Erfahrungen mit den Statistikprogrammen SAS und SPSS[1]

Gerhart W. Schnaitmann

In den beiden hier darzustellenden und in den Statistikanalysen verwendeten statistischen Programmpaketen SAS und SPSS besteht eine Gemeinsamkeit darin, daß sie mit (Variablen-) Namen für die Spalten der Datenmatrix arbeiten. Diese Namen bilden dann die Grundlage für die Manipulation des Datenfiles und für den Einsatz der statistischen Prozeduren. Beide Programmpakete erlauben die Generierung von neuen Variablen durch Transformation, die Bildung neuer Dateien durch Verknüpfung vorhandener Dateien, die Selektion von Fällen usw. Die Systeme erlauben eine Beschriftung des Output durch Überschriften, Labels für die Variablen und Labels für die Variablenausprägungen. Das Methodenspektrum ist einigermaßen vergleichbar, wobei SPSS das Schwergewsicht auf sozialwissenschaftliche Anwendungen und SAS keinen Anwendungsbereich speziell hervorhebt.

1. SAS

SAS bedeutet Statistical Analysis System und ist ein Software-System für Daten-Analyse.
SAS eignet sich nicht nur für Statistik-Analysen, sondern auch zur Datenverwaltung und verfügt darüber hinaus über die Eigenschaften einer problemorientierten höheren Programmiersprache.
Bei SAS ist es jedoch nicht notwendig, Programmiersprachen zu lernen.
SAS, wie auch SPSS, gibt es mittlerweile für Großrechner („host“ bzs. „mainfram“ in der Version 6.06) wie für PC (Version 6.03 und 6.04).
Das Programmsystem besteht im einzelnen aus einem Basispaket (Base SAS) und mehreren Zusatzpaketen für speziellere Anwendungszwecke.
Das Basispaket stellt die Basissoftware für alle Zusatzpakete dar und beinhaltet u.a. die Sprachelemente von SAS und einige grundlegende Anwendungsprozeduren, u.a. für deskriptive statistische Analysen zur Aufbereitung von Daten. Das für die statistische Analyse wichtigste Zusatzpaket ist SAS/STAT mit Analyseprozeduren für univariate Analysen (Berechnung deskriptiver statistischer Kernwerte für numerische Variablen, z.B. Mediane, Extremwerte der Verteilung, Mittelwerte, Standardabweichungen) und für multivariate statistische Analysen (z.B. multiple Korrelations- und Regressions-, Faktoren-, Cluster-, Diskriminanzanalysen).
Darüber hinaus gibt es weitere Programmteile für grafisches Gestalten (SAS/GRAPH), für die Erleichterung des Editierens von Dateien (SAS/FSP) oder für statistische Qualitätskontrollen (SAS/QC).
SAS wurde einmal deshalb bei der Auswertung der empirischen Daten in unserem Zentrum angewandt, weil das Universitätsrechenzentrum von Heidelberg eine sehr günstige Campus-Lizenz zur Verfügung stellt und weil man mit SAS sehr umfangreiche Datenanalysen ohne großen Aufwand der Speichererweiterung am PC durchführen kann.
Das SAS-System ist so komplex und die Anzahl der (vor allem amerikanischen) Handbücher sind so umfangreich, daß man sich mit der ersten Kenntnis einzelner weniger Anweisungen und Prozeduren am Anfang begnügen, jedoch keine Detailkenntnisse des Gesamtsystems haben muß, um damit zu arbeiten.
Wie im einzelnen die Statistikprozeduren und Programmstatements (in der Sprache von SAS heißen sie „Jobs“) aus den Rohdatenmengen aufgebaut werden, soll hier nicht dargestellt werden, sondern es werden im folgenden die verwendeten Analysen kurz beschrieben:

---

[1] Aus der Arbeit des Didaktischen Zentrums, Abt. Unterrichtsforschung, der Pädagogischen Hochschule Ludwigsburg

Mit SAS wurden vor allem folgende Statistikprozeduren gerechnet:

