Statistische Auswertungen mit Programmpaketen

Autor(en): **Rüst, H.**

Objekttyp: **Article**

Zeitschrift: **Mitteilungen / Vereinigung Schweizerischer Versicherungsmathematiker = Bulletin / Association des Actuaires Suisses = Bulletin / Association of Swiss Actuaries**

Band (Jahr): **75 (1975)**

PDF erstellt am: **27.05.2024**

Persistenter Link: https://doi.org/10.5169/seals-967116

Statistische Auswertungen mit Programmpaketen

Von H. Rüst

Nach dem an der Mitgliederversammlung 1975 gehaltenen Vortrag

1. Beispiele von statistischen Erhebungen

1.1 Grössenordnung der statistischen Erhebung

Es soll gezeigt werden, welche Grössenordnung von statistischen Studien und welche Arbeitsschritte der statistischen Auswertung für die Verwendung von Statistikprogrammpaketen geeignet sind. Ein Beispiel soll die Arbeitsabläufe und die Verwendungsmöglichkeiten solcher Programme erläutern.
Wir wollen annehmen, dass die Erhebung etwa 100 bis 2000 Fälle umfasst und dass pro Fall ein Fragebogen vorliegt, welcher die Ergebnisse der Untersuchung in einer Form enthält, die für die elektronische Datenverarbeitung geeignet ist. Die Fragebogen können Antworten, Messwerte oder Beobachtungen enthalten. Zählt man alle diese Angaben für sämtliche Fragebogen zusammen, so wird man mindestens einige hundert, meistens aber einige tausend und oft sogar einige zehntausend Angaben erhalten, welche in die Auswertung einzubeziehen sind. Damit ist man sicher in einer Grössenordnung, wo die Computerauswertung der Handauswertung mit Hilfe von Tischcomputern überlegen ist.
Müsste man die Auswertungsprogramme für eine solche Erhebung speziell erstellen, so müsste man etwa 2000 bis 10000 Befehle z. B. in der Programmiersprache FORTRAN oder PL/1 schreiben.

1.2 Beispiel: Klinische Studie

Abbildung 1 zeigt die Datenstruktur einer klinischen Studie.
In einer Multicenterstudie, also bei Patienten von mehreren Prüfern, soll ein neues Medikament mit einer bewährten Behandlungsmethode verglichen werden.

Die Patienten werden nach einem Zufallsverfahren in zwei Behandlungsgruppen eingewiesen und nach einem festgelegten Verfahren behandelt. Für die Auswertung werden nun folgende Angaben festgehalten:

- Angaben über Prüfer bzw. Klinik.
- Angaben über den Patienten: Personalien, Kriterien, die zur Aufnahme in die Studie geführt haben, Vorgeschichte der Erkrankung.
- Angaben über die Behandlung, d.h. insbesondere Abweichungen vom Behandlungsplan.
- Der Patient wird nach einem vorgegebenen Zeitplan untersucht, und dabei werden quantitative und qualitative Angaben zur Wirksamkeitsbeurteilung festgehalten. Anlässlich dieser Untersuchungen wird der Prüfer auch alle festgestellten oder vom Patienten angegebenen unerwünschten Wirkungen ermitteln, sowie sämtliche Medikamente, welche der Patient während der Prüfung neben dem Prüfmedikament eingenommen hat, angeben.
- Vor, während und nach der Prüfung werden Blut- und Urinuntersuchungen durchgeführt. Man erhält daraus etwa 20–30 sog. Laborwerte pro Analyse. Da die Analysemethoden in den Labors differieren, benötigt man von jedem Labor noch die sog. Normalwerte. (Diese können zusätzlich noch vom Alter und Geschlecht des Patienten abhängen.)

Auf Grund der Ergebnisse der statistischen Auswertung soll die bessere Behandlung festgestellt werden können. Massgebend dafür ist die Beurteilung der Wirksamkeit. Zu kontrollieren ist ferner die Verträglichkeit. Unerwünschte Wirkungen und abnormale Laborwerte müssen festgehalten werden.

