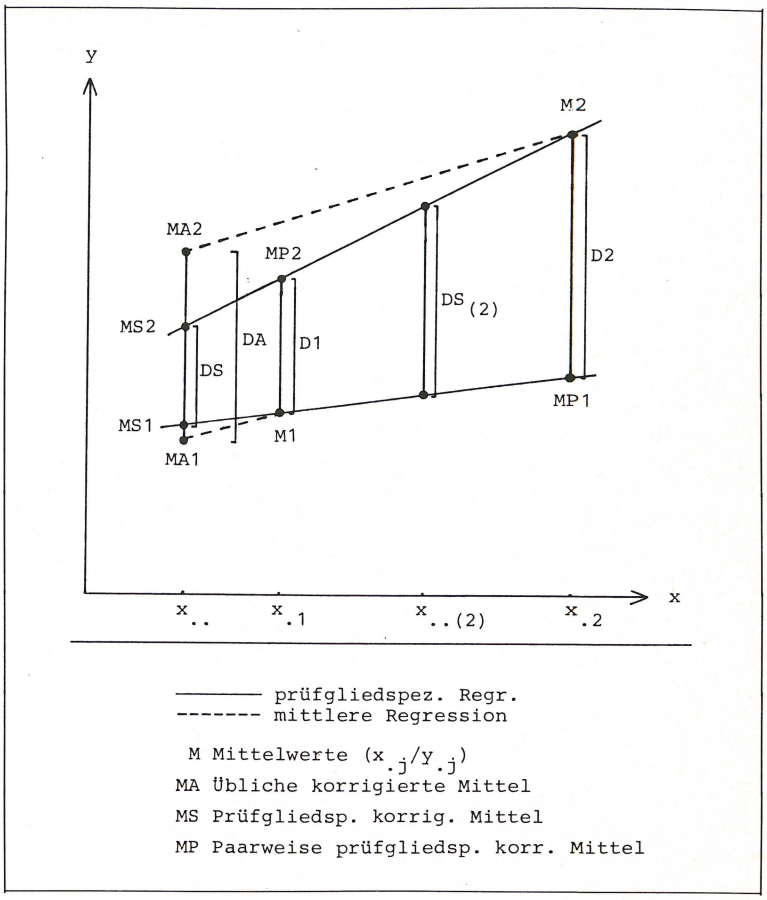


Abb. 1. Verschiedene Möglichkeiten der kovarianzanalytischen Korrektur von Mittelwerten.

Die Differenz DS der prüfgliedspezifisch korrigierten Mittel MS stellt den Vergleich wieder bei x.. an. Die Varianz der Differenz ergibt sich aus der Summe der Varianzen der MS und somit zu:

$$s^2_{DS} = MS'_f \left(\frac{1}{n_1} + \frac{(x._1 - x..)^2}{E_{xx_1}} + \frac{1}{n_2} + \frac{(x._2 - x..)^2}{E_{xx_2}} \right) \quad (14)$$

Eine Prüfgröße für DS läßt sich also wie folgt formulieren:

$$F_{DS} = DS^2/s^2_{DS} \quad (15) \quad ; \quad FG = 1, N_t - 2J, \text{ wobei}$$

$$DS = y._1 - \hat{\beta}_1 (x._1 - x..) - y._2 + \hat{\beta}_2 (x._2 - x..) = \hat{\tau}^*_1 - \hat{\tau}^*_2 \quad (16)$$

Der Erwartungswert für DS ist $(\tau^*_1 - \tau^*_2) + (f._1 - f._2)$ und ist somit – im Gegensatz zu denen von DA und D1 – bei Gültigkeit der Nullhypothese auch im Falle interner Heterogenität gleich Null.

Eine weitere Möglichkeit besteht darin, je 2 Mittel prüfgliedspezifisch zu korrigieren, und auf der Basis von

$$x..{}_{(2)} = (\sum_i x_{i1} + \sum_i x_{i2})/(n_1 + n_2)$$

zu vergleichen. Die Nullhypothese wird insoweit modifiziert, daß nun τ^*_1 bei $x..{}_{(2)}$ gleich τ^*_2 ist. Der entsprechende Testquotient ist:

$$F_{DS(2)} = DS^2_{(2)}/s^2_{DS(2)} \quad (17)$$

$DS^2_{(2)}$ und $s^2_{DS(2)}$ werden berechnet, indem in Formel (14) und (16) anstelle x.. der Wert $x..{}_{(2)}$ eingesetzt wird. Trifft die modifizierte Hypothese zu, ist der Erwartungswert von $DS_{(2)}$ ebenfalls gleich Null. Formell besteht dieses Vorgehen darin, für jedes Prüfgliedpaar die Parameter eines Formel (5) entsprechenden Modells zu schätzen, wobei dann aber für jeden Vergleich ein anderes Niveau $x..{}_{(2)}$ gewählt wird. Die Hypothese, daß τ^*_1 bei x.. gleich τ^*_2 ist, läßt sich mit $F_{DS(2)}$ nicht testen. Umgekehrt ist F_{DS} kein für den Test der Hypothese $\tau^*_1 = \tau^*_2$ bei $x..{}_{(2)}$ geeigneter Quotient.

Eine Variante des von WINER (1971) vorgeschlagenen F-Testes für interne Heterogenität ergibt sich, wenn in Formel (9) statt MS'_e der Schätzwert MS'_f eingesetzt wird und anstelle DA die Differenz DS genutzt wird.

Der Testquotient F_v hat nun folgendes Aussehen:

$$F_v = DS^2 / \left(MS'_f \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(x._1 - x._2)^2}{2E_{xx}/J} \right) \right) \quad (18)$$

$$FG = 1, N_t - 2J$$

Eine weitere Möglichkeit besteht darin, im Zähler die Differenz DA zu belassen, und nur im Nenner MS'_f anstelle MS'_e zu setzen.

4. Ergebnisse aus simulierten Versuchen

Um die Eignung der verschiedenen hier entwickelten Testquotienten bei interner Heterogenität miteinander vergleichen zu können, wurde eine Simulationsstudie durchgeführt, bei der nachstehend aufgeführte Quotienten für den Vergleich je zweier auf verschiedene Weise korrigierter Mittel berücksichtigt wurden.

1a) üblicher F-Test für korrigierte Prüfglieder F_K (vgl. STEEL und TORRIE, 1960; WINER, 1971)

1b) F-Test F_w zweier korrigierter Mittel nach WINER, Formel (9), (für J = 2 gilt $F_k = F_w$)

2) F_{w_2}: Zähler wie F_k und F_w, aber im Nenner nicht MS'_e, sondern MS'_f
3) F_v gemäß den Erläuterungen zu Formel (18)
4) F_{DS} für die prüfgliedspezifisch korrigierten Mittel (Formel [15])
5) je zwei F_{D_j} für die »paarweise prüfgliedspezifisch« korrigierten Mittel (Formel [13]).

Bei der Simulation wurden Regressionsheterogenität und Stichprobengrößen entsprechend den in Tabelle 2 gemachten Angaben variiert. Erzeugt wurden gemäß der Modellgleichung (5) mit $\tau_1^* = \tau_2^* = 0$ je 20 000 Fälle pro Konstellation, wodurch sich für die beiden F_{D_j}-Quotienten insgesamt eine Anzahl von 40 000 ergibt. Die Reste f_{ij} wurden zufällig einer normalverteilten Grundgesamtheit mit dem Mittelwert Null entnommen, so daß die $\hat{\tau}_j^*$ nur zufällig von Null verschieden sind, die Hypothese, daß bei x.. keine Prüfgliedeffekte τ_j^* bestehen, also gültig ist. Die x_{ij} entstammen ebenfalls einer Normalverteilung.

Die Ergebnisse sind in Tabelle 2 zusammengestellt, wobei in der Spalte unter E (%) die relativen Häufigkeiten aufgeführt sind, die ein auch bei interner Heterogenität F-verteilter Testquotient erwarten ließe. Da für gleiche Wiederholungszahlen und 2 Prüfglieder $F_k = F_w$ gilt, wurden deren Überschreitungshäufigkeiten hier nur jeweils einmal angegeben. Die Studie läßt nachstehend aufgeführte Aussagen über die Eigenschaften der Testquotienten zu:

Die Tests mit den Quotienten F_k und F_w werden mit steigender Heterogenität konservativer, wobei diese Verzerrung mit Vergrößerung der Wiederholungszahl bei den vorgegebenen Konstellationen geringfügig ansteigt. Umgekehrt zeigt der Testquotient F_{w_2} mit ansteigender Heterogenität vermehrt signifikante Differenzen an, und zwar bei kleinerer Wiederholungszahl in stärkerer Ausprägung. Die Testquotienten F_v und F_{DS} führten zu fast identischen Ergebnissen, und ihre Verteilungen lagen der F-Verteilung am nächsten, besonders bei der größeren Wiederholungszahl. Der Quotient F_{D_j} ist (erwartungsgemäß) nicht F-verteilt.

5. Beispiel für die Interpretation der verschiedenen Mittel

Die Unterschiede zwischen den verschiedenen Korrekturmöglichkeiten bei interner Heterogenität sollen noch einmal anhand eines einfachen Beispieles, das sich wie in Abbildung 2 skizzieren läßt, aufgezeigt werden (vgl. auch Abb. 1).

Es ist allgemein bekannt, daß Forellen ein größeres Sauerstoffbedürfnis haben als Karpfen. Der Regressionskoeffizient (Fischertrag/Jahr u. m^3 Wasser) pro (mg O_2/l Teichwasser) wird deshalb bei Forellen höher sein als bei Karpfen. Angenommen, ein Teichwirt hätte eine Teichanlage A, deren Teiche relativ schlecht mit Sauerstoff versorgt sind und in denen er deshalb Karpfen mästet, sowie eine gut durchlüftete Forellenteich-Anlage B. Es steht die Frage offen, mit welcher Fischart eine dritte Anlage C besetzt werden soll, die in ihrer Sauerstoffversorgung etwa in der Mitte zwischen A und B liegt, ansonsten aber mit A und B vergleichbar ist. Diese Frage kann mit einer KoVA der Daten aus den Anlagen A und B beantwortet werden, wobei ein einzelner Teich die Erhebungseinheit bzw. Parzelle darstellt. Die übliche Überführung der Mittel M in die korrigierten Mittel MA würde den Ertrag beider Fischarten in der Anlage C überschätzen, denn die richtigen spezifisch korrigierten Mittel MS liegen niedriger. Die Mittel MS zeigen, daß die Anlage C besser mit Karpfen zu besetzen ist, wenn angenommen wird, daß Produktionskosten und Preis pro kg Fisch für beide Arten gleich sind. Die übliche Korrektur könnte in diesem Fall ferner zu der falschen Schlußfolgerung führen, daß die Forellenanlage auch nur mit Karpfen zu besetzen ist. Das paarweise spezifisch korrigierte Karpfen-Mittel MP_k zeigt jedoch, daß die Karpfen, würden sie in der Forellenanlage gemästet, einen geringeren Ertrag als die Forellen erbrächten.

Tabelle 2. Ergebnisse einer Simulationsstudie zum Mittelwertvergleich bei interner Heterogenität.

Konstellation		$n_1 = n_2 = 12$ $\beta_1 = 1, \beta_2 = 2, J = 2$	$n_1 = n_2 = 12$ $\beta_1 = 0, \beta_2 = 3, J = 2$	$n_1 = n_2 = 33$ $\beta_1 = 1, \beta_2 = 2, J = 2$	$n_1 = n_2 = 33$ $\beta_1 = 0, \beta_2 = 3, J = 2$
Testquotient	E (%)	Häufigkeit (%)	Häufigkeit (%)	Häufigkeit (%)	Häufigkeit (%)
$F_k = F_w$	90	91,7	98,2	92,8	99,0
	5	4,3	1,2	4,3	0,7
	4	3,2	0,6	2,5	0,3
	1	0,8	0,1	0,5	0,1
F_{w_2}	90	88,4	85,4	89,4	88,0
	5	5,3	6,4	5,1	5,7
	4	4,9	5,7	4,2	5,1
	1	1,4	2,5	1,1	1,2
F_v	90	88,3	87,9	89,7	89,4
	5	5,5	5,6	5,0	5,2
	4	4,8	4,9	4,3	4,2
	1	1,4	1,7	1,1	1,3
F_{DS}	90	88,4	88,0	89,7	89,5
	5	5,5	5,5	4,9	5,2
	4	4,8	4,8	4,3	4,1
	1	1,4	1,7	1,1	1,3
F_{D_j}	90	85,7	65,4	85,4	64,9
	5	5,9	9,1	6,3	8,7
	4	6,1	12,6	6,2	13,0
	0,9	1,8	8,3	1,8	8,3
	0,1	0,5	4,6	0,3	5,0

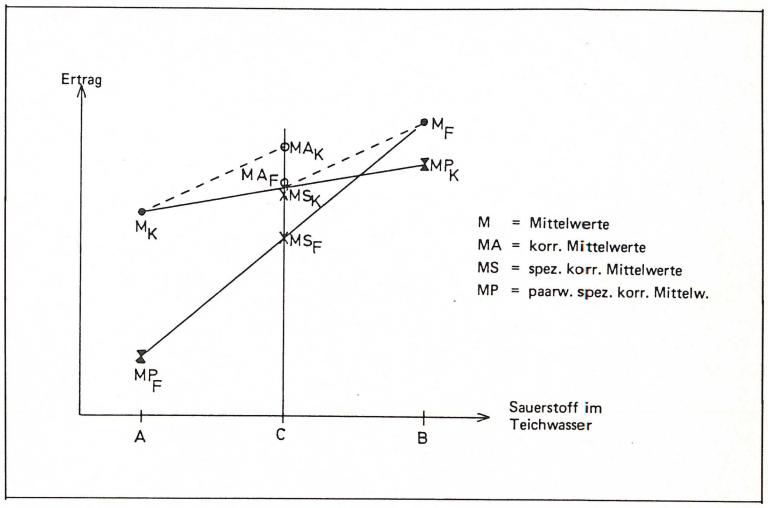


Abb. 2. Forellen- und Karpfenertrag dreier Teichanlagen.

In den Abschnitten 2.–4. sind die Auswirkungen der internen Heterogenität auf die bei der KoVA üblichen Tests aufgezeigt und besprochen worden. Zusammenfassend ergibt sich, daß die bei der »Standard-KoVA« angewandten F-Tests bei interner Heterogenität konservativ sind und daß Unterschiede zwischen den korrigierten Prüfgliedern oft übersehen werden. Eine Möglichkeit, diese Situation zu umgehen, besteht in der Korrektur der Prüfgliedmittel mit ihrer spezifischen Regression. Dabei tritt allerdings die Schwierigkeit auf, daß, wie die Simulationsstudie zeigt, für die Tests der Differenzen der so prüfgliedspezifisch korrigierten Prüfgliedmittel keine der dort angegebenen Testquotienten als F-verteilt angesehen werden konnten, wohl aber die Richtung der Verzerrung dieser Testquotientenverteilungen deutlich wurde. Verglichen mit der hohen vorgegebenen internen Heterogenität liegen die Verteilungen der beiden für diesen Spezialfall formulierten Testquotienten F_v und F_{DS} nahe der F-Verteilung und können deshalb zumindest bei der Auswertung von Versuchen und Erhebungen im Pflanzenbau Hilfestellung leisten. Dabei ist zu bedenken, daß diese Tests den Fehler der 1. Art nicht ganz einhalten, weshalb für die Entscheidung eine etwas kleinere Irrtumswahrscheinlichkeit α vereinbart werden muß.

6. Diskussion

Interne Heterogenität ist Ausdruck einer altbekannten Eigenschaft biologischer Daten: Die Wirkung des einen Merkmals, der x-Variablen, hängt ab von der Ausprägung des anderen, nämlich der des Prüffaktors, es handelt sich also um eine Wechselwirkung zwischen einem qualitativen und einem quantitativen Merkmal. Für derartige Interaktionen gibt es seit langem Beispiele: Aus der Pflanzenzüchtung ist bekannt, daß die verschiedenen Genotypen einer Kulturpflanzenart in der Regel unterschiedliche Fähigkeiten zeigen, sich veränderten Umweltbedingungen anzupassen. Bei der Tierproduktion hängt die Gewichtszunahme von der Fütterungsart ab. Die statistische Abhängigkeit des Körpergewichtes von der Körpergröße ist beim Menschen geschlechtsspezifisch. Gehölze zeigen artspezifische Empfindlichkeiten gegenüber Schadstoffen in der Luft. All dies sind im weitesten Sinne Beispiele, bei deren kovarianzanalytischer Auswertung mit von interner Heterogenität hervorgerufenen Problemen gerechnet werden muß.