- einfache Frequenzenanalysen (PROC SORT, PROC LIST, PROC FREQ, CROSSTABS) mit Statistikkennwerten (statistic options) wie „expected“ (Ausgabe der unter der Hypothese der Unabhängigkeit erwarteten Frequenzen), „deviation“ (Ausgabe der Abweichung der beobachteten Frequenzen vom Erwartungswert), „chisq“ (Chi-square-Test auf Homogenität bzw. Unabhängigkeit für jede Schicht, einschließlich phi und Cramer’s V als generelle Assoziationsmaße).
- Mittelwertberechnung (PROC MEANS). Um die Gesamtstatistik für die relevaten Variablen zu bekommen, werden die gewünschten Variablen in einem VAR-Statement aufgelistet. PROC MEANS druckt automatisch die Anzahl der Beobachtungen aus, den Mittelwert, die Standardabweichungen, das Minimum und Maximum für jede numerische Variable.
- Korrelationsanalysen (PROC CORR) für die Pearsonschen-Produkt-Moment-Korrelationskoeffizienten r. (Die Abkürzung r ist auf das Wort Regression zurückzuführen. Die ersten Anwendungen des Korrelationskoeffizienten stammen von Francis Galton und Karl Pearson, die mit diesem Zusammenhangsmaß die Beziehung von Körperbaumaßen zwischen Eltern- und Kindergeneration untersuchten. Den Korrelationskoeffizienten erhalten wir, indem wir die Kovarianz (die Kovarianz ist durch den Mittelwert aller Produkte von korrespondierenden Abweichungen gekennzeichnet) zweier Variablen durch das Produkt der Standardabweichungen der Variablen dividieren.
  Die Korrelation zweier Variablen entspricht somit der Kovarianz der z-transformierten Variablen bzw. dem durchschnittlichen Produkt korrespndierender z-Werte.
- Faktorenanalysen (PROC FACTOR) Die Faktor-Prozedur ermöglicht eine Reihe faktorenanalytischer Verfahren sowie Hauptkomponentenanalysen; als Eingabedaten können Rohdaten-, Korrelations- und Kovarianzmatrizen verwendet werden.
  Neben unrotierten Lösungen können verschiedene orthogonale und oblique Rotationslösungen bestimmt werden. Ferner können Koeffizienten ausgegeben werden, die – mittels PROC SCORE – zur Faktorscore-Berechnung verwendet werden können. Faktorenscores können auch direkt berechnet und ausgegeben werden.

Die FACTOR-Anweisungen sind in SAS wie folgt gegliedert:

PROC FACTOR	Optionen (Methodenoptionen, Optionen für die Kommunalitätenschätzung, Optionen zur Steuerung der Anzahl der Faktoren, Optionen für zusätzliche Statistiken und Druckausgaben)
PRIORS	Kommunalitäten
VAR	Variablen
Partial	Variablen
FREQ	Variable
WEIGHT	Variable (gewichtete Korrelationsmatrix)
BY	Variable(n)

SAS wird momentan auf Großrechnern sowie am PC unter DOS eingesetzt. Seit kurzem gibt es auch eine Windows-Version der Implementation von Relase 6.04 des SAS-Systems.
An der Pädagogischen Hochschule Ludwigsburg besteht im Didaktischen Zentrum, Abteilung Unterrichtsforschung, über einen Lizenzvertrag mit dem Universitätsrechenzentrum Heidelberg die Möglichkeit, sowohl die PC-Version als auch die Großrechner-Version per Internet-Verbindung (über das BELWUE Baden-Württemberg) zu verwenden.
Die SAS-Dateien und deren Anwendungen können zwischen den Rechnersystemen auch ausgetauscht werden.