2. Statistikprogrammpakete

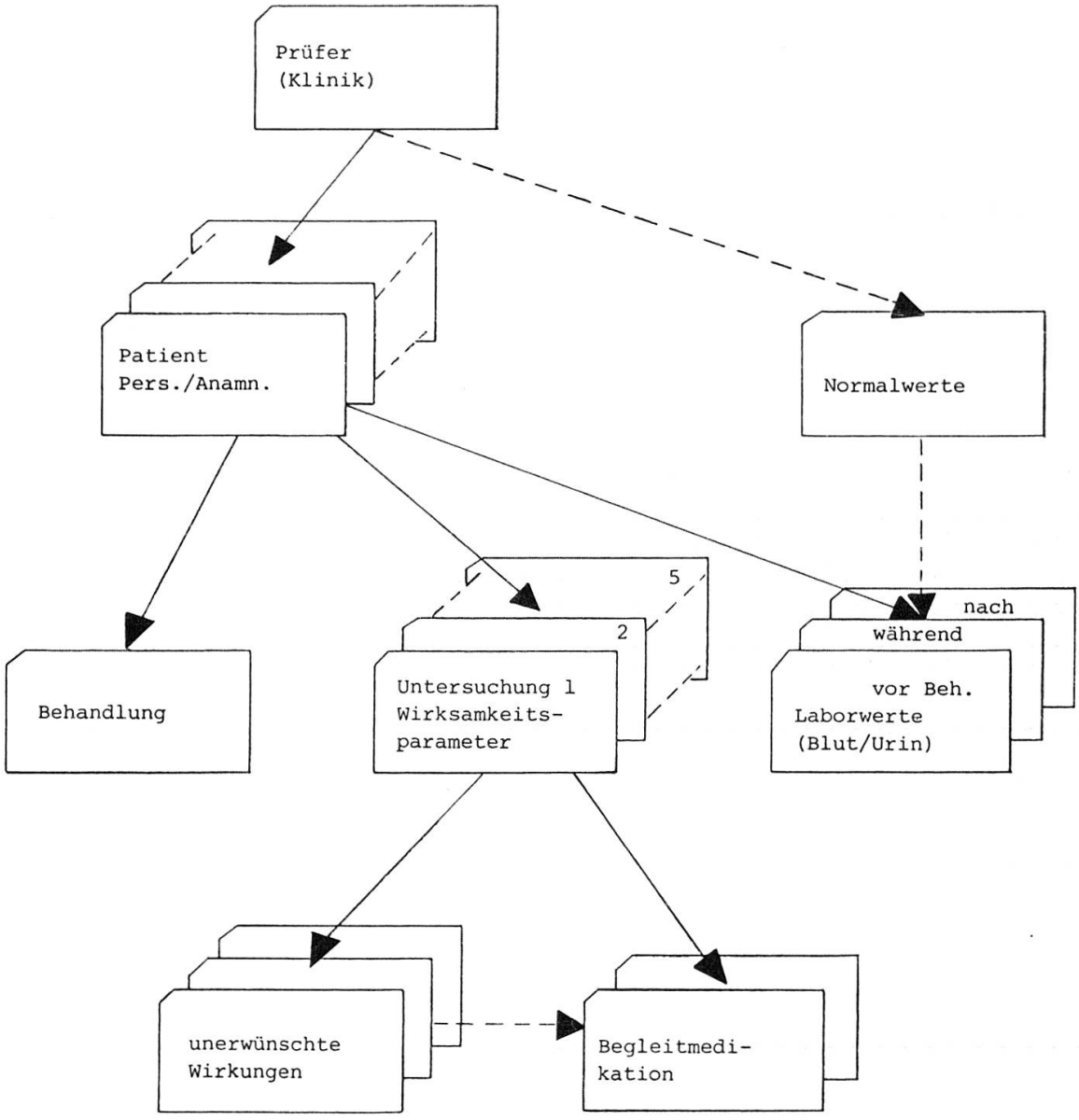
2.1 Auswahl von zwei Systemen

Unter einem Statistikprogrammpaket verstehen wir ein allgemein für statistische Auswertungen verwendbares, standardisiertes System von Computerprogrammen.

Solche Statistikprogrammpakete wurden hauptsächlich an Rechenzentren von amerikanischen Universitäten erarbeitet. Sie sind heute soweit entwickelt, dass sie für die statistische Auswertung verschiedener Studien eingesetzt werden können.

Abbildung 1

Datenschema einer klinischen Studie



Die Auswahl der folgenden beiden Programmpakete ist subjektiv. Es sind die einzigen, mit denen wir in unserer Firma eine praktische mehrjährige Erfahrung haben.