In der Literatur finden sich diese Thematik betreffend u. a. folgende Arbeiten:

COCHRAN (1969) gibt Modellgleichungen und ein Beispiel an, in dem der Effekt eines Französisch-Sprachkurses im Fernsehen geschätzt werden soll, schlägt aber keinen Test für die mit einer paarweise prüfgliedspezifisch korrigierten Differenz vergleichbare Differenz zwischen den Teilnehmern und Nichtteilnehmern vor. DE LURY nennt schon 1948 die Möglichkeit, prüfgliedspezifisch zu korrigieren, führt dies aber nicht weiter aus. FLEISS (1971) gibt ein Modell für interne Heterogenität im Blockversuch an. HENDERSON (1982) schlägt vor, unkorrigierte Differenzen $(\hat{\tau}_1^* - \hat{\tau}_2^*) + (\hat{\beta}_1 - \hat{\beta}_2)(x^* - x..)$ für verschiedene x^* zu berechnen, was einer Analyse der $y._j$ bei verschiedenen, sinnvollen x entspricht. Dabei wird die Summe $\tau_j^* + \beta_j (x^* - x..)$ als Prüfgliedeffekt verstanden. HENDRIX, CARTER und SCOTT (1982) regen an, die mit der prüfgliedspezifischen Regression korrigierten Mittel bei $x..$ und $x.. \pm s_x$ zu vergleichen, wobei s_x die Gesamt-Standardabweichung von x darstellt. ROGOSA (1980) bespricht die Möglichkeiten des Vergleichs zweier nichtparalleler Regressionsgeraden und geht dabei auch auf die Problematik der Schätzung und des Vergleichs der Prüfgliedeffekte ein. Er unterscheidet zwischen »pick-a-point-procedures« und der Berechnung einer »region of significance«. Im ersten Fall werden die zwei Prüfglieder nur bei vorgegeben bestimmten x-Werten verglichen, für deren Auswahl ROGOSA Kriterien diskutiert, im 2. Fall werden Intervalle für die x-Variable berechnet, innerhalb derer sich die Erwartungswerte für y (mit der Irrtumswahrscheinlichkeit α) unterscheiden. SEARLE (1979) zeigt Lösungswege im Fall interner Heterogenität bei 2-faktoriellen Kovarianzanalysen auf.

Danksagung

Für die Durchsicht des Manuskriptes und Diskussion danke ich Herrn Prof. Dr. H. Rundfeldt und Herrn Privatdozent Dr. W. E. Weber.

Literatur

BEYERBACH, M. (1983): Einsatzmöglichkeiten kovarianzanalytischer Verfahren bei der Auswertung von Erhebungen. Dissertation des Fachbereiches für Gartenbau der Universität Hannover.

BEYERBACH, M. (1985): Kovarianzanalyse mit heterogenen Regressionen A: Externe Heterogenität. EDV in Medizin und Biologie **16,** 49–60.
COCHRAN, W. G. (1969): The Use of Covariance in Observational Studies. Appl. Stat. **18,** 270–275.
DE LURY, D. B. (1948): The Analysis of Covariance. Biometrics **4,** 153–170.
FLEISS, J. L. (1971): Testing for Equal Slopes in Randomized Block Design with Covariance. Biometrics **27,** 225–229.
HENDERSON, C. R., Jr. (1982): Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions. Biometrics **38,** 623–640.
HENDRIX, L. J., CARTER, M. W. and SCOTT, D. T. (1982): Covariance analysis with heterogeneity of slopes in fixed models. Biometrics **38,** 641–650.
ROGOSA, D. (1980): Comparing nonparallel regression lines. Psychological Bulletin **88,** 307–321.
SEARLE, S. R. (1979): Alternative covariance models for the two-way crossed classification. Communications in Statistics, Series **A8,** 799–818.
STEEL, R. G. D. und TORRIE, J. H. (1960): Principles and Procedures of Statistics. McGraw Hill Book Company, Inc., New York.
WINER, B. J. (1971): Statistical Principles in Experimental Design. Second Edition, McGraw Hill, Kogakusha, Ltd.

Eingegangen am 6. November 1984
Anschrift des Verfassers: Dr. M. Beyerbach, Inst. f. Stat. u. Biometrie der Tierärztlichen Hochschule Hannover, Bischofsholer Damm 15, 3000 Hannover 1

EDV in Medizin und Biologie **16** (2), 60–64, ISSN 0300-8282

Parameterschätzung bei der quadratischen Kovarianzanalyse

M. Beyerbach

Zusammenfassung

Bei der quadratischen Kovarianzanalyse sind zur Schätzung der Modellparameter nach der Methode der kleinsten Abweichungsquadrate zwei Berechnungsarten gebräuchlich: a) die Methode der Koeffizientenanpassung und b) die konventionelle Lehrbuchmethode, oft unter »multipler Kovarianzanalyse« beschrieben. Weitere Schwierigkeiten entstehen durch verschiedene Möglichkeiten, die x-Variable für das quadratische Glied zu definieren. Die unterschiedlichen Methoden, Modelle und Bedeutungen der Schätzwerte werden diskutiert.

Summary

In the analysis of covariance with parabolic regression two calculation-methods were custumary to estimate the model parameters by the LS-method: a) the method of fitting constants and b) the conventional textbook method, often described under »multiple covariance«. Further complication arises from several ways to define the squared x-variable. The different methods, models and meanings of the estimators were under discussion.

1. Einleitung

Die Kovarianzanalyse ist ein wichtiges Hilfsmittel der Forschung im Bereich biologischer Wissenschaften. Sie dient

a) zur Bereinigung von Schätzwerten für Prüfgliedmittel eines untersuchten Merkmals (y) vom Einfluß weiterer Störgrößen quantitativer Natur (x) und
b) zur Beseitigung des Einflusses von Prüfgliedeffekten auf Schätzwerte für Parameter aus Regressionsmodellen,

mit anderen Worten also, zur Trennung der (vermischten) Wirkungen qualitativer und quantitativer Merkmale. Im folgenden wird der Fall einer Begleitvariablen x und eines Prüffaktors betrachtet. Der Zusammenhang zwischen dem untersuchten Merkmal (y) und der Begleitvariablen (x) wird dabei oft durch eine lineare Regression beschrieben. In vielen Fällen wird eine Gerade den gegebenen biologischen Verhältnissen jedoch nicht entsprechen und ein Polynom 2. Grades dem wahren Sachverhalt näherkommen.

In vielen Lehrbüchern (KIRK, 1968; SCHEFFÉ, 1959; SNEDECOR und COCHRAN, 1967; STEEL und TORRIE, 1960; WINER, 1971) ist die quadratische Kovarianzanalyse nur indirekt in Form der multiplen, zweifach linearen Kovarianzanalyse beschrieben. Einer der wenigen Hinweise auf die Möglichkeit, eine Kovarianzanalyse auch quadratisch durchführen zu können, findet sich bei SNEDECOR und COCHRAN (1967) auf Seite 460: »If the regression is found to be curved, the treatment means are adjusted for the parabolic regression. The calculations follow the method given in Sektion 14.8« (In Abschnitt 14.8 wird die einfaktorielle lineare Kovarianzanalyse mit zwei x-Variablen beschrieben.) Dem ist anzumerken, daß bei der multiplen Kovarianzanalyse im dreidimensionalen Raum korrigiert und interpretiert wird, die quadratische Kovarianzanalyse aber ein zweidimensionales Problem ist, weshalb die von SNEDECOR und COCHRAN (1967) angegebene Formel für diesen Fall einer Änderung bedarf.

Ein weiteres Problem der quadratischen Kovarianzanalyse besteht in der Zusammenfassung mehrerer quadratischer Regressionen, etwa aus Versuchsserien, zu einer für alle Prüfglieder bzw. Versuche gültigen quadratischen Regression.

2. Modell und Schätzgleichung

Die quadratische Kovarianzanalyse kann neben dem in den genannten Lehrbüchern unter multipler Kovarianzanalyse beschriebenen Verfahren auch mit der Methode der Koeffizientenanpassung (Method of fitting constants, vgl. SEARLE, 1971) durchgeführt werden.

Das Modell lautet:

$$y = D \cdot \delta + e \qquad (1)$$

Darin ist

- y der n_t-dimensionale Vektor der Beobachtungswerte von y mit $n_t = \sum_1^J n_j$, n_j = Anzahl Wiederholungen im j-ten Prüfglied, $j = 1,J$
- e der n_t-dimensionale Vektor der nichterklärbaren Reste bzw. Meßfehler
- δ ein J + 2-dimensionaler Vektor der Parameter mit $\delta' = (\alpha_1, \alpha_2, \ldots, \alpha_J, \beta_1, \beta_2)$, worin
  - α_j = Konstante des j-ten Prüfgliedes
  - β_1 = Regressionskoeffizient von y auf x_1 (linear)
  - β_2 = Regressionskoeffizient von y auf x_2 (quadratisch)

  (Auf die Bedeutung von x_1 und x_2 wird im folgenden noch eingegangen)
- D eine n_t* (J+2)-dimensionale Koeffizientenmatrix, bestehend aus den Untermatrizen Z und X.
  Z enthält J n_t-dimensionale Koeffizientenvektoren $z_1, \ldots, z_J$ mit den Koeffizienten für $\alpha_1, \ldots, \alpha_J$. Diese sind gleich 1, wenn die Beobachtung aus dem j-ten Prüfglied stammt, sonst 0. X enthält zwei n_t-dimensionale Koeffizientenvektoren x_1 und x_2 für die zwei Regressionskoeffizienten β_1 und β_2.

Die Reste e_{ij} werden als zufällig und normalverteilt vorausgesetzt, sie dürfen ferner nicht korreliert sein und müssen in allen Prüfgliedern homogene Varianz aufweisen. Ferner muß die Voraussetzung getroffen werden, daß für jedes Prüfglied die Regressionskoeffizienten β_1 und β_2 gleich sind und daß die Effekte der Prüfglieder nicht mit den zugehörigen Mitteln von x korreliert sind. Die x-Variablen sollten fehlerfrei gemessen sein, weil andernfalls die Beträge der Regressionskoeffizienten zu klein geschätzt werden. Bei der quadratischen Kovarianzanalyse besteht nun die Möglichkeit, x_1 und x_2 auf verschiedene Weisen zu definieren. Es sollen hier 4 denkbare, teils in der Regressionsanalyse Anwendung findende Modellvarianten gegenübergestellt werden, um aufzuzeigen, daß sich, je nach Definition der x-Variablen, die Schätzwerte für deren Parameter in Zahlenwert und Bedeutung unterscheiden.

Die Varianten lauten:

A)	$x_{ij1} = x_{ij}$;	$x_{ij2} = x_{ij}^2$
B)	$x_{ij1} = (x_{ij} - x..)$;	$x_{ij2} = (x_{ij} - x..)^2$
C)	$x_{ij1} = (x_{ij} - x..)$;	$x_{ij2} = x_{ij}^2 - (x..)^2$
D)	$x_{ij1} = (x_{ij} - x..)$;	$x_{ij2} = x_{ij}^2 - (x^2)..$

Darin ist

- x_{ij} der Meßwert von x in der i-ten Wiederholung im j-ten Prüfglied
- $x..$ das Mittel, $x.. = \sum_j \sum_i x_{ij}/n_t$, $i = 1,n_j$
- $(x..)^2$ das Quadrat des Mittels und
- $(x^2)..$ das Mittel der Quadrate, $(x^2).. = \sum_j \sum_i x_{ij}^2/n_t$

Prüfgliedmittel werden im folgenden mit $x._j$ (bzw. $x._{j1}$ und $x._{j2}$ für die beiden aus x abgeleiteten Variablen) und mit $y._j$ bezeichnet, die Gesamtmittel mit $x..1$, $x..2$ und $y..$.

Der J+2-dimensionale Vektor $\hat{\delta}$ der nach der Methode der kleinsten Abweichungsquadrate berechneten Schätzwerte für die Parameter ergibt sich (vergl. SEARLE; 1971) aus der Vorschrift:

$$(\hat{\alpha}_1, \ldots, \hat{\alpha}_J, \hat{\beta}_1, \hat{\beta}_2)' = \hat{\delta} = (D'D)^{-1}D'y \qquad (2)$$

Die einzelnen für y an der Stelle x_{ij} zu erwartenden Werte lassen sich aus $\hat{y} = D\hat{\delta} = (ZX)\hat{\delta}$ schätzen.

3. Beispiel

Die weiteren Erläuterungen sollen an folgendem Zahlenbeispiel mit 3 Prüfgliedern und $n_1 = 10$, $n_2 = 6$, $n_3 = 4$ demonstriert werden:

Für Variante A seien y, Z, X:

$$y = \begin{bmatrix} 74 \\ 72 \\ 76 \\ 74 \\ 74 \\ 72 \\ 75 \\ 73 \\ 71 \\ 75 \\ 67 \\ 67 \\ 69 \\ 71 \\ 65 \\ 67 \\ 69 \\ 71 \\ 73 \\ 77 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad X_A = \begin{bmatrix} 0 & 0 \\ 2 & 4 \\ 0 & 0 \\ 2 & 4 \\ 1 & 1 \\ 1 & 1 \\ 3 & 9 \\ 1 & 1 \\ 2 & 4 \\ 2 & 4 \\ 2 & 4 \\ 3 & 9 \\ 3 & 9 \\ 4 & 16 \\ 1 & 1 \\ 1 & 1 \\ 2 & 4 \\ 3 & 9 \\ 3 & 9 \\ 4 & 16 \end{bmatrix}$$

Als Mittelwerte ergeben sich: $y.. = 71{,}6$ $x.. = 2{,}0$ $(x^2).. = 5{,}3$. Bei den Varianten B, C, D verbleiben y und Z in unveränderter Form, die Matrix X hat entsprechend den vorhergehenden Definitionen das Aussehen:

$$X_B = \begin{bmatrix} -2 & 4 \\ 0 & 0 \\ -2 & 4 \\ 0 & 0 \\ -1 & 1 \\ -1 & 1 \\ 1 & 1 \\ -1 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 2 & 4 \\ -1 & 1 \\ -1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 2 & 4 \end{bmatrix} \quad X_C = \begin{bmatrix} -2 & -4 \\ 0 & 0 \\ -2 & -4 \\ 0 & 0 \\ -1 & -3 \\ -1 & -3 \\ 1 & 5 \\ -1 & -3 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 5 \\ 1 & 5 \\ 2 & 12 \\ -1 & -3 \\ -1 & -3 \\ 0 & 0 \\ 1 & 5 \\ 1 & 5 \\ 2 & 12 \end{bmatrix} \quad X_D = \begin{bmatrix} -2 & -5{,}3 \\ 0 & -1{,}3 \\ -2 & -5{,}3 \\ 0 & -1{,}3 \\ -1 & -4{,}3 \\ -1 & -4{,}3 \\ 1 & 3{,}7 \\ -1 & -4{,}3 \\ 0 & -1{,}3 \\ 0 & -1{,}3 \\ 0 & -1{,}3 \\ 1 & 3{,}7 \\ 1 & 3{,}7 \\ 2 & 10{,}7 \\ -1 & -4{,}3 \\ -1 & -4{,}3 \\ 0 & -1{,}3 \\ 1 & 3{,}7 \\ 1 & 3{,}7 \\ 2 & 10{,}7 \end{bmatrix}$$

Tabelle 1. Komponenten der Schätzwertvektoren $\hat{\delta}$ für das Beispiel

Variante	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$
A	75	68	72	–3	1
B	73	66	70	1	1
C	73	66	70	–3	1
D	74,3	67,3	71,3	–3	1

Die Schätzwerte sind für die 4 Varianten in Tabelle 1 aufgelistet. Zu erkennen ist, daß sich sowohl die $\hat{\alpha}_j$ wie auch $\hat{\beta}_1$ und $\hat{\beta}_2$ bei den 4 Modellvarianten unterscheiden.