Das SAS-Produkt auf dem PC gliedert sich in mehrere unabhängige Teilpakete/Module:
- SAS/Base
SAS/Base beinhaltet die grundlegenden Funktionen, die für alle weiteren SAS-Anwendungen Voraussetzung sind. Zudem werden einfache deskriptive Prozeduren (z.B. SUMMARY, FREQUENCY) und Utilityprozeduren (z.B. Prozedur COPY oder SORT) angeboten.
- SAS/STAT
SAS/STAT bietet alle wesentlichen Statistikprozeduren wie z.B. für Virianzanalyse (ANOVAS und GLM) oder Regressionsanalyse (REG).
- SAS/IML
SAS/IML ermöglicht die Verarbeitung von SAS-Dateien im Matrizenkalkül. Dieses Produkt ist auch interessant, um Algorithmen im Matrixkalkül zu besprechen: z.B. lineare Modelle in der Statistik.
- SAS/FSP
SAS/FSP dient dem maskengesteuerten Editieren von SAS-Dateien. Mit der darin enthaltenen Prozedur FSEDIT können eigene Eingabefelder für SAS-Dateien geschrieben werden, welche zur Dateneingabe benutzt werden. Zudem können dort Plausibilitätsprüfungen der Eingabewerte vorgenommen werden.
- SAS/GRAPH
SAS/GRAPH ermöglicht die Erstellung hochauflösender Graphik in SAS/PC-Umgebung. Hierzu sind allerdings zusätzliche Hardwarekomponenten notwendig.
- SAS/AF
SAS/AF ermöglicht das Schreigen von menügesteuerten Lernprogrammen und Endbenutzermenüs in SAS/PC-Umgebung.

Alle hier beschriebenen Produkte sind im SAS/BASE-Rahmen anwendbar, der eine benutzerfreundliche Arbeitsumgebung in SAS/PC zur Verfügung stellt.

2. SPSS
Für die Mehrzahl der Correlationsanalysen, bei T-Test- und F-Test-Analysen, bei der Berechnung des Reliabilitätskoeffizienten und für einige Faktorenanalysen wurde das Statistikkprogramm SPSS verwendet.
SPSS ist die Abkürzung für Superior Performing Statistical Software, früher für Statistical Package for the Social Sciences. Es ist das am weitesten verbreitete Computer-Programmpaket für die Datenverarbeitung und Datenanalyse.
„Programmpaket“ bedeutet dabei, daß SPSS in einem einzigen integrierten System alle beim Datenverarbeitungsprozeß notwendigen Schritte anbietet.
SPSS ist auch deshalb ein integriertes System, weil es eine Vielzahl von Computerprogrammen zur statistischen analyse sozialwissenschaftlicher Daten anbietet. Die besondere Leistung liegt nicht allein im Umfang und der Art seiner statistischen Prozeduren, sondern in der Einheitlichkeit und Flexibilität des Systems, das durch eine quasi-natürliche Steuersprache gehandhabt wird. Außerdem ist eine leicht verständliche Fehlerdiagnostik eingeschlossen.
SPSS ermöglicht neben statistischen Analysen eine Vielzahl von Datentransformationen, Datenselektionen und Dateimanipulationen und über eine ausführliche Dateibeschreibung die Erzeugung einer leicht lesbaren Druckausgabe. Ergebnisse, aber auch Zwischenergebnisse können am Ende jeder beliebigen Phase der Analyse von Auswertungsvorhaben auf verschiedenen Datenträgern, z.B. im Word-Format, gespeichert werden.
SPSS setzt wie SAS keine Kenntnisse einer Programmiersprache voraus. Damit wird auch der Anfänger in Datenverarbeitung der häufig beobachteten Abhängigkeit von EDV-Spezialisten enthoben und kann somit auch einen weiteren wichtigen Teilabschnitt seines Projekts, nämlich den der Auswertung, selbständig durchführen und kontrollieren.