2.2 SPSS (Statistical Package for the Social Sciences)

SPSS wurde seit 1965 von einer Arbeitsgruppe um Norman H. Nie zunächst an der Standford University und später an der University of Chicago auf IBM-Maschinen entwickelt. Seit 1970 ist auch eine Version für CDC-Maschinen verfügbar. Diese Version wird vom Rechenzentrum der Northwestern University betreut.
Es sollen jedoch heute Versionen für etwa 20 verschiedene Computer und Betriebssysteme in 600 Installationen existieren. In Zürich ist SPSS insbesondere an den Rechenzentren der beiden Hochschulen verfügbar.
Die folgenden Ausführungen beziehen sich auf die neueste Version 6.0, installiert auf einem Computer CDC 6400.
Das Programm ist in FORTRAN und Assembler (Maschinensprache) geschrieben. Das Programm-Listing umfasst etwa 2800 Seiten. Das Programm besitzt eine komplizierte Struktur mit 51 Overlays.

2.3 SAS (Statistical Analysis System)

Das SAS-Programmpaket (Statistical Analysis System) wurde seit 1966 von einer Arbeitsgruppe um A.J.Barr an der North Carolina State University entwickelt.
Es ist in Assembler (Maschinensprache), PL/1 und FORTRAN geschrieben und kann nur auf grossen IBM-Maschinen betrieben werden.
Es soll jedoch auch bereits an über 100 Rechenzentren installiert sein.
Die folgenden Ausführungen beziehen sich auf die Version 75.1. Diese Version ist von den Erstellern allerdings erst provisorisch freigegeben, da sie noch fertig ausgetestet wird. Sie weist jedoch gegenüber der Version 72 einige interessante Neuerungen auf.
Das Programm-Listing umfasst etwa 500 Seiten.

3. Bereitstellung und Überprüfung der Daten für die statistische Auswertung

3.1 Der Datenfile

Die Grundeinheit für die spätere statistische Bearbeitung ist der Datenfile. Dieser besteht zunächst aus der Datenmatrix, den Variablenwerten geordnet nach Fällen.
Weiter können in diesem Datenfile für alle oder für einen Teil der Variablen noch folgende Angaben eingegeben werden:
Variablenbenennung: Ausführliche Bezeichnung der Variablen.
Variablenwertbenennung: Bezeichnung der Variablenwerte (Codezahlen).
Kennzeichnung fehlender Werte: Diese gekennzeichneten Werte werden in der späteren Auswertung gesondert behandelt.
Diese zusätzlichen Angaben stehen bei den statistischen Auswertungen zur Verfügung. Auf den ausgedruckten Tabellen erscheinen sie dann automatisch in den Tabellenüberschriften oder an entsprechender Stelle der Ergebnisdarstellung.
Die Benennung von Variablenwerten ist nur beim SPSS-Programm möglich. Damit wird die Verständlichkeit der Resultattabellen wesentlich erhöht.

3.2 Steuerung des Programmablaufs

Für die Steuerung des Programmablaufs hat man beim SPSS-System etwa 80 und beim SAS-System etwa 50 verschiedene Anweisungen zur Verfügung. Diese Anweisungen sind sehr flexibel aufgebaut. Insbesondere besteht in der Regel die Möglichkeit, sie auf eine ganze Liste von Variablen wirken zu lassen. Im einfachsten Fall werden diese aneinandergereihten Anweisungen sequentiell durchlaufen und ausgeführt. Sie nehmen so das Aussehen eines Computerprogramms an und funktionieren in vielen Punkten auch analog wie Computerprogramme. Wir können sie deshalb auch als problemorientierte Makroprogrammiersprachen bezeichnen.
In der Folge sollen solche Aneinanderreihungen von SPSS- bzw. SAS-Anweisungen, welche zur Steuerung des Systemdurchlaufs verwendet werden, als SPSS- bzw. SAS-Programme bezeichnet werden.
Interessant ist der sequentielle Programmablauf beim SPSS-Programm. Dieser Ablauf kann lediglich durch die Verwendung von Schlaufen gelockert werden.