4. Vergleich der Schätzwerte

Für die Verschiedenheit der Schätzungen findet sich durch deren Gegenüberstellung eine Erklärung. Nachfolgend sind alle 4 Schätzgleichungen für die y_{ij} in eine gut mit der Gleichung (3) vergleichbare Form gebracht:

Variante A:

$$\hat{y}_{ij} = \hat{\alpha}_{j(A)} + \hat{\beta}_{1(A)} x_{ij} + \hat{\beta}_{2(A)} x_{ij}^2 \tag{3}$$

Variante B:

$$\begin{aligned}\hat{y}_{ij} &= \hat{\alpha}_{j(B)} + \hat{\beta}_{1(B)}(x_{ij} - x..) + \hat{\beta}_{2(B)}(x_{ij} - x..)^2 \\ &= \left(\hat{\alpha}_{j(B)} - \hat{\beta}_{1(B)} x.. + \hat{\beta}_{2(B)}(x..)^2\right) + \left(\hat{\beta}_{1(B)} - 2\hat{\beta}_{2(B)} x..\right) x_{ij} \\ &\quad + \hat{\beta}_{2(B)} x_{ij}^2\end{aligned} \tag{4}$$

Variante C:

$$\begin{aligned}\hat{y}_{ij} &= \hat{\alpha}_{j(C)} + \hat{\beta}_{1(C)}(x_{ij} - x..) + \hat{\beta}_{2(C)}\left(x_{ij}^2 - (x..)^2\right) \\ &= \left(\hat{\alpha}_{j(C)} - \hat{\beta}_{1(C)} x.. - \hat{\beta}_{2(C)}(x..)^2\right) + \hat{\beta}_{1(C)} x_{ij} + \hat{\beta}_{2(C)} x_{ij}^2\end{aligned} \tag{5}$$

Variante D:

$$\begin{aligned}\hat{y}_{ij} &= \hat{\alpha}_{j(D)} + \hat{\beta}_{1(D)}(x_{ij} - x..) + \hat{\beta}_{2(D)}\left(x_{ij}^2 - (x^2)..\right) \\ &= \left(\hat{\alpha}_{j(D)} - \hat{\beta}_{1(D)} x.. - \hat{\beta}_{2(D)}(x^2)..\right) + \hat{\beta}_{1(D)} x_{ij} + \hat{\beta}_{2(D)} x_{ij}^2\end{aligned} \tag{6}$$

Wie die Gleichungen (3)–(6) zeigen, lassen sich die geschätzten Parameter der Variante A durch eine lineare Transformation aus den Schätzwerten der Varianten B, C und D erzeugen.

Es gilt:

$$\hat{\beta}_{2(A)} = \hat{\beta}_{2(B)} = \hat{\beta}_{2(C)} = \hat{\beta}_{2(D)} \tag{7}$$

$$\hat{\beta}_{1(A)} = \hat{\beta}_{1(B)} - 2\hat{\beta}_{2(B)} x.. = \hat{\beta}_{1(C)} = \hat{\beta}_{1(D)} \tag{8}$$

$$\begin{aligned}\hat{\alpha}_{j(A)} &= \hat{\alpha}_{j(B)} - \hat{\beta}_{1(B)} x.. + \hat{\beta}_{2(B)}(x..)^2 \\ &= \hat{\alpha}_{j(C)} - \hat{\beta}_{1(C)} x.. - \hat{\beta}_{2(C)}(x..)^2 \\ &= \hat{\alpha}_{j(D)} - \hat{\beta}_{1(D)} x.. - \hat{\beta}_{2(D)}(x^2)..\end{aligned} \tag{9}$$

Den Gleichungen (7)–(9) ist u. a. zu entnehmen:

1. $\hat{\beta}_2$ ist von der Definition der x-Variablen unabhängig.
2. $\hat{\beta}_1$ ist von der Definition der x-Variablen abhängig, bei quadratischer Regression ist also insbesondere bei der Interpretation des linearen Koeffizienten auf die Definition der x-Variablen zu achten.
3. Aus den Gleichungen 7 und 8 folgt:

$$-\hat{\beta}_{1(B)} x.. + \hat{\beta}_{2(B)}(x..)^2 = -\hat{\beta}_{1(C)} x.. - \hat{\beta}_{2(C)}(x..)^2$$

Daher ist $\hat{\alpha}_{j(B)} = \hat{\alpha}_{j(C)}$, während $\hat{\alpha}_{j(A)} \neq \hat{\alpha}_{j(B)} = \hat{\alpha}_{j(C)} \neq \hat{\alpha}_{j(D)}$ und $\hat{\alpha}_{j(A)} \neq \hat{\alpha}_{j(D)}$; (siehe auch Tab. 1).

5. Interpretation der $\hat{\alpha}_j$

Die Bedeutung der Schätzwerte für die Varianten A und B geht aus Abbildung 1 hervor. In dieser Abbildung gelten die durchgezogene Ordinate und die über der Abszisse angegebenen x-Werte für die Variante A, für Variante B sind die unterbrochen gezeichnete y-Achse und die x-Werte unter der x-Achse gültig. Bedingt durch die Definition der x-Variablen sind die Schätzwerte zu den Varianten C und D zweidimensional nicht graphisch darstellbar, weil sich x_2 nicht auf derselben Achse wie x_1 kennzeichnen läßt, da hier $x_{ij2} \neq (x_{ij1})^2$ ist.

Deutlich wird, daß die $\hat{\alpha}_{j(A)}$ den y-Achsenabschnitten der j-ten Parabel für das j-te Prüfglied mit der Regressionsfunktion $\hat{y}_{ij} = \hat{\alpha}_j + \hat{\beta}_1 x_{ij} + \hat{\beta}_2 x_{ij}^2$ entsprechen. Da $\mathbf{x} = 0$ aber oft keine sinnvolle Vergleichsbasis darstellt, ist eine Betrachtung bei x.. besser. Den entsprechenden Funktionswert auf der Parabel $y_{j\,korr.}$ (siehe Abb. 1) erhält man, indem in Gleichung 3 (für Variante A) für x_{ij} der Wert x.. eingesetzt wird. Werden die Funktionswerte an der Stelle x.. für Gleichung 4 und 5 berechnet, ergibt sich $y_{j\,korr.} = \hat{\alpha}_{j(B)}$ bzw. $\hat{\alpha}_{j(C)}$, d. h., die $\hat{\alpha}_{j(B)}$ und $\hat{\alpha}_{j(C)}$ entsprechen schon den gewünschten Schätzwerten. Im Beispiel ist

$$y_{1\,korr.} = 73,\ y_{2\,korr.} = 66,\ y_{3\,korr.} = 70.$$

Bei Variante D ist $\hat{\alpha}_{j(D)}$ noch nicht der gewünschte Funktionswert im zweidimensionalen Sinne. Hier stellen die y-Achsenabschnitte $\hat{\alpha}_{j(D)}$ die im dreidimensionalen korrigierten Werte dar.

6. Die mittlere Regression

Bei Betrachtung der in Abbildung 1 dick gezeichneten mittleren Regressionsparabel wird deutlich, daß bei der Schätzung der Koeffizienten auch die Differenzen der Prüfglieder in der x-Variablen mitberücksichtigt werden. Im linearen Fall genügt es, die mittlere Regression aus den Abweichungen der einzelnen x_{ij}- und y_{ij}-Werte von ihren Prüfgliedmitteln zu berechnen. Im vorliegenden quadratischen Ansatz darf nun bei Anwendung der in den zitierten Lehrbüchern (vgl. Abschnitt 1) beschriebenen Methode nicht $(x_{ij} - x._j)^2$ als Argument des quadratischen Gliedes benutzt werden, weil sich dann ein – in Abbildung 2 gestrichelt skizziertes – Polynom mit einer anderen Bedeutung ergibt. In die Schätzung der Regressionskoeffizienten für dieses mittlere Polynom gehen die Prüfglieddifferenzen in der x-Variablen nicht ein, weshalb die Schätzwerte nicht für eine Korrektur der Prüfgliedmittel genutzt werden dürfen.

Wie im Beispiel und der Abbildung 1 angedeutet, ergeben sich in der Praxis, wenn für jedes Prüfglied eine quadratische Regression – die prüfgliedspezifische Regression – bestimmt wird, nicht numerisch gleiche Schätzwerte für die Regressionskoeffizienten. Dies tritt, u. a. bedingt durch Versuchs- und Stichprobenfehler, auch dann auf, wenn die Parameter β_1 und β_2 für jedes Prüfglied gleich sind. Die in Abbildung 1 gepunktet dargestellten prüfgliedspezifischen Kurvenäste verlaufen dann etwas steiler ($j = 3$) oder flacher ($j = 2$) als die mittlere Regression. Die prüfgliedspezifischen Polynome unterscheiden sich in den Abbildungen 1 und 2 durch ihre Lage bezüglich x. Während in Abbildung 1 alle Polynome eine ähnliche Form haben und die mittlere Parabel diese über die gesamte Variationsbreite von x repräsentiert, haben in Abbildung 2 die prüfgliedspezifischen Polynome recht unterschiedliche Formen, was durch die unzulässige Verschiebung der Polynome in Richtung der x-Achse verursacht wird. Das Übergehen der Prüfglieddifferenzen in x ist im linearen Fall zulässig, weil die Steigung einer Geraden von x unabhängig ist, im quadratischen jedoch nicht.

Die für die Korrektur und die Angabe einer mittleren Beziehung geeignete Regressionsfunktion ist die in Abbil-

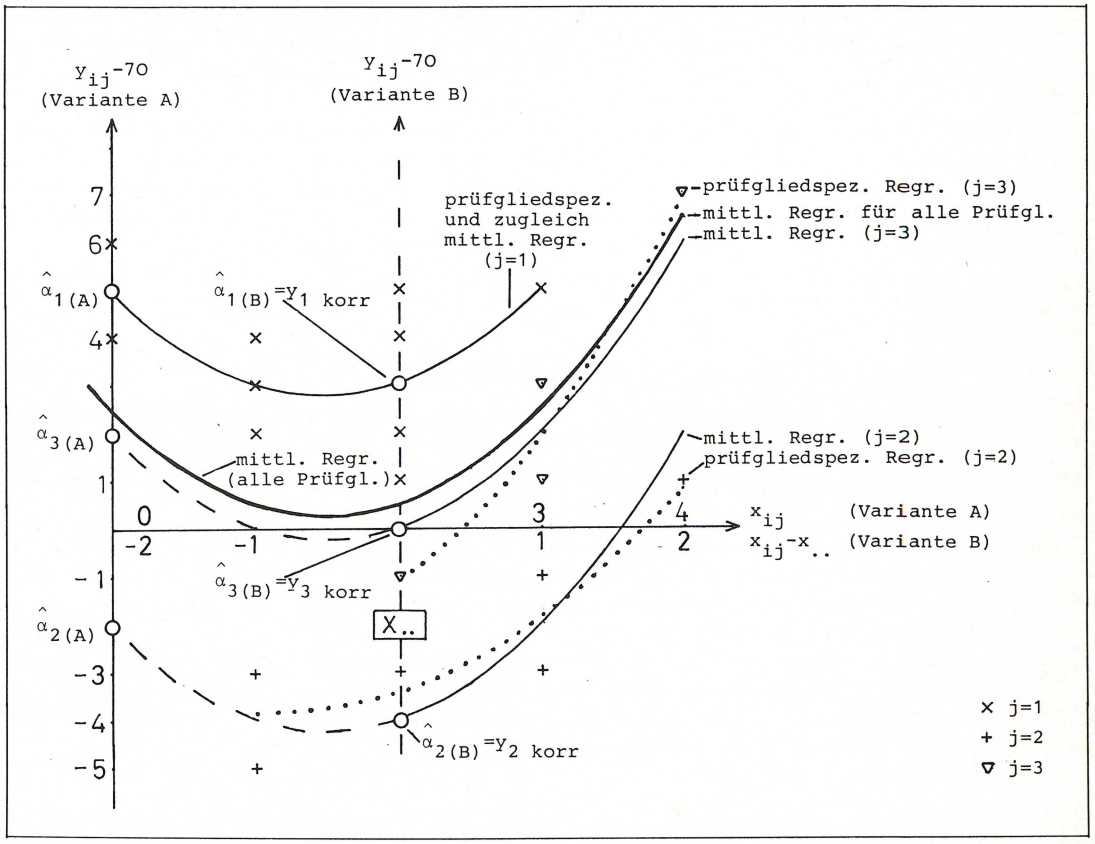


Abb. 1. Graphik zum Zahlenbeispiel.

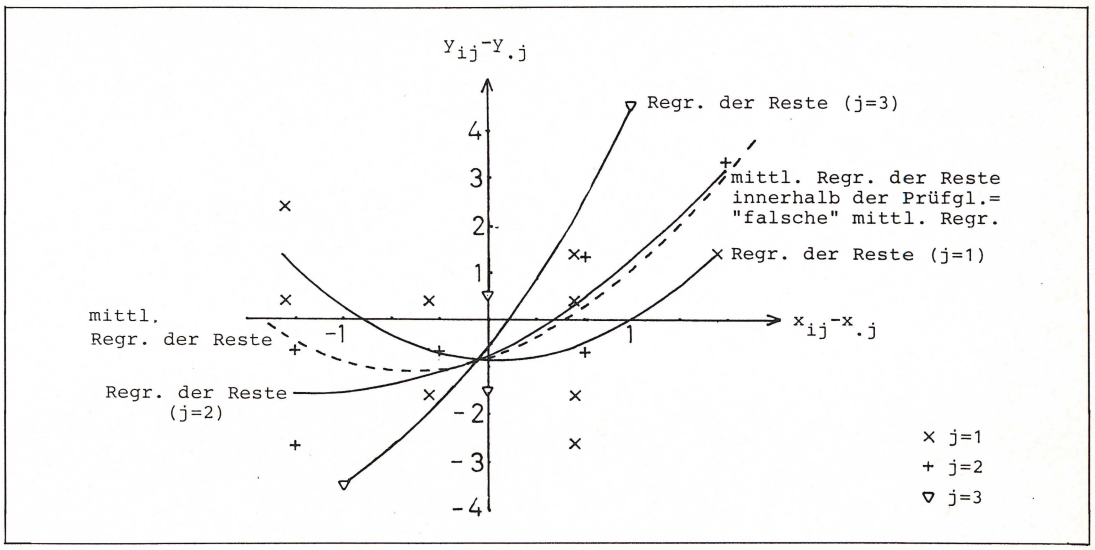


Abb. 2. Graphische Darstellung der zur Schätzung der mittleren Regression ungeeigneten Berechnungsweise.

dung 1 skizzierte Form. Es handelt sich um eine Parabel, die durch die einzelnen Prüfgliedeffekte parallel zur y-Achse, aber nicht zur x-Achse verschoben wird. Um zu dieser richtigen Form einer mittleren Regression zu gelangen, sind im quadratischen Fall die Abweichungen $(x_{ij2} - x_{.j2})$ mit zu berücksichtigen, wenn die Berechnung nach der in den oben genannten Lehrbüchern beschriebenen Form erfolgt. In diesen sind gewissermaßen auch die Differenzen zwischen den Prüfgliedmitteln von x mit enthalten, sie gehen also auf diesem Wege mit in die Analyse ein. Bei Verwendung der FITTING-CONSTANTS-Methode müssen als Schätzwerte für die Regressionskoeffizienten der mittleren Regression die sich – für alle 4 Varianten – aus Gleichung (2) ergebenden Werte herangezogen werden.

Das konstante Glied ist wie folgt zu schätzen:

$$\hat{\alpha}_{.} = y_{..} - \hat{\beta}_1 x_{..1} - \hat{\beta}_2 x_{..2} \qquad (10)$$

$\hat{\alpha}_{.}$ läßt sich für jede einzelne Variante auch aus

$$\hat{\alpha}_{.} = \sum_j n_j \hat{\alpha}_j / n_t$$

berechnen. Somit liegt die mittlere Regressionsparabel genau in der Mitte zwischen den einzelnen, in Richtung der y-Achse verschobenen Parabelästen für die Prüfglieder. Die mittlere Regression stellt also eine Regression der um die Prüfgliedeffekte auf y bereinigten Werte dar.

Der »Prüfgliedeffekt« kann für jede der 4 Varianten aus $\hat{\alpha}_j - \hat{\alpha}_{.}$ bestimmt werden.

Die quadratische Kovarianzanalyse arbeitet graphisch gedeutet also folgendermaßen: Um die Beziehungen im j-ten Prüfglied mittels der mittleren Regression zu beschreiben, wird die mittlere Regressionsparabel so lange parallel zur y-Achse verschoben, bis sie optimal, d. h. mit geringster Abweichungsquadratsumme im Meßpunkteschwarm des j-ten Prüfgliedes liegt. Die $y_{j\,korr.}$ sind die Funktionswerte dieser mittleren Regressionsparabeln für die J-Prüfglieder an der Stelle x...