SPSS wird seit nunmehr über 22 Jahren ständig ereweitert und erheblich verbessert. Es wird außer von Soziologen für die Auswertung von Umfragen auch von Psychologen, Medizinern, Biologen, Volks- und Betriebswissenschaftlern, Pädagogen u.v.a. an fast allen Hochschulen und Forschungsinstituten der Bundesrepublik Deutschland zur Auswertung ihrer Experimente und Untersuchungen verwendet.
Die wichtige Bedeutung von SPSS als zentrales Auswertungssystem für statistische Probleme findet ihren Niederschlag u.a. darin, daß andere Programmpakete z.B. CLUSTAN und das Datenbanksystem SIR SPSS-Dateien bearbeiten bzw. erstellen.
Die bisher aufgeführten Programmsysteme gibt es mittlerweile nicht nur für den Großrechner, es stehen auch Versionen für Personal Computer zur Verfügung.
Die ersteren Versionen sind die Programmversion SPSS-X 3.0, die PC-Version ist die Version 4.0. Seit ungefähr einem Vierteljahr bietet SPSS auch eine Windows-Version und eine neue Version 5.0 an.
Die bei unseren Analysen verwendete SPSS-Version ist die PC-Version SPSS/PC+, Version 4.0, die u.a. folgende Prozeduren enthält. Diese werden nach der Systematik innerhalb des Programmsystems übersichtsartig aufgeführt:

a) Descriptive Statistics:
- Frequencies
- Descriptives
- Corosstabs
- Means
- Examine

b) Reports and Tables:
- List
- Report
- Tables

c) Correlation and Regression:
- Correlations
- Regression – multiple lineare Regressionen

d) Comparing group Means:
- T-Test: Diese Prozedur testet durch die Berechnung von t-Werten nach Student, ob zwei Stichprobenmittelwerte signifikant verschieden sind. Dabei können zwei Typen von Texts durchgeführt werden:
    - Texts für unabhängige Stichproben und
    - Texts für abhängige Stichproben.

Für unabhängige Stichproben wird sowohl ein F-Test auf Varianzhomogenität als auch je ein t-Test für homogene und inhomogene Varianzen durchgeführt. Die Irrtumswahrscheinlichkeit wird für die zweiseitige Fragestellung ausgegeben; teilt man diese durch 2, erhält man die für die einseitige Fragestellung.
Für abhängige Stichproben wird neben den Verteilungskennwerten für Differenzwerte und dem T-Wert mit zweiseitigem Signifikanzniveau auch die Korrelation der Variablen ausgegeben.

- ANOVA rechnet Varianzanalysen für faktorielle Versuchspläne mit maximal 10 Faktoren (unabhängigen –Variablen)
- ONEWAY berechnet einfaktorielle Varianzanalysen (jeweils nur eine unbhängige Variable, Trends über die Kategorien der unabhängigen Variablen)

- MANOVA (diese Prozedur berechnet alle im allgemeinen im linearen Modell üblichen Signifikanzen und Parameterschätzungen)

f) Classification and Clustering
Hierzu zählen folgende Analysen:
- Facotor Analysis
- Cluster (verschiedene agglomerative hierarchische Clusteranalysen)
- Discriminant (diese Analyse wird angewandt, wenn einerseits ein Satz von quantitativen unabhängigen Variablen (= beschreibende oder diskriminierende Variablen) für jeden Fall erhoben wurde, und wenn andererseits eine damit zusammenhängende Klassifikationsvariable oder gleichwertig eine Gruppeneinteilung der Fälle vorliegt)

g) Reliability
Diese Prozedur dient zur Durchführung von Itemanalysen bei Tests und Fragebögen und ermöglicht die Berechnung gebräuchlicher Reliabilitätskoeffizienten auf der Basis der sogenannten „Klassischen Testtheorie“. Bei dieser Analyse berechnet man für die einzelnen Fragen verschiedene Itemkennwerte, u.a. Itemschwierigkeiten und –trennschärfen, z.B. um den Zusammenhang zwischen einzelnen Items oder Aufgaben mit dem Gesamtwert zu ermitteln. Das Testgütekriterium Reliabilität gibt Auskunft darüber, wie genau ein Merkmal durch den Test oder Fragebogen erfaßt wird.
Bei homogenen Tests fällt die Reliabilität umso höher aus, je mehr Items ein Test oder Fragebogen enthält. Mit der Prozedur Reliability kann eine Schätzung der Testreliabilität mit der Methode der Testhalbierung (split-half-Reliabilität) und über die Berechnung der inneren Konsistenz (Testhomogenität) vorgenommen werden.
Außerdem berechnet RELIABILITY alle üblichen Itemanalyse-Kennwerte wie z.B. Item-Schwierigkeiten, Trennschärfen, Interkorrelationen, aber auch spezielle Tests wie zweifaktorielle Varianzanalysen, T^2-Test auf Gleichheit der Mittelwerte bei Meßwiederholungsdesigns.