Bei allen Berechnungen kann die Ausführung an die Erfüllung von Bedingungen geknüpft werden.
Beim SAS-Programm sind Verzweigungen möglich und mit Hilfe von sog. MACROS können eine Art Unterprogramme eingesetzt werden. Weiter hat man bei SAS-Programmen die Möglichkeit, gleichzeitig mehrere Datenfiles zu bearbeiten und sie auch miteinander zu verknüpfen.

3.3 Berücksichtigung der Datenstruktur

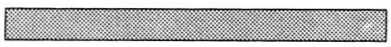
Die Daten der klinischen Studie sind hierarchisch aufgebaut (siehe Abbildung 1).
Ohne die Normalwerte hätte man eine einfache Baumstruktur. Diese zusätzlichen Beziehungen machen die Struktur zu einem sog. Netzwerk.
Dies sind jedoch bereits Begriffe aus der Theorie der Datenbanken. Ein Hauptproblem bei der Bearbeitung der Datenbanken ist die Bewältigung von komplizierten Datenstrukturen. Die EDV-Fachleute haben auch bereits Programmiersysteme entwickelt, welche den Umgang mit Datenstrukturen ermöglichen.
Leider ist dieses Problem bei den Statistikprogrammpaketen noch nicht gelöst. (Sowohl die SPSS- wie auch die SAS-Projektgruppen schreiben in ihren periodischen Mitteilungen, dass sie an diesen Problemen arbeiten.)
Bei unserem Beispiel muss man sich zunächst entscheiden, welche Stufe man als Zeile in der Datenmatrix, also als Fall betrachten will.
Man kann z. B. die Variablen nach Patienten gliedern oder aber auch nach Untersuchung oder innerhalb der Untersuchung nach unerwünschter Wirkung usw. (siehe Abbildung 2).
Wählt man z. B. den Patienten als Zeilengliederung, so kann man trotzdem noch Angaben höherer Stufe im Datenfile, etwa Angaben über den Prüfer, festhalten. Diese Angaben müssen dann jedoch bei sämtlichen Patienten desselben Prüfers wiederholt werden.
Wird der File auf einer hohen Stufe der Hierarchie gegliedert, so muss man mit vielen Variablen rechnen; gliedert man auf einer tiefen Stufe, so werden Angaben aus höheren Stufen häufig repetiert werden müssen.
In der Praxis wird man in der Regel mehrere Datenfiles erstellen müssen, welche zum Teil gleichartige, zum Teil disjunkte Variablenklassen enthalten werden.
Beim SPSS ist man sehr eingeschränkt bei der Bearbeitung von Datenstrukturen. Zunächst muss man die Daten nach einem starren Format, welches zwar

Abbildung 2

Organisation der Datenmatrix

Fall-Nr.	Pat. Nr.	Untersuchung 1	Untersuchung 2	.....
		Wirksamk.param.	Wirksamk.param.	....
1	101			
2	102			
:	:			

Fall-Nr.	Pat. Nr.	Unt. Nr.	Wirsamkeits-parameter	1.unerw.W.	2.unerw.W.
				Angaben	Angaben
1	101	1			
2	101	2			
3	101	3			
4	101	4			
5	101	5			
:					

Fall-Nr.	Pat. Nr.	Unt. Nr.	Angaben über unerw. Wirkung	Angaben über Untersuchung
1	101	2		
2	101	2		
3	101	4		

vom Benützer eingegeben werden kann, aber dann für alle Fälle fest ist, einlesen. Dann kann man während desselben SPSS-Durchlaufs jeweils nur mit einem einzigen Datenfile arbeiten.
Beim SAS-System hat man wesentlich mehr Möglichkeiten. Man kann das Einleseformat während des Einlesens verändern. Man kann mehrere Einleseanweisungen kombinieren. Man kann mit mehreren Datenfiles im selben Programmdurchlauf arbeiten.