7. Diskussion

Zwischen linearer und quadratischer Kovarianzanalyse besteht folgender nur scheinbarer Unterschied:
Bei der linearen Kovarianzanalyse wird allgemein von korrigierten »Mittelwerten« gesprochen. Im quadratischen Fall jedoch liegen die Prüfgliedmittel nicht auf der Regressionsfunktion, sondern etwa im Fall $\hat{\beta}_2 < 0$ darüber. Was der Anwender der Kovarianzanalyse im quadratischen Fall aber wissen möchte, ist die Größe von y an der Stelle x.., also gilt dem Funktionswert an der Stelle x.. das Interesse und nicht einem bei der aktuellen Fragestellung eventuell völlig irrelevanten korrigierten »Mittelwert«. Dies ist auch im linearen Fall so, nur entspricht hier der Funktionswert an der Stelle $x_{.j}$ zugleich dem Prüfgliedmittel $y_{.j}$. Im quadratischen Fall aber sollte, um Verwirrungen vorzubeugen, besser von korrigierten »Funktionswerten« gesprochen werden.

Das an sich naheliegende und auch in der Empfehlung von SNEDECOR und COCHRAN (1967) enthaltene Verfahren, nämlich die dreidimensionalen Mittelwerte gemäß der Formel

$$y'_{.j} = y_{.j} - \hat{\beta}_1 (x_{.j1} - x_{..1}) - \hat{\beta}_2 (x_{.j2} - x_{..2}) \qquad (11)$$

zu korrigieren, ist im quadratischen Fall nicht möglich, da mit der Festlegung auf $x_{..1}$ der Wert $x_{..2}$ nicht mehr frei gewählt werden kann. Vielmehr ist für die Varianten A, B, C in Formel (11) $x_{..2}$ durch $(x_{..1})^2$ zu ersetzen, um zu den $y_{j\,korr.}$ zu gelangen. Für Variante D gilt:

$$y_{j\,korr.} = \hat{\alpha}_{j\,(D)} + \hat{\beta}_{2\,(D)} ((x..)^2 - (x^2)..) \qquad (12)$$

Dem Anwender der Kovarianzanalyse sei die Schätzung der Parameter der Modellvariante A unter Anwendung der hier dargestellten Methode der Koeffizientenanpassung empfohlen, da bei Variante A keine Transformation der x-Werte notwendig ist, was u. a. die Interpretation erleichtert. Tests zu Hypothesen bezüglich der Parameter sind speziell für die Kovarianzanalyse u. a. von BEYERBACH (1983) und allgemein für linear-additive Modelle u. a. von SEARLE (1971) beschrieben worden.

Danksagung

Für die Durchsicht und Diskussion des Manuskriptes danke ich Herrn Prof. Dr. H. Rundfeldt und Herrn Privatdozent Dr. W. E. Weber.

Literatur

BEYERBACH, M. (1983): Einsatzmöglichkeiten kovarianzanalytischer Verfahren bei der Auswertung von Erhebungen. Dissertation, Fachbereich Gartenbau der Universität Hannover.

KIRK, R. E. (1968): Experimental Design: Procedures for the Behaviorial Sciences. Brooks/Cole Publishing Company, Belmont, California.

SCHEFFÉ, H. (1959): The analysis of variance. John Wiley & Sons, Inc., New York, London.

SEARLE, S. R. (1971): Linear Models. John Wiley & Sons, New York.

SNEDECOR, G. W. and W. G. COCHRAN (1967): Statistical Methods. Sixth edition, Iowa-State University Press, Ames, Iowa, USA.

STEEL, R. G. D. and J. H. TORRIE (1960): Principles and Procedures of Statistics. McGraw Hill Book Company, Inc., New York.

WINER, B. J. (1971): Statistical Principles in Experimental Design. Second Edition, McGraw Hill, Kogakusha, Ltd.

Eingegangen am 6. November 1984.
Anschrift des Verfassers: Dr. M. Beyerbach, Institut für Statistik und Biometrie der Tierärztlichen Hochschule Hannover, Bischofsholer Damm 15, D-3000 Hannover 1.

EDV in Medizin und Biologie **16** (2), 65–73, ISSN 0300-8282

Numerische Gruppierung und graphische Darstellung von Daten: Ein Methodenvergleich

G. Ohmayer, H. Seiler

Zusammenfassung

Die Fett-, Eiweiß-, Lactose- und Aschegehalte der Milchen von 50 Lebewesen dienten als Daten für eine vergleichende Beschreibung der Methoden Hauptkomponentenanalyse, Biplot, nichtlineare Abbildung, Vernetzungsdiagramm, Dendrogramm, Austauschverfahren und Bestimmung unscharfer Gruppen. Es wird gezeigt, wie graphische Verfahren sowohl zur Aufbereitung von Gruppierungsergebnissen als auch zur optischen Wiedergabe der wesentlichsten Datenstrukturen herangezogen werden können. Möglichkeiten zur Abschätzung der Güte einer Gruppierung werden diskutiert.

Summary

Composition of the milk constituents fat, protein, lactose and ash from 50 mammals were used as data for comparing principal component analysis, biplot, non-linear mapping, linkage-maps, dendrograms, relocation-techniques and fuzzy partitions. The usefulness of applying graphical methods is demonstrated not only for giving a visual representation of the most important aspects of the data, but also for determining group structures within the data. The possibility of estimating the validity of a particular grouping is discussed.

1. Einleitung

Die Vielfalt der heute bekannten Gruppierungsmethoden, die unter den Begriffen «Clusteranalyse», «automatische Klassifikation» oder «numerische Taxonomie» geführt werden, ist kaum noch zu überblicken. Ständig werden neue Verfahren, zumeist Modifikationen der bekannten Methoden, vorgestellt. In der vorliegenden Arbeit soll deshalb ein Überblick über einige grundsätzliche Gruppierungsmethoden und deren wesentlichste Charakteristika gegeben werden. Auf eine ausführliche Darstellung aller mathematischen Algorithmen wird verzichtet. Vielmehr sollen anhand eines vom Umfang her einfachen, strukturell aber hinreichend komplexen Beispiels die Möglichkeiten und Probleme bei der Anwendung der Methoden in den Vordergrund gestellt werden. Besondere Berücksichtigung erfährt die graphische Darstellung der ermittelten Strukturen.

2. Beschreibung der Daten

Als Demonstrationsbeispiel wurde ein bewußt einfaches, allgemeinverständliches Datenmaterial verwendet (Tabelle 1, S. 66). Die zu gruppierenden Objekte sind Milchen von Land- und Meeressäugetieren, Primaten sowie Homo sapiens. Als Gruppierungskriterien dienten die vier charakteristischen Merkmale Fett, Gesamteiweiß, Lactose und Asche der Muttermilchen. Es sei betont, daß sich alle Resultate hinsichtlich der Ähnlichkeitsbeziehungen dieser Lebewesen ausschließlich auf deren Milchzusammensetzung beziehen. Phylogenetische Verwandtschaftsbeziehungen können nur mit äußerster Vorsicht abgeleitet werden, da die Zusammensetzung der Milch keine konservative Eigenschaft ist.

Die Merkmalswerte sind als «weiche Daten» zu betrachten. Zum einen sind sie Datenerhebungen mehrerer Autoren entnommen, was sie nur beschränkt vergleichbar macht (JENNESS, 1974). Außerdem geben sie, da es sich um durchschnittliche Werte handelt, nicht die mit Sicherheit vorhandene unterschiedliche Varianz einzelner Populationen wieder. Erschwerend kommt hinzu, daß die Milchzusammensetzung in Abhängigkeit von Ernährung, Jahreszeit, Alter der Tiere, Wurfgröße, Lactationszyklus usw. erheblichen Schwankungen unterworfen ist. Trotz dieser Einschränkungen eignen sich die Daten sehr gut für den durchgeführten Methodenvergleich. Bei der sachlichen Interpretation der Ergebnisse soll jedoch die Unschärfe der Ausgangsdaten Berücksichtigung finden.

3. Abbildungsverfahren

Die in der englischsprachigen Literatur unter den Namen «Mapping-» bzw. «Ordination-methods» bekannten Verfahren bilden multivariate Daten in einem zweidimensionalen Diagramm ab (EVERITT, 1978). Ziel dieser Verfahren ist es, die einzelnen Beobachtungen so auf Punkte in der Ebene zu verteilen, daß deren tatsächliche Struktur im mehrdimensionalen Raum möglichst gut wiedergegeben wird. Ist die Abbildungsgüte hinreichend, d. h., hält sich der Informationsverlust durch die Datenreduktion in Grenzen, sind solche Abbildungen eine gute Ergänzung zu den rechnerischen Ergebnissen einer Clusteranalyse, da eventuell vorhandene Gruppenstrukturen leicht zu erkennen sind. Man unterscheidet lineare Abbildungsverfahren oder Projektionstechniken, wie zum Beispiel Hauptkomponentenanalyse und Biplot, und nichtlineare Abbildungen (Nonlinear mapping).

3.1 Hauptkomponentenanlayse und Biplot

Ziel der Hauptkomponentenanalyse ist die Bestimmung der wichtigsten Richtungen im Datenraum. Dazu werden die

Tabelle 1. Daten des Demonstrationsbeispiels: Prozentuale Milchzusammensetzung verschiedener Lebewesen

Lebewesen	Fett	Eiweiß	Lactose	Asche
Herrentiere				
Menschen				
1 Mensch	3,8	1,0	7,0	0,2
Menschenaffen				
2 Orang Utan	3,5	1,5	6,0	0,2
3 Schimpanse	3,7	1,2	7,0	0,2
Meerkatzenartige				
4 Zwergmeerkatze	2,9	2,1	7,2	0,3
5 Pavian	5,0	1,6	7,3	0,3
Krallenaffen				
6 Tamarin	3,1	3,8	5,8	0,4
Unpaarhufer				
Pferdeartige				
7 Esel	1,4	2,0	7,4	0,5
8 Hauspferd	1,9	2,5	6,2	0,5
9 Wildpferd	2,2	2,0	6,1	0,4
10 Zebra	2,1	2,3	8,3	0,4
Paarhufer				
Schweine				
11 Wildschwein	6,8	4,8	5,5	1,7
Kamele				
12 Lama	2,4	7,3	6,0	0,5
13 Kamel	5,4	3,9	5,1	0,7
14 Dromedar	4,5	3,6	5,0	0,7
Hirsche				
15 Sikahirsch	19,0	12,4	3,4	1,4
16 Rothirsch	19,7	10,6	2,6	1,4
17 Ren	20,0	9,5	2,6	1,4
Hornträger, Unterfamilie Gazellenartige				
18 Edmigazelle	19,0	12,4	3,3	1,5
19 Thompson Gazelle	19,6	10,5	2,7	1,4
20 Schwarzfersenantilope	20,4	10,8	2,4	1,4
Hornträger, Unterfamilie Rinder				
21 Hausrind	3,7	3,4	4,8	0,7
22 Zebu	4,7	3,2	4,9	0,7
23 Yak	6,5	5,8	4,6	0,9
24 Wasserbüffel	7,4	3,8	4,8	0,8
25 Bison	3,5	4,5	5,1	0,8
Hornträger, Unterfamilie Ziegenartige				
26 Moschusochse	5,4	5,3	4,1	1,1
27 Hausziege	4,5	2,9	4,1	0,8
28 Hausschaf	7,4	5,5	4,8	1,0
Fleischfresser				
Hundeartige				
29 Haushund	12,9	7,9	3,1	1,2
30 Wolf	9,6	9,2	3,4	1,2
31 Kojote	10,7	9,9	3,0	0,9
32 Schakal	10,5	10,0	3,0	1,2
33 Afrik. Wildhund	9,5	9,3	3,5	1,3
Großbären				
34 Schwarzbär	24,5	14,5	0,4	1,8
35 Grizzly Bär	22,3	11,1	0,6	1,5
36 Braunbär	22,6	7,9	2,1	1,4
37 Eisbär	33,1	10,9	0,3	1,4
Ohrenrobben				
38 Nördlicher Seebär	53,3	8,9	0,1	0,5
Nagetiere				
Biber				
39 Biber	11,7	8,1	2,6	1,1
Wühler				
40 Goldhamster	4,9	9,4	4,9	1,4
Mäuse				
41 Wanderratte	10,3	8,4	2,6	1,3
42 Hausmaus	13,1	9,0	3,0	1,3
Hasentiere				
Hasenartige				
43 Hauskaninchen	18,3	13,9	2,1	1,8
44 Florida-Waldkaninchen	13,9	23,7	1,7	1,5
45 Wildkaninchen	17,9	12,5	1,0	2,0
Waltiere				
Furchenwale				
46 Blauwal	42,3	10,9	1,3	1,4
47 Finnwal	32,4	17,8	0,3	1,0
48 Buckelwal	33,0	12,5	1,1	1,6
Delphine				
49 Delphin	33,0	6,8	1,1	0,7
Insektenfresser				
Igel				
50 Braunbrustigel	10,1	7,2	2,0	2,3

ursprünglichen Merkmale $x_1,\ldots,x_p$ ersetzt durch die Hauptkomponenten $y_1,\ldots,y_r$ $(r \leq p)$, die folgende Eigenschaften besitzen:

a) Die Hauptkomponenten sind Linearkombinationen:

$$y_j = \sum_{k=1}^{p} a_{jk} x_k \quad (j = 1,\ldots,r)$$

b) Sie sind unkorreliert: $r_{y_j y_{j'}} = 0$ für $j \neq j'$

c) Sie stellen die Richtungen maximaler Varianzerklärung dar; d. h. in Richtung von y_1 variieren die Daten insgesamt am stärksten, y_2 erklärt im zu y_1 senkrechten Unterraum maximale Varianz, usw.

Dies bedeutet, daß die Projektion der Beobachtungen in die durch die beiden ersten Hauptkomponenten aufgespannte Ebene im „Kleinst-Quadrate-Sinn" die bestmögliche lineare Abbildung mit planarer Darstellung liefert. Die Berechnung der Hauptkomponenten erfolgt über die Bestimmung der Eigenwerte und Eigenvektoren der Varianz-/Kovarianzmatrix bzw. der Korrelationsmatrix. Während die Eigenvektoren die Koeffizienten der Linearkombination liefern, d. h. die Richtung der Hauptkomponenten bestimmen, kann an den Eigenwerten λ_j die Höhe der Varianzerklärung einzelner Hauptkomponenten abgelesen werden.

Für die Approximationsgüte der zweidimensionalen Darstellung gilt:

$$F = (\lambda_1 + \lambda_2) / \sum_{j=1}^{r} \lambda_j .$$

Im Vergleich zur Hauptkomponentenanalyse werden mit der Biplot-Methode außer den Beobachtungen auch die Merkmale sowie deren Beziehungen zueinander graphisch dargestellt (GABRIEL, 1971). Dieses Verfahren beruht auf folgender Zerlegung der Datenmatrix $X_{n,p}$, welche bedeutet, daß sich jeder Beobachtungswert x_{ik} (ite Beobachtung am kten Merkmal) als Skalarprodukt zwischen g_i (iter Zeilenvektor von G) und h_k (kter Zeilenvektor von H) darstellen läßt:

$$X_{n,p} = G_{n,r} \cdot H^{T}_{p,r} \Leftrightarrow x_{ik} = g_i' \cdot h_k \qquad \begin{pmatrix} i = 1,\ldots,n \\ k = 1,\ldots,p \end{pmatrix}$$

Falls nun r – der sog. Rang der Datenmatrix X – gleich 2 ist, lassen sich die Daten ohne Informationsverlust in einem zweidimensionalen Diagramm durch die n + p Vektoren g_i und h_k als Biplot darstellen. Andernfalls ($r > 2$), und dies dürfte in praktischen Anwendungen die Regel sein, wird X durch eine Matrix $\tilde{X}$, die den Rang 2 besitzt, approximiert und deren Biplot gezeichnet. Dabei kann wiederum ein Maß für die Approximationsgüte berechnet und damit die Relevanz des erzeugten Biplots abgeschätzt werden.

Hinsichtlich der Berechnung der oben erwähnten Zerlegung, die über eine SVD (Singular value decomposition) durchgeführt werden kann, sei auf spezielle Literatur verwiesen (GABRIEL, 1971; GABRIEL et al., 1976).