An der Pädagogischen Hochschule Ludwigsburg wird zur Zeit die SPSS/PC+ Version 4.0 verwendet. Hierzu müssen folgende Voraussetzungen erfüllt sein:

- Betriebssystem PC-DOS oder MS-DOS Version 2.0 oder höher
- Mindestens 512 KB RAM, falls nur das BASE-System installiert wird
- Mindestens 640 KB RAM, falls mehr als das BASE-System installiert wird
- Extended Memory wird nicht unterstützt
- Expanded Memory wird bis zu 64 KB von einem Teil der Prozeduren unterstützt
- Festplatte mit mindestens 4,5 Megabyte freiem Speicherplatz (BASE-System). Der von SPSS/PC+ benötigte Speicherplatz ist abhängig von den installierten Modulen.

Wenn alle Module installiert sind, sollte man einen Gesamtspeicherplatz von ca. 20 MB zur Verfügung haben. Für SPSS/PC+BASE, Statistics, Advanced Statistics und Tablex benötigt man zusammen ca. 9,7 MB Speicherplatz.
Die Module umfassen im einzelnen u.a. folgende Prozeduren:
- SPSS/BASE enthält u.a. die Prozeduren AGGREGATE, PLOT, SORT, RANK, CROSSTABS, EXAMINE
- SPSS/STATISTICS umfaßt u.a. die Prozeduren CLUSTER, FACTOR, RELIABILITY, ANOVA, T-TEST, CORRELATION, REGRESSION
- SPSS/ADVACED STATISTICS ermöglicht u.a. die Prozeduren LOGISTIC REGRESSION, DISCRMINANT, SURVIVAL, HILOGLINEAR, MANOVA
- SPSS/TABLES, PRINT TABLES

- DATA ENTRY enthält die Prozedur DE, mit der sich beliebige Dateien eingeben und verändern lassen.
- SPSS/TRENDS enthält die Prozeduren CREATE, READ, MODEL, CASEÖLOT, CURVEFIT, SEASON u.a.
- SPSS/CATEGORIES ermöglicht die Prozeduren CONJOINT, ORTHOPLAN, ANACOR, PRINCALS u.a.

Das zentrale System mit den Prozeduren DATA LIST, GET, SAVE, DESCRIPTIVES, FREQUENCIES, MEANS, TRANSLATE u.a. muß immer installiert sein. Die anderen Teilmodule können aber installiert oder weggelassen werden.

Literatur

Bundschuh, W. und Hefner, J. (1987): Statistik der SAS-Prozeduren. Heidelberg: Skript des Universitätsrechenzentrums

Frenzel, G. (1987): Elementare Statistik mit SAS und SPSS. Heidelberg, Skript des Universitätsrechenzentrums.

Gogolok, J., Schuermer, R. und Ströhlein, G. (1992): Datenverarbeitung undf statistische Auswertung mit SAS. Stuttgart: Gustav Fischer Verlag. Band I: Einführung in das Programmsystem, Datenmanagement und Auswertung. Band II: Komplexe statistische Analyseverfahren.

Heller, K. und Rosemann, B. (1981): Planung und Auswertung empirischer Untersuchungen. Wissenschaftsmethodik und Forschungsstatistik für Pädagogen, Psychologen und Soziologen. Stuttgart: Klett.

Schubö, W., Uehlinger, H.-M., Perletz, Chr., Schroger, E. und Sierwald, W. (1991): SPSS. Handbuch der Programmversionen 4.0 und SPSS-X 3.0. Autorisierte deutsche Bearbeitung des SPSS Reference Guide. Stuttgart: Gustav Fischer.

Voss, W. (1988): Statistische Methoden und PC-Einsatz. Opladen: Leske und Budrich.

Zöfel, Peter (1988[2]) Statistik in der Praxis. Stuttgart: UTB.