4. Die statistische Auswertung

4.1 Aufteilung des Ablaufs der Auswertungsarbeiten

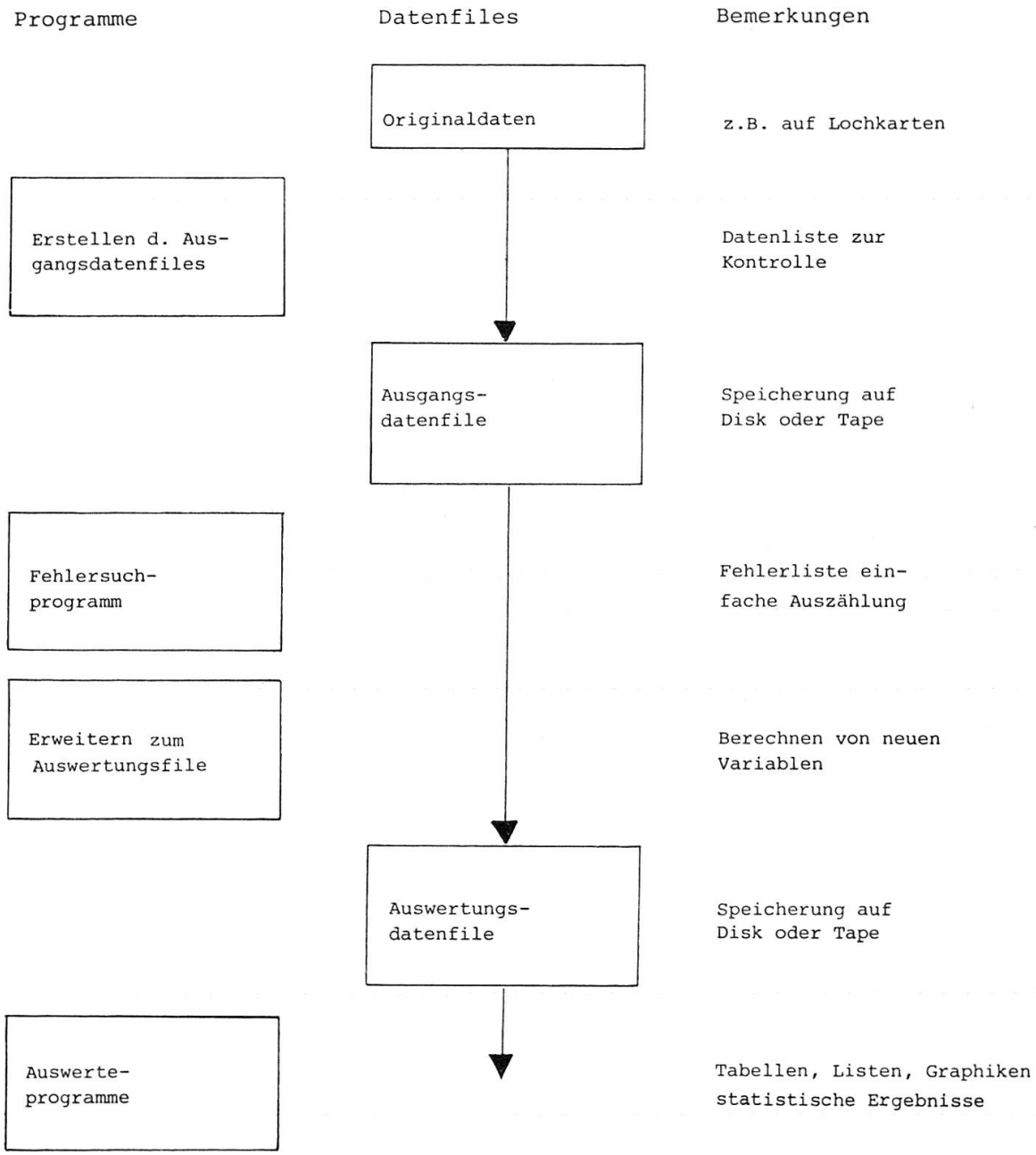
Für die Auswertung einer Studie wird man einige hundert bis einige tausend SPSS- bzw. SAS-Anweisungen benötigen.
Man hat damit den Programmieraufwand gegenüber der Verwendung einer gewöhnlichen Programmiersprache auf etwa 20–50% verringert.
Es wäre nun zwar möglich, alle diese Befehle in einem einzigen Programmdurchlauf zu bewältigen. Dieses Vorgehen wäre aber sicher nicht zweckmässig. Das Programm wäre unübersichtlich und müsste wegen eines einzigen Fehlers nochmals gerechnet werden.
Eine mögliche und zweckmässige Aufteilung zeigt die Abbildung 3. Der Begriff des Datenfiles und die Probleme bei seiner Erstellung wurden bereits im Kapitel 3 besprochen.