Zur Interpretation eines Biplots sind noch folgende Punkte zu beachten:

a) die Länge der Vektoren h_k entsprechen den Varianzen ($\| h_k \|^2 \sim s_k^2$), die Winkel zwischen diesen Vektoren den Korrelationen zwischen den Merkmalen ($r_{kk'} \sim \cos \sphericalangle h_k h_{k'}$).

b) Die Mahalanobis-Distanzen zwischen den Beobachtungen (wirkliche Abstände unter Berücksichtigung von Korrelationen) werden approximiert durch die Distanzen der Punkte im Biplot.

Folgende kritische Anmerkung zur Anwendung der Hauptkomponentenanalyse und Biplotmethode im Rahmen von

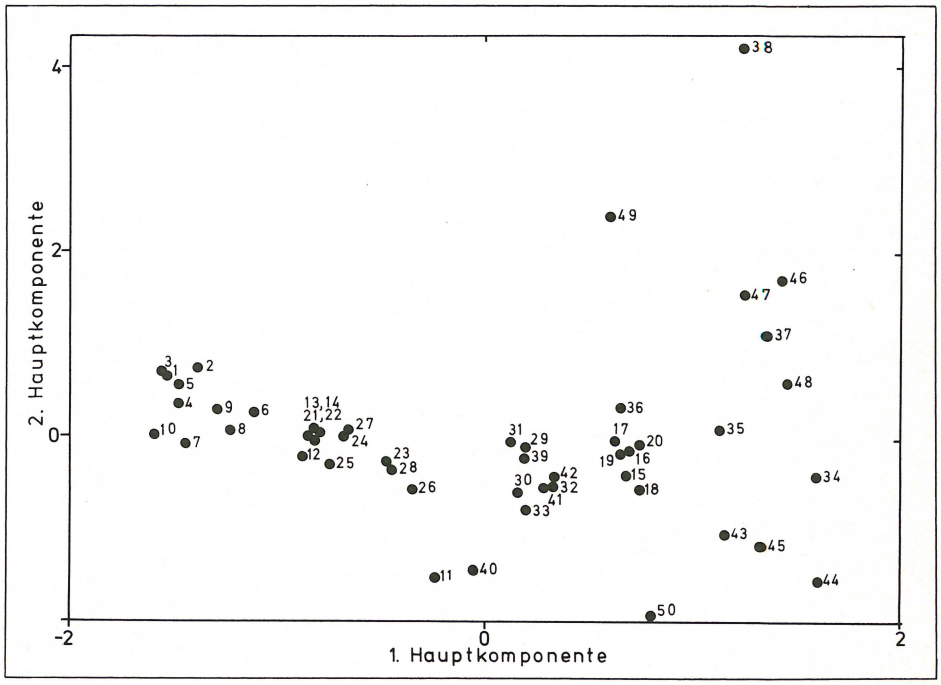


Abb. 1. Abbildung der Ähnlichkeitsbeziehungen von Milchen verschiedener Lebewesen mittels Hauptkomponentenanalyse (Approximationsgüte 92 %).

Gruppierungsproblemen sollte berücksichtigt werden: Diese Methoden wurden für den Fall entwickelt, daß homogene Daten im Sinn einer Stichprobe aus nur einer Population vorliegen. Genau dies kann jedoch bei Gruppierungsproblemen nicht unterstellt werden, denn der Anwender vermutet ja gerade die Herkunft der Daten aus verschiedenartigen, allerdings unbekannten Teilpopulationen. Dieser Effekt wird sich im Einzelfall um so negativer auswirken, je mehr die aus allen Daten geschätzte Varianz-/Kovarianzmatrix S – wichtigster Ausgangsparameter der Methoden – von den Varianz-/Kovarianzmatrizen Σ_l (l = 1,....m) der m Populationen abweicht. Falls Gleichheit der Matrizen Σ_l unterstellt werden kann ($\Sigma_1 = \Sigma_2 .... \Sigma_m = \Sigma$), müßte eine Schätzung von Σ in den Verfahren verwendet werden; diese Schätzung könnte die sog. „Varianz-/Kovarianzmatrix innerhalb der Gruppen" sein, die nach einer Gruppierung der Daten bestimmt werden kann.

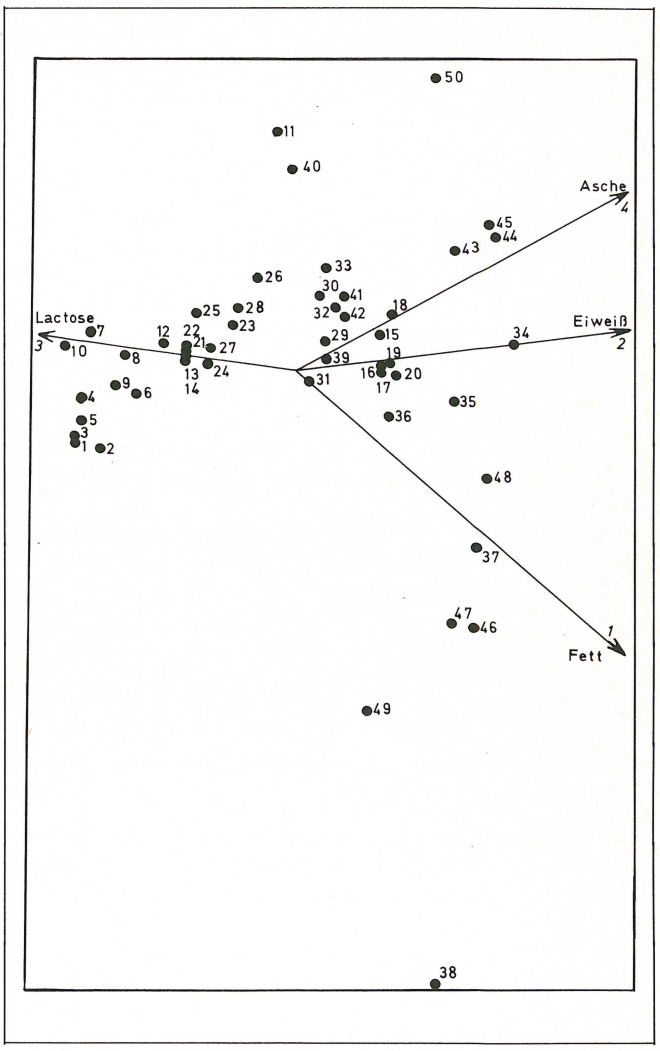
Die Abbildungen 1 und 2 zeigen für unser Anwendungsbeispiel die Ergebnisse der Hauptkomponentenanalyse und des Biplotverfahrens.

Auffallend ist, daß bezüglich der Position der Lebewesen in den beiden Diagrammen kaum Unterschiede bestehen. Im Biplot – hier allerdings berechnet an den auf die Merkmalsvarianzen 1 standardisierten Daten – sieht man aber noch zusätzlich die Wirkung der Merkmale. Es wird deutlich, daß Trokkensubstanz und Fett sowie Gesamteiweiß und Asche positiv korreliert sind (Winkel klein), während das Merkmal Lactose zu den erstgenannten eine negative Korrelation aufweist (Winkel zwischen 90° und 180°).

3.2 Nichtlineare Abbildung und Vernetzungsdiagramm

SAMMON beschrieb 1969 die „Nonlinear mapping" Methode. Das Verfahren setzt Schätzwerte d_{ij} für die Unähnlichkeit zweier Beobachtungen i und j voraus. Welches aus der Palette verfügbarer Distanzmaße adäquat ist, hängt vom aktuellen Anwendungsfall ab (BOCK, 1974). Ziel der nichtlinearen Abbildung ist es, alle Beobachtungen in einem zweidimensionalen Diagramm so zu positionieren, daß deren euklidische Abstände im Diagramm d_{ij}^* möglichst gut mit den Distanzen d_{ij} übereinstimmen. Das bedeutet, daß die Abstandsverhält-

Abb. 2. Abbildung der Ähnlichkeitsbeziehungen mit der Biplot-Methode (Darstellungsgüte = 89,2 %).



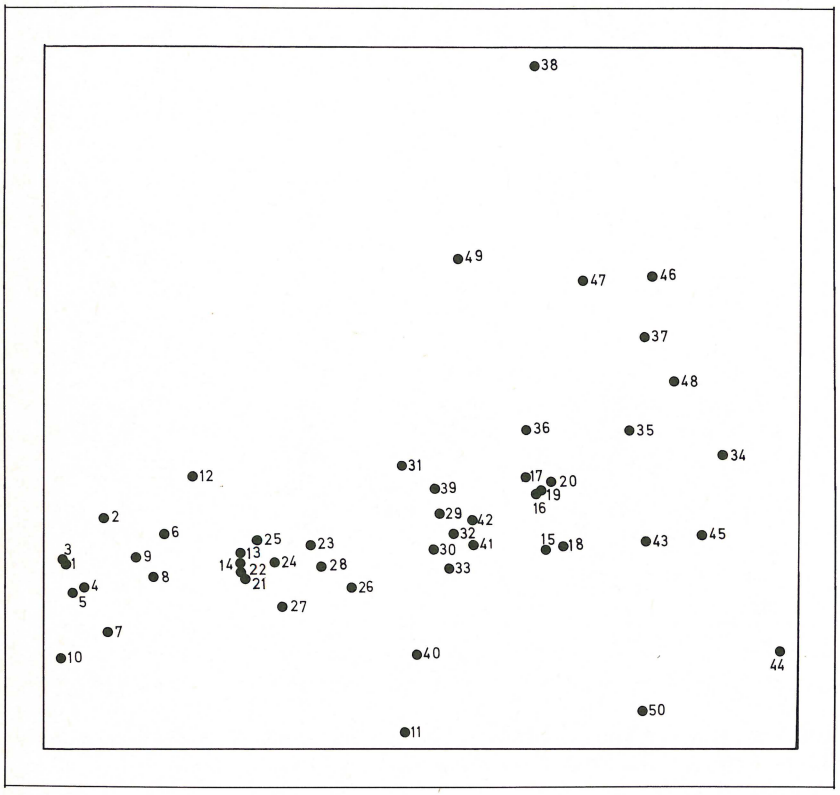


Abb. 3. Darstellung der Milchen mit Hilfe der «nichtlinearen Abbildung» (Darstellungsfehler = 5,3 %).

nisse bestmöglich erhalten werden sollen. Man erreicht dies durch Minimierung des Faktors

$$E = \frac{1}{\sum_{i<j}^{n} d_{ij}} \cdot \sum_{i<j}^{n} \frac{(d_{ij} - d^*_{ij})^2}{d_{ij}}$$

Diese Minimierung wird in einem Iterationsverfahren mit folgender Grundidee erzielt. Ausgehend von einer im allgemeinen beliebigen Anfangskonfiguration der Beobachtungen in der Ebene werden schrittweise neue, bessere Konfigurationen gebildet, indem die Beobachtungen in Richtung des steilsten Abstiegs der Funktion E geringfügig verschoben werden. Ein solches Gradientenverfahren garantiert selbstverständlich nicht, daß das nach vielen Iterationen erreichte Minimum globaler Art ist. Es kann daher empfohlen werden, den Minimierungsprozess mit verschiedenen Anfangskonfigurationen zu starten.

Zur Beurteilung der Güte einer Endkonfiguration ist folgender mittlerer Darstellungsfehler $\bar{e}$ besser geeignet als der zu minimierende Fehler E

$$\bar{e} = \sum_{i<j}^{n} e_{ij}/N \text{ wobei } e_{ij} = \frac{|d_{ij} - d^*_{ij}|}{d_{ij}}$$

(relativer Fehler der Darstellung des Beobachtungspaares i, j)

$$N = \frac{n \cdot (n-1)}{2}$$ (Zahl der Distanzen bei n Beobachtungen).

Der mittlere Darstellungsfehler sollte nach Erfahrungen aus umfangreichen empirischen Studien unter 0,2 (20 %) liegen; andernfalls wird auf eine Interpretation des Diagramms besser verzichtet.

Die durch «Nonlinear mapping» gewonnene Darstellung kann modifiziert und in ihrem Aussagewert verbessert werden, indem die höchsten Ähnlichkeitsbeziehungen zwischen den Beobachtungen graphisch eingetragen werden. Dieser Vorschlag geht zurück auf BUSSE (1970) und wurde von OHMAYER et al. (1980) im Zusammenhang mit dem Verfahren «Nonlinear mapping» aufgegriffen. Die Darstellungen wurden Vernetzungsdiagramme oder «Linkage maps» genannt. Voraussetzung ist dabei die Vorgabe einer sinnvollen Schichtung der Distanzen, d. h. die Einteilung der Distanzen in Klassen durch Festlegung sog. Vernetzungsniveaus. Durch Zuordnung einer Linierungsart zu jeder Klasse können die entsprechenden Vernetzungen eingezeichnet werden; Gruppen ähnlicher Beobachtungen werden im «Linkage map» durch hohen Vernetzungsgrad deutlich.

Die Abbildungen 3 und 4 zeigen nichtlineare Abbildung und Vernetzungsdiagramm für die Milchdaten. Der mittlere Darstellungsfehler von 5,3 % ist hinreichend niedrig und bedeutet, daß die Ähnlichkeiten der Lebewesen hinsichtlich der Milchzusammensetzung im Mittel gut durch die Abstände im Diagramm wiedergegeben werden.

4. Gruppierungsverfahren

Die besprochenen Abbildungsverfahren projizieren die Objekte nach verschiedenen Algorithmen in die Ebene, ohne daß eine Entscheidung über Gruppierungen getroffen wird. Die vielfach unter dem Sammelbegriff «Clusteranalyse» bekannten Verfahren gehen einen Schritt weiter und ermitteln als Ergebnis Cluster, d. h. Gruppen von Beobachtungen. Ziel dieser Gruppenbildung ist, daß Beobachtungen innerhalb einer Gruppe möglichst ähnlich (Homogenitätsforderung), Beobachtungen in verschiedenen Gruppen dagegen möglichst unähnlich (Isolationsforderung) sind. Die vielen verfügbaren Methoden (BOCK, 1974; EVERITT, 1974) unterscheiden sich

- durch verschiedene Gewichtung von Homogenität und Isolation
- durch unterschiedliche Gruppierungsstrategien (agglomerative, divisive, iterative, graphentheoretische u. a. Verfahren)

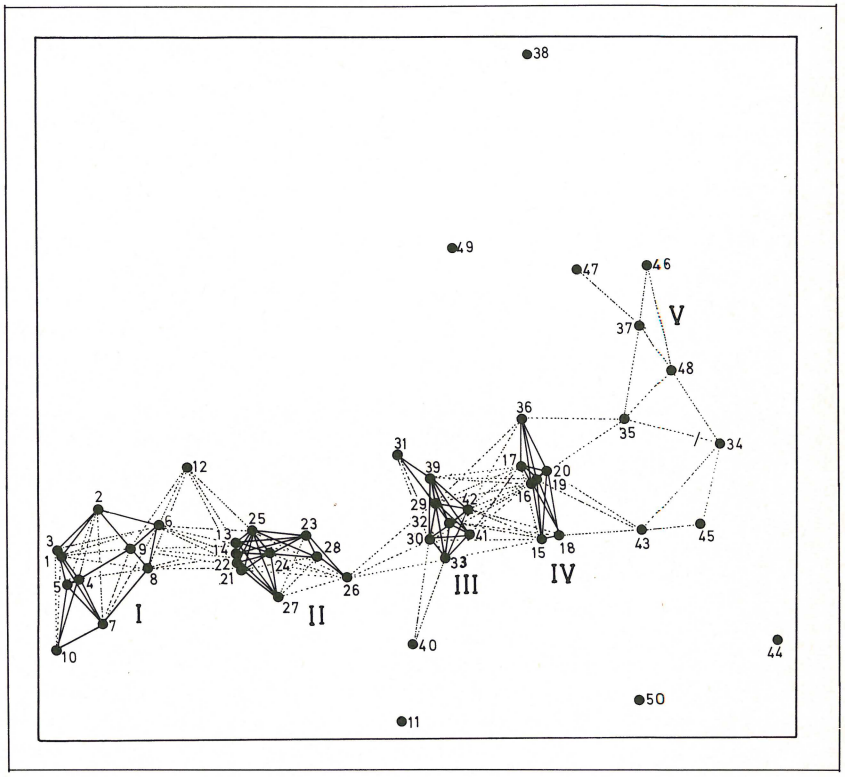


Abb. 4. Vernetzungsdiagramm (Vernetzungsniveaus ——— $0 \leq d_{ij} \leq 0{,}15$; . . . . . $0{,}15 < d_{ij} \leq 0{,}25$).

– durch verschiedene Gruppierungsformen (disjunktive, hierarchische, unscharfe Gruppierungen)
– durch unterschiedliche A-priori-Vorgaben (Gruppenanzahl bekannt, Verteilungsannahmen u. a.).

An dieser Stelle sollen nur drei Typen von Methoden kurz besprochen und auf den Beispielsdatensatz angewendet werden.

4.1 Hierarchische Gruppierungsmethoden

Die hierarchischen Verfahren sind innerhalb der Clusteranalyse die bekanntesten und am häufigsten verwendeten Methoden. Sie liefern als Ergebnis nicht nur eine, sondern eine ganze Hierarchie von Gruppierungen, die graphisch als Dendrogramm dargestellt werden kann. Man unterscheidet aufgrund des verschiedenen Konstruktionsprinzips agglomerative und divisive Verfahren. Während die ersten durch schrittweise Fusion von ähnlichen Beobachtungen bzw. Gruppen immer größere Gruppen bilden, operieren letztere durch schrittweise Spaltung der jeweils inhomogensten Gruppe in umgekehrter Richtung. Die einzelnen agglomerativen Verfahren unterscheiden sich in der Berechnung der Distanzen zwischen Gruppen aus den Distanzen zwischen Beobachtungen in den Gruppen. Die wichtigsten Vertreter dieser Methodenklasse sind die unter den Namen Single-, Complete- und Average-linkage bekannten Verfahren sowie die Ward- und Centroid-Methode. Die divisiven Verfahren sind gegenüber den agglomerativen Methoden von geringerer Bedeutung.

Zur Beurteilung der Güte von Dendrogrammen gibt es verschiedene Maßzahlen, die auf der Berechnung der ultrametrischen Distanzen basieren (OHMAYER, 1982). Da zwischen einem Dendrogramm und der zugeordneten ultrametrischen Distanz eine eindeutige Beziehung besteht, kann beispielsweise deren Korrelation mit den Ausgangsdistanzen als Beurteilungskriterium dienen. Tabelle 2 zeigt diese Korrelationskoeffizienten für die Dendrogramme der Milchdaten. Unter den agglomerativen Verfahren liefert das Average-linkage-Verfahren (Abb. 5) das im beschriebenen Sinne beste Resultat.

Abb. 5. Gruppierung der Milchen mit dem «Average-linkage-Verfahren» (euklidischer Koeffizient, standardisierte Daten).

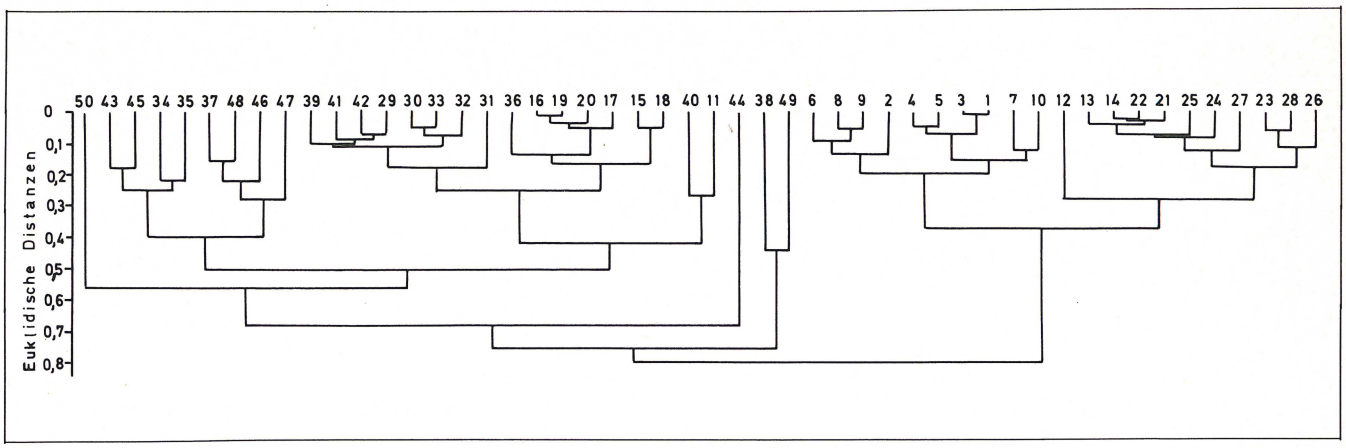


Tabelle 2. Gütewerte verschiedener Dendrogramme.

Verfahren	Güte
Average linkage	0,75
Centroid	0,75
Complete linkage	0,71
Ward	0,67
K-diameter (k = 2)	0,67
Single linkage	0,59

4.2 Austauschverfahren

Unter den Verfahren, die als Ergebnis keine Hierarchie von Gruppierungen, sondern eine einfache Partition der zu klassifizierenden Objekte liefern, sind hauptsächlich iterative Methoden hervorzuheben, die nach dem «Austauschprinzip» arbeiten. Es wird dabei eine Anfangsklassifikation vorgegeben, die zufällig sein kann oder aufgrund von Vernetzungen bzw. als Ergebnis anderer Gruppierungsverfahren entsteht. Damit wird gleichzeitig die Zahl der Gruppen fixiert. Jedes Objekt wird nun iterativ daraufhin überprüft, ob ein Austausch, d. h. Wechsel in eine andere Gruppe, die Gruppierung «verbessert». Das Iterationsverfahren wird abgebrochen, sobald keine «Verbesserung» durch Austausch mehr möglich ist.

Die bekannten Verfahren unterscheiden sich insbesondere in der Definition der zu optimierenden «Güte einer Gruppe», wobei im wesentlichen das Varianz-, Determinanten- und Spur-Kriterium zu erwähnen sind. Die einfachste Möglichkeit ist jedoch die Anwendung der «Minimaldistanzregel», nach der ein Austausch immer dann vorzunehmen ist, wenn ein Objekt zum Mittelpunkt einer anderen Gruppe eine geringere Distanz hat als zur eigenen Gruppe. Unterschiede ergeben sich auch durch die Wahl des Zeitpunktes für die Neuberechnung der Gruppenmittelpunkte; diese erfolgt entweder nach jedem durchgeführten Objekttausch oder nach jedem kompletten Durchlauf.

Eine Schwierigkeit bei der Anwendung solcher Austauschverfahren besteht in der Vorgabe einer Anfangsklassifikation, hauptsächlich aber in der Festlegung der Gruppenzahl, die auch bei Bildung einer zufälligen Ausgangsklassifikation erforderlich ist.

Da in vielen Fällen die «richtige» Gruppenzahl unbekannt ist, bestimmt man üblicherweise die Gruppierungen für zunehmende Klassenzahlen und entscheidet sich mit Hilfe der Werte des Gruppenkriteriums für eine Klassenzahl.

Abbildung 6 zeigt die Gruppierung der Milchen – graphisch dargestellt mit Hilfe der nichtlinearen Abbildung – mit 5 sowie auch 8 Klassen, da sich letztere als nächstbeste Gruppierung hier zufällig durch Aufspaltung von Gruppen ergibt.

4.3 Unscharfe Gruppierungen

Die bisher besprochenen Gruppierungsmethoden geben keine Information darüber, welche Beobachtungen im Zentrum, am Rand oder im Übergangsbereich zwischen Gruppen liegen. Es wird deshalb im folgenden ein unter den Anwendern noch ziemlich unbekanntes Verfahren zur Bestimmung «unscharfer Gruppen» (Fuzzy sets) beschrieben (BOCK, 1979). Durch Minimierung eines Zielkriteriums werden iterativ Gewichte oder Zugehörigkeitsindizes u_{ik} berechnet, die bei der vorgegebenen Gruppenzahl m für jede Beobachtung i die Höhe der Zugehörigkeit zur Gruppe k prozentual angeben. D. h., es muß für die Matrix $U = (u_{ik})$, welche unscharfe Gruppierung oder unscharfe Partition genannt wird, gelten:

$$0 \leq u_{ik} \leq 1 \text{ für } \begin{matrix} i = 1,....n \\ k = 1,....m \end{matrix} \text{ und } \sum_{k=1}^{m} u_{ik} = 1$$

Durch Vorgabe einer Schranke (z. B. $u_{ik} \geq 0{,}7$) können Kernpunkte von Gruppen bzw. Rand- und Übergangsbereiche ermittelt werden. Das üblicherweise verwendete Kriterium, welches minimiert wird, ist:

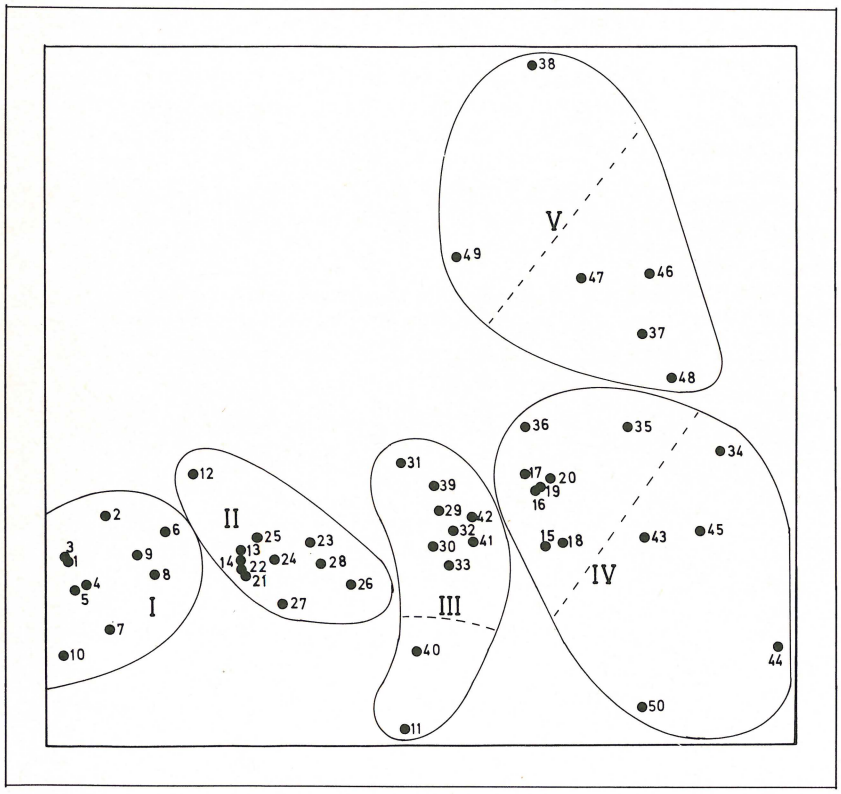
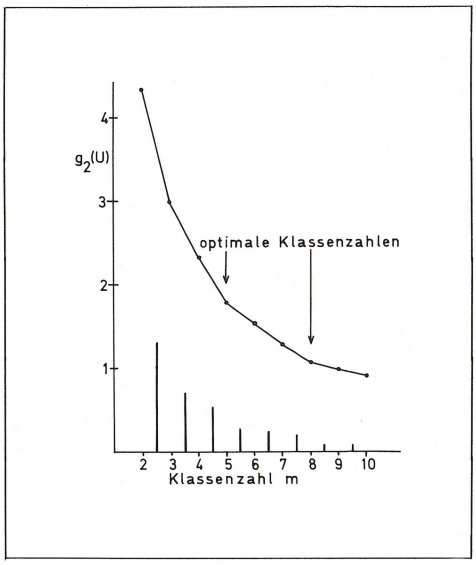
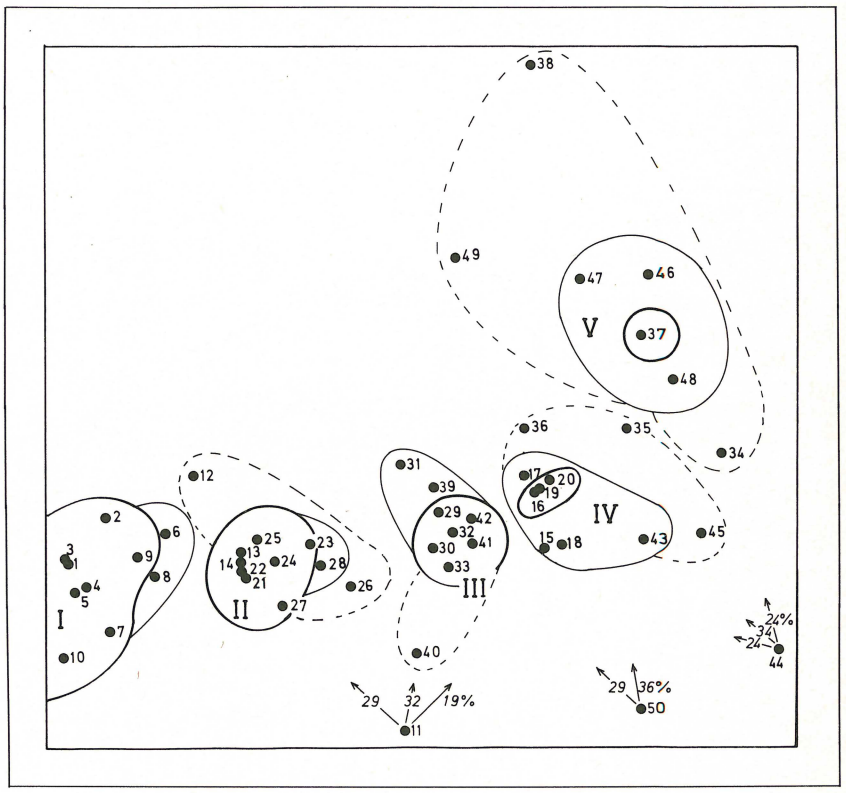


Abb. 6. Gruppierung der Milchen nach dem Austauschverfahren (Klassenzahl 5 und 8).

Abb. 7 (rechts). Darstellung der «unscharfen Gruppierung» für fünf Klassen (Schranken: $u_{ik} \geq 0{,}8$; 0,6; 0,4).

Abb. 8 (unten). Zielfunktionsverlauf in Abhängigkeit der Gruppenzahl bei «unscharfen Gruppierungen».





$$g_r(U) = \sum_{k=1}^{m}\sum_{i=1}^{n} u_{ik}^r \cdot \| x_i - \bar{x}_k \|^2 \text{ mit } \bar{x}_k = \sum_{i=1}^{n} u_{ik}^r x_i / \sum_{i=1}^{n} u_{ik}^r.$$

Dabei kann über den Parameter r die ,,Unschärfe" der Gruppierung gesteuert werden; denn die Wahl von r = 1 führt noch zu einer ,,scharfen" Partition, d. h. zum Ergebnis $u_{ik} = 0$ oder = 1. Erst Werte r > 1 liefern unscharfe Gruppierungen. Empfohlen wird die Verwendung der Werte r = 2 oder r = 3.

Um anzugeben, wie unscharf eine Partition ist, werden das Entropiemaß H(U) bzw. das Fuzzy-Maß F(U) vorgeschlagen:

$$H(U) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} u_{ik} \log u_{ik}/\log m,$$

$$F(U) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} u_{ik}^2$$

Beide Maße liefern Werte im Intervall [0,1], sind allerdings im folgenden Sinn gegenläufig:
H(U) = 0 <=> F(U) = 1 <=> U ist eine gewöhnliche, d. h. scharfe Gruppierung
H(U) = 1 <=> F(U) = 0 <=> U ist die unschärfste Gruppierung mit $u_{ik} = \frac{1}{m}$

H(U) und F(U) können insbesondere in Verbindung mit $g_r(U)$ verwendet werden, um bei unbekannter Gruppenzahl durch Vergleich der unscharfen Gruppierungen für ein Intervall $m^- \leq m \leq m^+$ den optimalen Wert von m zu finden.

Zur graphischen Darstellung der Ergebnisse einer unscharfen Gruppierung kann die im Abschnitt 3.2 beschriebene nichtlineare Abbildung dienen. Durch Einzeichnen der Gruppenkerne und Angabe der Zugehörigkeitsindizes für die übrigen Punkte entsteht eine Abbildung, die die Struktur in der Regel gut und leicht erfaßbar wiedergibt (Abb. 7). Abbildung 8 zeigt den Verlauf des Zielkriteriums $g_r(U)$ (absolute Werte und Veränderungen) für $1 \leq m \leq 10$ und $r = 2$ bei Anwendung des Verfahrens auf die Milchdaten. Die Klassenzahlen 5 und 8 erweisen sich auch hier als optimal, da die Funktion $g_2(U)$ deutliche Knickstellen bei diesen Werten zeigt.