4.2 Das Fehlersuchprogramm

Es ist auch bei sorgfältig geplanten und durchgeführten Studien damit zu rechnen, dass ein Teil der Angaben falsch oder unvollständig ist (oder im Verlauf der Verarbeitung falsch interpretiert und weitergegeben wird). Es muss versucht werden, diese Fehler zu entdecken oder wenigstens dafür zu sorgen, dass sie das Ergebnis nicht ungebührlich beeinflussen können. Weiter ist zu überprüfen, ob die Vorschriften des Versuchsplanes und für das Ausfüllen des Fragebogens eingehalten wurden.
Das Erstellen eines SPSS- oder SAS-Fehlersuchprogrammes erfordert etwa denselben Aufwand, wie wenn es als FORTRAN- oder PL/1-Programm erstellt

Abbildung 3

Ablauf einer Auswertung



würde. Trotzdem dürfte es zweckmässig sein, dieselbe Programmiersprache zu verwenden, da gleichzeitig auch einige systeminterne Funktionen wie Variablenbenennung, fehlende Werte usw. überprüft werden können.

4.3 Erweiterung des Datenfiles

In der Regel genügen die eingelesenen Variablen noch nicht für die statistische Auswertung.
Durch Umkodierung, Transformation oder Verknüpfung von Variablen werden neue Variablen berechnet oder bereits definierte modifiziert.
Weiter können die Fälle nach gewissen Kriterien sortiert werden; Fälle können aus dem File entfernt werden.
Sowohl im SPSS- wie auch im SAS-System sind solche Rechnungen und Umformungen leicht durchzuführen.

4.4 Deskriptiver Teil der statistischen Auswertung

Der deskriptive Teil der statistischen Auswertung ist meistens der aufwendigste. Die grossen und unübersichtlichen Datenmengen müssen kondensiert und so dargestellt werden, dass sie verständlich werden und dass unsere Aufmerksamkeit auf wichtige Zusammenhänge und Beziehungen gelenkt wird. Man verwendet dafür Tabellen, Listen und Graphiken und berechnet statistische Masszahlen wie Durchschnitt, Streuung oder Extremwerte.
Für diesen deskriptiven Auswertungsteil stehen bei beiden Programmsystemen ähnliche Verfahren zur Verfügung:

- Kreuztabellen: ein- und mehrdimensionale Häufigkeitsauszählungen, Prozentzahlen, verschiedene Assoziationsmasse, χ^2-Test;
- Mittelwerte, Streuungen und andere Statistiken von Zufallsvariablen;
- Korrelationskoeffizienten (Produkt-Moment, Spearman, Kendall): Berechnung und Test auf Null;
- graphische Darstellungen: Histogramme, Scattergramme;
- Listen: Variablenwerte fallweise gelistet für ausgewählte oder für alle Fälle.

Meistens kann einem jedoch das Programmpaket nicht die ganze Arbeit abnehmen. Die Ergebnisse müssen oft noch in kompakteren Tabellen wieder zusammengefasst werden.

4.5 Analytischer Teil der statistischen Auswertung

Die Früchte der umfangreichen Vorarbeiten erntet man erst beim analytischen Teil der statistischen Auswertung, bei der Durchführung von komplizierten Verfahren der Datenanalyse.
Die verfügbaren Verfahren für einfache Tests sind beinahe disjunkt.

SPSS:
- Vergleich von Mittelwerten: *t*-Test für unabhängige und für gepaarte Stichproben, Ein-Faktor-Varianzanalyse mit Rangordnungstest, Test auf Trend und Kontraste;
- Berechnung von partiellen Korrelationskoeffizienten und Test auf Null;
- nichtparametrische Tests: Kolmogorov-Smirnov, McNemar, Wilcoxon, Friedman und andere Verfahren.

SAS:
- Vergleich von Mittelwerten mit multiplem Rangordnungstest;
- Maximum-Likelihood-Schätzungen der Parameter der Probit-Gleichung.

Die Möglichkeiten für Regressionsrechnung und Varianzanalyse stimmen wieder eher überein:
- Regressionsrechnung: multiple Regression, schrittweise Verfahren, Varianzanalyse, graphische Darstellung der Residuen, Schätzen von Parametern nichtlinearer Regressionsgleichungen;
- Varianzanalyse: bei SPSS Modelle mit mehrfacher Klassifikation (nicht hierarchisch) mit Interaktionen und Kovariablen auch für nichtorthogonale Versuchsanlagen; bei SAS Modelle mit mehrfacher Klassifikation nur für orthogonale Versuchsanlagen und Modelle mit hierarchischer Klassifikation.