5. Darstellung von Gruppierungsergebnissen

Die meisten Gruppierungsverfahren bilden in jedem Fall, selbst bei homogenem Datenmaterial, Gruppen, was leicht zu Fehlinterpretationen führen kann. Es ist deshalb ratsam, die Gruppierungsergebnisse kritisch zu prüfen. Folgende Maßnahmen bieten sich an, um zufällige und somit falsche Gruppierungen aufzudecken:

a) Einsatz mehrerer, möglichst verschiedener Gruppierungsverfahren und Vergleich der Ergebnisse

b) Prüfung mittels varianz- und diskriminanzanalytischer Methoden

c) Graphische Darstellung der Gruppen-Eigenschaften.

Liegen echte Gruppenstrukturen vor, sind in der Regel die Gruppierungen über die Methodenunterschiede hinweg in den wesentlichsten Punkten stabil. Bereiche unterschiedlicher Gruppierung sollten als Grauzone gesehen und als solche interpretiert werden, wobei das Verfahren «unscharfe Gruppierung» geeignet ist, Gruppenkerne und Randbereiche quantitativ zu erfassen.

Varianz- und Diskriminanzanalyse werden mit dem Ziel eingesetzt, die Ungleichheit der Gruppen zu bestätigen und eventuell einzelne Beobachtungen umzuklassifizieren. Einige Vorbehalte gegen dieses Vorgehen sind jedoch anzumelden. Zum einen fällt eine konfirmatorische Analyse (hier Diskriminanzanalyse) prinzipiell viel zu positiv aus, falls die zu prüfende Hypothese (hier Gruppierung) vorher an demselben Datenmaterial in einer explorativen Analyse (hier Clusteranalyse) gewonnen wurde. Zum anderen stellen Varianz- und Diskriminanzanalyse, zumindest die klassischen parametrischen Verfahren, gewisse Anforderungen an die Daten bezüg-

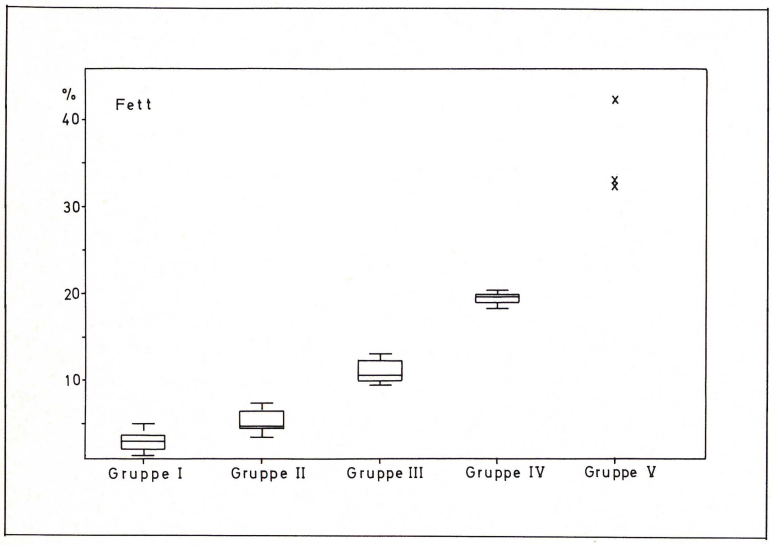


Abb. 9. Darstellung der Gruppeneigenschaften für das Merkmal Fettgehalt mit dem Verfahren «Schematic box-plots».

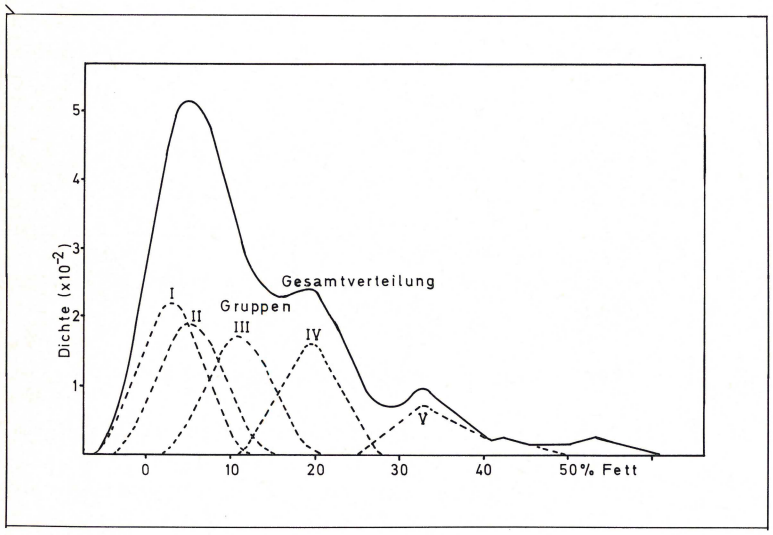


Abb. 10. Darstellung der Gruppeneigenschaften für das Merkmal Fettgehalt mit dem Verfahren «nichtparametrische Dichteschätzung».

lich Normalverteilung, Homoskedastizität etc., was in vielen Fällen sicherlich nicht gegeben ist.

Die graphische Darstellung der Gruppen-Eigenschaften ist in jedem Fall ein sinnvolles Hilfsmittel zur Interpretation der Gruppen und zur Charakterisierung ihrer Unterschiede. Exemplarisch sollen an dieser Stelle zwei Möglichkeiten erläutert werden, die geeignet sind, pro Merkmal die Gruppen-Eigenschaften zu erfassen. Betrachten wir zunächst die «Schematic plots» oder «Box and whisker plots» für die Milchdaten und deren im Abschnitt 4 ermittelte Gruppierung (Abb. 9). Jede Box umfaßt das Intervall 25%- bis 75%-Quantile und kennzeichnet damit den mittleren Datenbereich. Innerhalb jeder Box ist die 50%-Quantile, d. h. der Median, eingezeichnet. Die Definition der Länge der Whiskers (Barthaare) hängt von der gewählten Methode ab (DIETLEIN, 1981). Nach MCNEIL (1977) werden diese Whiskers so festgelegt, daß sie ca. 95 % des Datenbereiches einschließen. Werte, die außerhalb dieser Grenze liegen, werden einzeln geplottet. Eine zweite Form der graphischen Darstellung von Gruppen-Unterschieden ergibt sich durch Schätzung der Wahrscheinlichkeitsverteilung pro Merkmal und Gruppe. Dazu wird ein nichtparametrisches Verfahren gewählt, mit dem über dreiecksverteilte Kerne die Dichte eines Merkmals geschätzt werden kann (VICTOR, 1978). Abbildung 10 zeigt die geschätzten Dichtefunktionen des Merkmals Fettgehalt für die einzelnen Gruppen sowie für alle Daten.

6. Interpretation des Demonstrationsbeispiels

Die untersuchten Säugetiere der Ordnungen Primaten, Unpaarhufer, Paarhufer, Fleischfresser, Nagetiere, Hasentiere, Wale und Insektenfresser zeigen im Vernetzungsdiagramm (Abb. 4) eine differenzierte Gliederung. Die Milchen von Igel (50), Seebär (38), Delphin (49), Wildschwein (11) und Forida-Waldkaninchen (45) liegen relativ isoliert, während die übrigen Tiere zu Gruppen geordnet sind. Primaten und Unpaarhufer bilden zusammen die Gruppe I. Dies läßt auf eine große quantitative Ähnlichkeit der Milchkomponenten schließen. Im Übergangsbereich von I nach II ist die Lama-Milch (12) lokalisiert. Die übrigen Paarhufer trennen sich in zwei Gruppen; erstere umfaßt die Familie Kamele und die Rinderartigen der Unterfamilien Rinder und Ziegen (Gruppe II), zweitere bilden die Familien Hirsche und Antilopen (Gruppe IV). Die freizehigen Fleischfresser gliedern sich in Hunde und Bären. Die Milchen der Hunde ergeben zusammen mit denen der Nagetiere die Gruppe III; die Milchzusammensetzung der Bären tendiert mehr in Richtung Wale (Gruppe V). Letztere Gattung sowie die Hasentiere bilden nur andeutungsweise Vernetzungsschwerpunkte. Beim Vergleich mit Tabelle 1 wird klar, daß in der nichtlinearen Abbildung (Abb. 3), die als Basis für das Vernetzungsdiagramm dient, im allgemeinen die Milchen von links nach rechts mit zunehmender Gesamttrockenmasse geordnet wurden. Die Gliederung

der Objekte in der Hauptkomponentenanalyse und im Biplot (Abb. 1, 2) zeigt gegenüber den Abbildungen 3 und 4 nur geringe Abweichungen. Die Gruppierungen I bis IV sind auch hier relativ gut zu erkennen. Ebenso bilden die Wale und die Hasentiere nur andeutungsweise Gruppierungen. In Abbildung 1 wird das Objekt 12 (Lama-Milch) – ebenso wie beim Austauschverfahren (Abb. 6) – den Paarhufern (Gruppe II) zugeordnet. Während in den Verfahren Hauptkomponentenanalyse, Biplot, nichtlineare Abbildung und zum Teil auch im Vernetzungsdiagramm keine Entscheidung über Gruppierungen getroffen wird, ergeben das Dendrogramm, das Austauschverfahren und die unscharfe Gruppierung (Abb. 5, 6, 7) mehr oder weniger klare Entscheidungshilfen für die Definition von Clustern. Überprüft man jedoch die Güte der gebildeten Gruppen nach den in Abschnitt 5 ausgeführten Kriterien, so wird klar, daß bei den Tiergruppen Wale (46–48), Bären (34–37) und Hasenartige (43–45) eine eindeutige Grenzziehung nicht möglich ist.

Dieses Demonstrationsbeispiel umfaßt einerseits Gruppen von Beobachtungen, welche sich über alle Verfahren hinweg als stabil erweisen, und andererseits Beobachtungen, welche relativ isoliert liegen bzw. als Zwischen- oder Randpunkte zu betrachten sind. Diese Struktur der Daten wird nach Meinung der Autoren am adäquatesten durch die Abbildungen 4 und 7, d. h. durch das Vernetzungsdiagramm und die «unscharfe Gruppierung», beschrieben.

Literatur

BOCK, H. H.: Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen 1974.

BOCK, H. H., 1979: Clusteranalyse mit unscharfen Partitionen. In BOCK, H. H. (Hrsg.): Klassifikation und Erkenntnis III, Studien zur Klassifikation, Bd. 6, Proc. der 3. Fachtagung der Gesellschaft für Klassifikation e.V., Frankfurt, 137–163.

BUSSE, M., 1970: Eine neue Methode zur Untergliederung eng verwandter Bakteriengruppen am Beispiel der Enterobakterien. In DELLWEG, H. (Hrsg.): Zweites Symposium Technische Mikrobiologie, Institut für Gärungsgewerbe und Biotechnologie, Berlin, 243–257.

DIETLEIN, G., 1981: Schematic Plots – eine Alternative zur Darstellung von mittleren Verlaufskurven. Stat. Software Newsletter **7,** 100–103.

EVERITT, B.: Cluster Analysis. Heinemann Educational Books Ltd., London 1974.

EVERITT, B.: Graphical Techniques for Multivariate Data. Heinemann Educational Books Ltd., London 1978.

GABRIEL, K. R., 1971: The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. Biometrika **58,** 453–467.

GABRIEL, K. R., G. RAVE, E. WEBER, 1976: Graphische Darstellung von Matrizen durch das Biplot. EDV in Medizin und Biologie **7,** 1–15.

JENNESS, R., 1974: The Composition of Milk. In LARSON, B. L., V. R. SMITH (Eds.): Lactation. Comprehensive Treatise. Volume III: Nutrition and Biochemistry of Milk, Academic Press, London, 3–107.

MCNEIL, C., 1977: Interactive Data Analysis. John Wiley and Sons, New York.

OHMAYER, G., M. PRECHT, H. SEILER, M. BUSSE, 1980: Linkage-maps and their Relations to Linkage Cluster Procedures. Zbl. Bakt. II. Abt. **135,** 22–37.

OHMAYER, G., 1982: Ein einfaches Wahrscheinlichkeitsmodell «Klassifikation binärer Daten». Dissertation, TU München-Weihenstephan.

SAMMON, J. W., 1969: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers, C-18, 401–409.

VICTOR, N., 1978: Alternativen zum klassischen Histogramm. Meth. Inform. Med. **17,** 120–126.

Eingegangen am 28. März 1985

Anschrift der Verfasser: Dr. G. Ohmayer, Abt. Mathematik und stat. Methodenlehre, Datenverarbeitungsstelle, TU München-Weihenstephan, D-8050 Freising 12
Dr. H. Seiler, Bakteriologisches Institut der Süddeutschen Versuchs- und Forschungsanstalt für Milchwirtschaft, D-8050 Freising 12

BUCHBESPRECHUNGEN

KÖPCKE, W.
Zwischenauswertungen und vorzeitiger Abbruch von Therapiestudien
Gemischte Strategien bei gruppensequentiellen Methoden und Verfahrensvergleiche bei Lebensdauerverteilungen. Reihe Medizinische Informatik und Statistik Nr. 53, Springer-Verlag Berlin – Heidelberg – New York 1984.

Das Buch behandelt statistische Verfahren zur (block-)sequentiellen Auswertung therapievergleichender klinischer Studien. Um eine tatsächlich vorhandene Differenz bestimmter Größe zwischen zwei Vergleichstherapien im Versuch zu entdecken, benötigt man bei einem solchen Vorgehen zwar in der einzelnen Studie maximal größere, bei wiederholter Anwendung in zahlreichen Studien im Mittel jedoch geringere Patientenzahlen.

Nach der einleitenden Beschreibung des Problems werden drei Beispiele von vorzeitig abgebrochenen Therapiestudien diskutiert und allgemeine Überlegungen zu Zwischenauswertungen und vorzeitigem Studienabbruch angestellt. Kapitel 3 enthält eine Übersicht über die vorhandenen Verfahren. Die eingehende Beschreibung des jeweiligen Vorgehens mit Anfügung der notwendigen Tabellen (ausführlicher Tabellenanhang) und zahlreichen Beispielen ermöglicht dem Anwender die konkrete Konstruktion entsprechender Sequentialpläne.

In Kapitel 4 erweitert der Autor die auf dem Prinzip des wiederholten Signifikanztests beruhenden gruppensequentiellen Pläne um eine Form mit stückweise ansteigendem und stückweise konstantem nominellen Einzeltestniveau. Die verschiedenen gruppensequentiellen Pläne werden per Stimulation verglichen. Dies orientiert sich am Bedarf des Anwenders an einer vergrößerten Verfahrensauswahl, um die Eigenschaften des Sequentialplanes den Anforderungen im einzelnen Anwendungsfall besser anpassen zu können. Weitere Entwicklungen in der eingeschlagenen Richtung könnten – zur Ergänzung solcher Pläne mit fixierten Einzelniveaus – die Möglichkeit vorsehen, diese Niveaus aus Anwendungsvorgaben zu bestimmen (etwa aus den Beta-Fehlern zu den einzelnen Zwischenauswertungen).

Ein weiteres Kapitel ist dem simulativen Vergleich von sequentiellen Tests bei zensierten Überlebenszeiten gewidmet. H. Schäfer

ERBS, H.-E. und STOLZ, O.
Einführung in die Programmierung mit PASCAL
2. überarb. Aufl. 1984, 240 S., DM 24,80
B. G. Teubner, Stuttgart

Es bedarf eigentlich keiner weiteren Herausstellung, wenn nach knapp zwei Jahren jetzt die zweite Auflage dieser Einführung vorliegt. Die Autoren haben es eben verstanden, den Stoff so anzubieten, daß der Leser damit arbeiten kann. Die zweite Auflage wurde dem in der Zwischenzeit erschienenen Normentwurf DIN 66256 angepaßt. – Das vorliegende Buch kann vorbehaltlos einem interessierten Leserkreis empfohlen werden. Ge.

ZALEWSKI, T.
Originäre Nachfrage nach medizinischen Leistungen und Steuerungspotentiale in der ambulanten ärztlichen Versorgung
1984, 131 S., DM 48,–
Asgard-Verlag, Dr. W. Hippe, Sankt Augustin

Die Etablierung ökonomischer Steuerungsinstrumente auf der Nachfrageseite im Gesundheitswesen setzt die Analyse von Steuerungspotentialen voraus. Während die Entscheidung zum Arztkontakt, zumindest sofern es sich um einen Primärkontakt handelt, weitgehend im Ermessensbereich der Patienten liegt, ist über Steuerungspotentiale und damit das Verhalten der Patienten im Leistungsgeschehen, also beim Arztkontakt selbst, nur wenig bekannt.