Für die multivariate Datenanalyse kommen grösstenteils dieselben Verfahren zum Einsatz:
- Faktorenanalyse;
- Diskriminanzanalyse;
- Kanonische Korrelation.

Beim SAS sind zudem noch verfügbar:
- Clusteranalyse;
- Schätzen der Koeffizienten bei Systemen von linearen Regressionsgleichungen;
- Spektralanalyse von mehrdimensionalen Zeitreihen.

Ein Teil der Kritik gegen Statistikprogrammpakete richtet sich gegen die missbräuchliche Verwendung solcher analytischer Verfahren.
Grundsätzlich wird man verlangen müssen, dass ein Anwender ein Verfahren nur dann benützen soll, wenn er es mit allen Einzelheiten versteht. Diese Arbeit kann ihm von den Verfassern der Programmdokumentation erleichtert werden.
Wir haben erst einen Teil der Verfahren studiert und angewendet. Insbesondere die Neuerungen der neuesten Programmversionen haben wir erst zu einem kleinen Teil ausprobiert.
Das SPSS-Handbuch ist vollständig und übersichtlich gestaltet. Beim SAS-Manual bereitet einem vor allem die Übersicht Mühe. Zudem wird anstelle von ausführlichen Verfahrensbeschreibungen auf die Literatur verwiesen.

5. Zusammenfassende Beurteilung

Bei der statistischen Auswertung von grösseren Studien (100–2000 Fälle mit insgesamt einigen hundert bis einigen zehntausend Angaben) kann der Programmieraufwand durch die Verwendung von Statistikprogrammpaketen auf 20–50% gesenkt werden.
Bei komplizierteren Datenstrukturen hat man Mühe, die Daten in der erforderlichen Form bereitzustellen.
Bei beiden Programmen hat man sehr vielseitige Möglichkeiten für die Bearbeitung und Umformung der Ausgangsdaten.
Für den Ablauf der statistischen Auswertung ist es zweckmässig, die Verwendung der Programmpakete in mehreren Schritten durchzuführen. Die Folge der Anweisungen zur Steuerung der Programmdurchläufe bezeichnen wir als SPSS- bzw. SAS-Programm.
Für den deskriptiven Teil der statistischen Auswertung stellen beide Programmpakete ähnliche Verfahren zur Verfügung. Die Ergebnisse der SPSS-Auswertung werden jedoch übersichtlicher dargestellt und vollständiger beschrieben.
Beide Programmpakete weisen vielseitige und verschiedenste Verfahren für den analytischen Teil der statistischen Auswertung auf, die sich nur teilweise dekken. Die beiden Programmpakete ergänzen sich. Jedes besitzt hinsichtlich der Auswertung von statistischen Erhebungen Vorzüge gegenüber dem anderen Paket. Zu bedauern ist eigentlich nur, dass die guten Eigenschaften beider Pakete nicht in einem einzigen Paket vereinigt sind.
Im SAS-System ist jedoch bereits ein Verfahren vorgesehen, um einen Datenfile, welcher mit dem SPSS-System erzeugt wurde, in einen SAS-Datenfile umzuwandeln.

Literatur

Nie, Normann H., Hull Hadlai, C., Jenkins, Jean G., Steinbrenner, Karin, Bent, Dale H.: SPSS Statistical Package for the Social Sciences, 2nd Edition, McGraw Hill, New York, 1975.
Service Jolayne: SAS Version 75. 1 Prov. Release User's Guide Statistical Analysis System designed and implemented by *Anthony J. Barr/James H. Goodnight*, Institute of Statistics, North Carolina State University, 1975.
Francis, Ivor, Heiberger, Richard M., Velleman, Paul F.: Criteria and Considerations in the Evaluation of Statistical Program Packages. The American Statistician *29*, 1975, 52–56.
Nelder, J. A.: A User's Guide to the Evaluation of Statistical Packages and Systems. International Statistical Review *42*, 1974, 291–298.
Martin, James: Computer Data Base Organization, Prentice-Hall, Inc., Englewood Cliffs, 1975.

Dr. Hanspeter Rüst,
c/o Wirtschafts-Mathematik AG
Mühlebachstrasse 38,
8032 Zürich