Im theoretischen Teil dieser Arbeit wird aufgezeigt, daß eine aktive Rolle der Patienten über die Entscheidung zum Arztkontakt hinaus allgemein, insbesondere aber im Leistungssystem der gesetzlichen Krankenversicherung, plausibel unterstellt werden kann.

Auf der Grundlage einer vom Autor durchgeführten Patientenbefragung in der Praxis von Allgemeinärzten findet der Bereich der originären Nachfrage seine empirische Bestätigung. Die Mehrzahl der Patienten gibt über den Besuchsgrund hinaus konkrete Wünsche nach einzelnen medizinischen Leistungen an. Gleichzeitig liefert die empirische Analyse Anhaltspunkte für einen von den Patienten ausgehenden Nachfragedruck sowie für eine qualitative Würdigung der Patientenwünsche.

Im Ergebnis werden Steuerungspotentiale auf seiten der Patienten sowohl im Vorfeld des Arztkontaktes als auch im Leistungsgeschehen zwischen Arzt und Patient aufgezeigt und die Etablierung ökonomischer Steuerungsinstrumente insbesondere im Bereich der Verordnung medizinischer Leistungen empfohlen.

BURHENNE, W. E. und PERBAND, K. (Hrsg.)
EDV-Recht
Systematische Sammlung der Rechtsvorschriften, organisatorische Grundlagen und Entscheidungen zur elektronischen Datenverarbeitung
Ergänzbare Ausgabe, einschl. der 40. Lieferung, 3686 Seiten und 3 Ausschlagtafeln, DIN A5, DM 148,–, zuzügl. 3 Spezialordner je DM 11,80. Ergänzungen folgen von Fall zu Fall.
Erich Schmidt Verlag, Berlin–Bielefeldt–München

Der Textteil dieses stets aktuellen Grundlagenwerkes umfaßt die einschlägigen Rechtsvorschriften, Dokumente, insbesondere die Ausführungsbestimmungen und Gerichtsentscheidungen sowie die Materialien zur Organisation und Zuständigkeit für die EDV in der Verwaltung jeweils für den Bund und die einzelnen Länder.

Der Erläuterungsteil enthält Kommentierungen zum Bundesdatenschutzgesetz und zu den Vertragsbedingungen für EDV-Anlagen und -Geräte. Durch diese Erläuterungen wird die komplizierte Materie des EDV-Rechts auch Nichtjuristen verständlich gemacht.

Systematisierte Nomenklatur der Medizin SNOMED. Band II.
Alphabetischer Index. Herausgeber der amerikanischen Ausgabe: Roger A. Côté. Deutsche Ausgabe bearbeitet und adaptiert von Friedrich Wingert. Berlin: Springer 1984.
und
SNOMED Manual von Friedrich Wingert. Berlin: Springer 1984

Mit dem Erscheinen des Manuals und des Bandes II zu SNOMED ist nunmehr die Voraussetzung für eine praktische Anwendung des SNOMED im deutschen Sprachraum geschaffen.

Auch Band II ist, wie schon Band I, nicht etwa eine einfache Übertragung aus dem Englischen, sondern eine erhebliche Erweiterung und Ergänzung.

Es ist ein echter alphabetischer Index entstanden. Viele Begriffe tauchen nur in diesem Teil auf und sind insoweit häufig eine Interpretation zu Band I. Zu den sieben systematischen Dimensionen sind vier alphabetische Indices entstanden, wobei die Dimensionen „Morphologie", „Funktion", „Ätiologie" und „Krankheit" der besseren praktischen Verwendbarkeit wegen zusammengefaßt wurden. Vor allem idiomatisierte Begriffe, also Begriffe, deren Komponenten nicht oder nur schwer isolierbar sind, lassen sich nämlich leichter über diesen kombinierten alphabetischen Index eindeutig zuordnen. Gleichzeitig kann auf diese Weise auch am jeweiligen Beispiel die Indexierung geübt werden, wenn nämlich aus Band II in Band I gegangen wird und andere Möglichkeiten überprüft werden.

Die Anwendung sei an einem Beispiel erläutert:

Subarachnoidalblutung ist ein Begriff, der nur im zweiten Band auftaucht. Der angegebene Code ist M 37000 (das bedeutet laut Band I Blutung) und T X1500 (= Subarachnoidalraum); in Band I findet sich außerdem als Modifikation für eine akute Blutung, um die es sich wohl stets handeln wird: M 37001. Aber: „Subarachnoidalblutung" wird in der Klinik häufig als Diagnose, die eigentlich nur eine vorläufige Symptombeschreibung ist, verstanden, und zwar ist sie die Beschreibung für die Folge einer Grundkrankheit, nämlich einer Ruptur eines Aneurysma einer Arterie (zumeist zur Versorgung des Cortex). Das ergibt nun aber den Code M 32401 (= Ruptur eines Aneurysma; Aneurysma ohne Ruptur wäre: M 32400) und T 45000 (= Arterie im Kopfbereich; eventuell, falls nachgewiesen, statt der Null der Code einer bestimmten Arterie).

Man kann also, wie man sieht, in SNOMED wie in der Klinik zu demselben Sachverhalt zu ganz unterschiedlichen Aussagen kommen. Aber da die zweite Aussage quasi synonym zur ersten benutzt wird, sind auch die beiden Verschlüsselungen quasi synonym. Das wie-

derum würde beim Retrieval Probleme machen, wenn das entsprechende medizinische Wissen in der Datenbank nicht gespeichert ist. Es ist daher empfohlen, den SNOMED nicht nur als Codierungsvorschrift, sondern als Sprache zu benutzen, wozu auch das Vorhandensein einer Reihe einfacher Modifikatoren und syntaktischer Links gehört. Man kann dementsprechend die beiden Aussagen unseres Beispiels miteinander verknüpfen, und zwar, da ein Begründungszusammenhang besteht, mit dem syntaktischen Link DT, und erhält also: M 37001 T X1500 DT M 32401 T 45000 NL. NL am Schluß markiert dabei das Ende eines syntaktischen Zusammenhangs.

Das Manual dient der systematischen wie auch praktischen Einführung in SNOMED. Es ist weitgehend von Wingert neu geschrieben. Lediglich der praktische Teil ist nur adaptiert.

Die Kapitel 1 bis 5 sind zu einem sehr kompakten Lehrbuch der Dokumentation mit einigen Randbemerkungen zur medizinischen Dokumentation geraten. Für Leute vom Fach mit guter formaler Ausbildung ist dieser Teil eine gute Wiederholung (und evtl. eine Vertiefung) des Stoffgebietes, für andere aber weder lesbar noch zur praktischen Anwendung des SNOMED notwendig. Hierzu dienen die aus der amerikanischen Ausgabe übernommenen allgemeinen Hinweise mit Codierübungen. Sie sind nach steigendem Schwierigkeitsgrad geordnet und sicherlich für ihren Zweck sehr geeignet, wenn man sie durch analoge Beispiele zur weiteren Einübung ergänzt. Somit dürfte einer praktischen Anwendung des SNOMED nichts im Wege stehen.

Gerhard K. Wolf (Heidelberg)

SAVORY, S. E. (Hrsg.)
Künstliche Intelligenz und Expertensysteme
Ein Forschungsbericht der Nixdorf Computer AG
1985, 248 S., DM 39,80
R. Oldenbourg Verlag, München–Wien

In dem vorliegenden Buch sind die Referate eines KI-Workshops zusammengestellt. Dadurch erhält der Leser einen ausgezeichneten Überblick über den Stand der Forschung und Entwicklung auf diesem aktuellen Gebiet. Bei der Lektüre gewinnt man den Eindruck, daß »Expertensysteme« bald in vielen Bereichen Realität sein werden.

Inhalt:

Artificial Intelligence – State of the Art 1984; Expertensysteme für den kommerziellen Einsatz; Funktionen und Arbeitsweise der Expertensystem-Shell TWAICE; PROLOG-Implementierungssprache der künstlichen Intelligenz; Das Wesen des Knowledge Engineering; Wissenserwerb und maschinelles Lernen; Lernen aus Beispielen; Induktives Lernen von Grammatikregeln aus ausgewählten Beispielen; Skizze einer Beschäftigungstheorie in PROLOG; Punify – ein assoziativer Prozessor für die Unifikation; Nixdorf und die KI-Fördervorhaben des Bundes. Ge.

FERSTL, O. K. und SINZ, E. J.
Software-Konzepte der Wirtschaftsinformatik
1984, 286 S., DM 38,–
W. de Gruyter, Berlin–New York

Das vorliegende Buch gibt eine Einführung in Software-Konzepte der Wirtschaftsinformatik, die als Grundlage für die Programmentwicklung benötigt werden. Es erfolgt eine Anlehnung an PASCAL und Modula-2 als Programmiersprachen. Unabhängig von der Fachrichtung Wirtschaftsinformatik kann die Darstellung allgemein als Einführung in die Datenverarbeitung benutzt werden. Ge.

WALDSCHMIDT, E. H. und WALTER, H. K.-G.
Grundzüge der Informatik I
Reihe Informatik, Bd. 43
1984, 338 S., DM 38,–
Bibliographisches Institut, Mannheim/Wien/Zürich

Im Mittelpunkt dieser Einführung steht der Algorithmusbegriff, dabei wurde bewußt vermieden, eine Einführung in die Programmierung zu schreiben. Die aber dennoch erforderliche Anlehnung an eine problemorientierte Programmiersprache wurde unter Verwendung von PASCAL vorgenommen. Die Darstellung erscheint insbesondere für Studierende der Informatik, der Mathematik und der Wirtschaftsinformatik geeignet. Ge.

WISHART, D.
CLUSTAN – Benutzerhandbuch (3. Ausgabe)
Aus dem Englischen übersetzt von J. B. SCHÄFFER
1984, 244 S., DM 48.–
G. Fischer Verlag, Stuttgart–New York

Softwarepakete erhalten erst eine genügende Verbreitung, wenn brauchbare Benutzerhandbücher vorliegen. Es ist zu begrüßen, daß nun für das Programmpaket CLUSTAN auch eine deutsche Übersetzung des Benutzerhandbuches herausgekommen ist. Eine Ergänzung geht dabei auf die Version 2.1 ein. Damit sollten eigentlich die Voraussetzungen gegeben sein, daß Methoden der Clusteranalyse ohne größere Schwierigkeiten angewandt werden können. Ge.

SCHAFFLAND, H.-J. und WILTFANG, N.
Bundesdatenschutzgesetz (BDSG)
Ergänzbarer Kommentar nebst einschlägigen Rechtsvorschriften
Ergänzbare Ausgabe, einschl. 14. Lieferung, 1134 Seiten, DIN A5, DM 86,–, zuzügl. Ordner DM 11,80. Ergänzungen folgen von Fall zu Fall.
Erich Schmidt Verlag, Berlin–Bielefeld–München

Die Bedeutung des Bundesdatenschutzgesetzes (BDSG) gewinnt in der Praxis und in der Rechtsprechung immer mehr an Bedeutung.

Mit dem vorliegenden Kommentar steht ein zuverlässiger Ratgeber zur Verfügung, der auf die Bedürfnisse der Praxis abgestellt ist und in die Hand eines jeden Datenschutzbeauftragten gehört.

SPÄTH, H.
Cluster-Formation und -Analyse
Theorie, FORTRAN-Programme, Beispiele
1983, 236 S., DM 84,–
R. Oldenbourg Verlag, München–Wien

Nach einer ausgezeichneten Einführung in die Methoden werden anschließend FORTRAN-Subroutinen dargestellt und mit diesen durchgerechnete Beispiele eingehend erläutert.

So ergeben sich nützliche Hinweise auf die Methodenauswahl beim Einsatz für praktische Probleme. Zur Benutzung auf Großrechnern ist ein Magnetband mit allen Programmen und Beispielen erhältlich.

Der Einsatz der Programme auf Mikrorechnern ist nach geringfügigen Änderungen, die beschrieben sind, möglich.

REPGES, R. und TOLXDORFF, Th. (Hrsg.)
Strukturen und Prozesse – Neue Ansätze in der Biometrie
Med. Informatik und Statistik Bd. 56
1984, 138 S., DM 29,50
Springer-Verlag, Berlin–Heidelberg–New York

In diesem Buch sind Übersichtsreferate des 28. Kolloquiums der Biometrischen Gesellschaft zusammengestellt, die insbesondere neue mathematische Ansätze für typisch biologische Fragestellungen behandeln. Die einzelnen Referate sind jedes für sich ausgezeichnet und sollten allen Biometrikern eigentlich als Pflichtlektüre empfohlen werden.

Behandelt werden u. a.: mathematische Methoden in Biologie und Ökologie, neue Entwicklungen in der Versuchsplanung, sequentielle Verfahren und Entscheidungsprozesse.

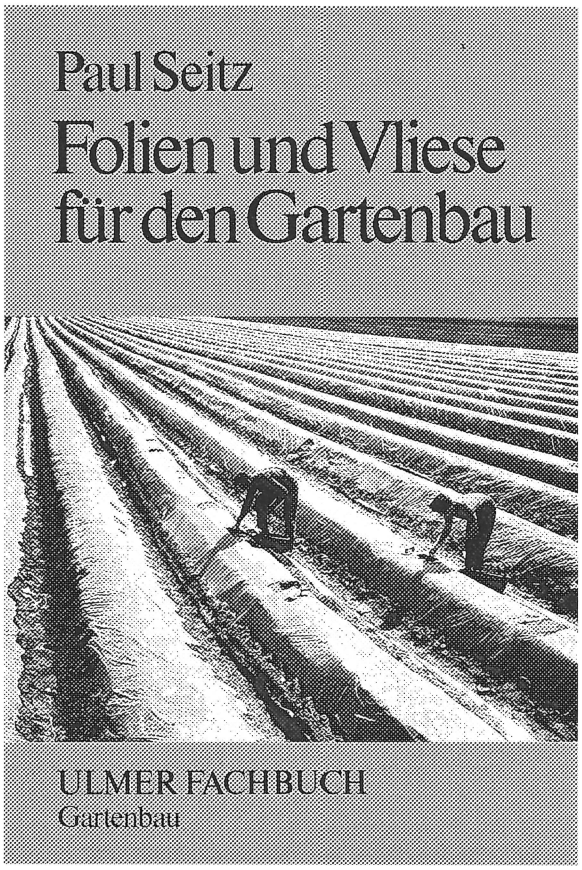
Berichtigung

In EDV Heft 1/1985, Band 16, fehlt im Beitrag der Autoren Gerald Morawe und Wolfgang Horst „Kompartimentierung von Räumen mit Randbedingungen" die Anschrift der Verfasser.
Wir bitten dieses Versehen zu entschuldigen.

Die Anschrift der Autoren lautet:
Gerald Morawe, Wolfgang Horst
Abteilung für Biomathematik
Zentrum der medizinischen Informatik
Fachbereich Humanmedizin
Johann-Wolfgang-Goethe-Universität
Theodor-Stern-Kai 7, 6000 Frankfurt

NEU!

Folien und Vliese für den Gartenbau. Der Autor geht in diesem → **Fachbuch** allen Fragen nach, die mit dem Einsatz von Folien und Vliesen im Gartenbau entstehen. Er beschreibt Material, Herstellung, Eigenschaften und Unterscheidungsmerkmale der Folien und Vliese, geht ein auf ihre → **Einsatzbereiche** und zeigt die Konsequenzen für die → **Kulturführung** auf. Übersichtliche Tabellen und eine vorzügliche Bebilderung vervollständigen das Werk. Ein umfassendes Buch für den → **umweltbewußten und fortschrittlichen** Gärtner. Von → **Dr. Paul Seitz,** Frankfurt. Neubearb. und erweiterte 2. Aufl. 244 Seiten mit 84 Schwarzweißfotos, 43 Zeichn. und 43 Tab. Kst. → **DM 58,–** (Ulmer Fachbuch Gartenbau).



Buch-Coupon an:
Ihre Buch(Fach)handlung oder
Verlag Eugen Ulmer,
Postfach 70 05 61, 7000 Stuttgart 70

52711 Folien DM 58,—

Name, Vorname

Straße/Nr.

PLZ, Ort

Unterschrift, Datum

E.U.
VERLAG EUGEN ULMER