Statistik mit SAS

Von
Professor Dr. Andreas Pfeifer
und
Dipl.-Math. Marco Schuchmann

R. Oldenbourg Verlag München Wien

Die Informationen in dieser Dokumentation wurden mit größter Sorgfalt erstellt. Trotzdem können Fehler nicht ausgeschlossen werden. Für fehlerhafte Angaben und deren Folgen werden weder juristische Verantwortung noch irgendeine Haftung übernommen. Für eine Mitteilung eventueller Fehler sind die Autoren dankbar.

Das Buch ist nach den neuen Rechtschreibregeln (vgl. Duden, Rechtschreibung der deutschen Sprache, Bd. 1, 21. Aufl. 1996) abgefasst.

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Pfeifer, Andreas:
Statistik mit SAS / von Andreas Pfeifer und Marco Schuchmann. - München ; Wien : Oldenbourg, 1997
ISBN 3-486-23953-8
NE: Schuchmann, Marco:

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0, Internet: http://www.oldenbourg.de

Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: R. Oldenbourg Graphische Betriebe GmbH, München

ISBN 3-486-23953-8

Vorwort

Dieses Buch gibt einen Einstieg in SAS. Zielsetzung dieser Einführung ist es, den Leser zu befähigen, selbständig statistische Auswertungen mit SAS durchführen zu können. Alle in diesem Buch beschriebenen Möglichkeiten werden durch kommentierte Beispiele und SAS-Ergebnisausgaben dargestellt. Dieses Buch eignet sich sowohl zum selbständigen Einstieg in SAS, als auch als Begleitbuch zu Kursen und Seminaren. Wir beschreiben in diesem Buch die Durchführung statistischer Analysen mit der SAS-Programmierung. Somit sind alle Verfahren unabhängig vom Betriebssystem bzw. der Hardwareumgebung durchführbar. Im ersten Kapitel beziehen wir uns zwar auf die Version 6.11 von SAS, die beschriebenen Programme laufen aber auch in früheren Versionen.

Natürlich können in dieser kompakten Dokumentation nicht alle Möglichkeiten von SAS für Windows umfassend beschrieben werden; dafür wird auf die umfangreichen Original-Handbücher von SAS verwiesen. Jedoch werden die wichtigsten Eigenschaften ausführlich mit Beispielen und Ergebnisinterpretationen erläutert. Ebenso werden der statistische Hintergrund dieser Befehle und die notwendigen Voraussetzungen der benutzten Verfahren dargestellt.

Kapitel 1 schneidet die Problematik bei der Anwendung von Statistik-Software an und gibt eine Übersicht über die Möglichkeiten von SAS für Windows. Im zweiten Kapitel wird die Planung einer empirischen Studie bis zur Eingabe der Daten in SAS anhand einer Fragebogenauswertung geschildert. Kapitel 3 beschreibt das Datenmanagement mit SAS. Wie Sie zu ersten deskriptiven Statistiken und Grafiken gelangen, wird in den Kapiteln 4 und 5 beschrieben.
Kapitel 6 geht auf statistische Tests näher ein. Im nächsten Kapitel werden einfache Methoden der schließenden Statistik erläutert. Kapitel 8 geht dann auf komplexere Methoden der Statistik, wie Varianz-, Kovarianz-, Faktoren- oder Clusteranalysen ein. Im Kapitel 9 zeigen wir Möglichkeiten, auch kategorielle Variablen, die oft bei der Auswertung von Fragebögen auftreten, mit linearen und loglinearen Modellen der kategoriellen Datenanalyse auszuwerten und diese Ergebnisse einfach darzustellen.

Den Anhang bilden ein Verzeichnis englischer Ausdrücke, eine Tabelle der benutzten mathematischen Zeichen und Abkürzungen sowie das Literaturverzeichnis.

Zu dem Buch gibt es eine Diskette mit Daten und Programmen. Weitere Informationen dazu stehen im Anhang D.

Wir danken Herrn Dipl.-Math. Werner Sanns für die Durchsicht des Manuskripts und für Korrektur- und Verbesserungsvorschläge.

Andreas Pfeifer, Marco Schuchmann

Inhaltsverzeichnis

1 Übersicht

1.1 Vorsicht bei Statistik-Programmpaketen

Bei statistischen Auswertungen und Datenanalysen werden Statistik-Programmsammlungen eingesetzt, um die benötigten statistischen Verfahren nicht selbst programmieren zu müssen. Sie brauchen für eine Datenanalyse keine Kenntnisse in einer Programmiersprache und auch kaum Kenntnisse der statistischen Verfahren, die Sie verwenden wollen.

Statistik-Programmsammlungen verhalten sich wie ein "schwarzer Kasten". Auf der einen Seite kommen die Daten und wenige Steueranweisungen hinein, auf der anderen Seite erhalten Sie die fertigen Ergebnisse. Dabei können Sie viel falsch machen, wenn Sie nicht ungefähr wissen, was der schwarze Kasten macht. Dies soll an einem einfachen Beispiel verdeutlicht werden:

Jemand möchte testen, ob er übernatürliche Fähigkeiten besitzt. Dazu geht er folgendermaßen vor:

Zunächst lässt er fünfzig Personen mit einem ganz normalen, symmetrischen Würfel würfeln. Dann nimmt er die Würfelergebnisse von zehn Personen, die besonders wenig Augenzahlen hatten:

1, 1, 1, 1, 1, 2, 1, 1, 1 und 2. (Mittelwert: 1,2)

Diese Personen bittet er nun zu sich und spricht zu ihnen einen Zauberspruch. Er behauptet, dass wenn diese Personen jetzt nochmals würfeln, sie kein solches Pech mehr haben. Folgende Ergebnisse liegen nach dem zweiten Würfeln dieser zehn Personen vor:

4, 2, 6, 2, 1, 5, 3, 5, 6 und 3. (Mittelwert: 3,7)

Jetzt testet er mit einem Statistikprogramm, ob durch den Zauberspruch die Augenzahl erhöht wurde; genauer ausgedrückt, er gibt die beiden obigen Zahlenreihen mit den Würfelergebnissen in den Computer ein und führt einen t-Test durch, um signifikante Mittelwertsunterschiede nachzuweisen, d.h. um nachzuweisen, dass er übernatürliche Fähigkeiten besitzt. Diesen Test kann er mit dem Statistikpaket SAS durchführen. Der Test wird bei diesem Beispiel signifikante Unterschiede "mathematisch" bestätigen. Doch diese Vorgehensweise ist aus mehreren Gründen falsch:

Ein Fehler liegt darin, dass die Testpersonen nicht zufällig ausgewählt wurden. Es wurden nämlich nur diejenigen zum Testen gewählt, die schlechte Testergebnisse

(d.h. niedrige Augenzahlen beim erstmaligen Würfeln) hatten. Ein anderer Fehler kommt dadurch zustande, dass der t-Test als eine Voraussetzung benötigt, dass die beiden Stichproben aus normalverteilten Grundgesamtheiten stammen. Dies ist nicht gegeben.

An diesem Beispiel sollen Sie folgendes erkennen: Jedes Testverfahren benötigt gewisse Voraussetzungen. Wenn nun diese Voraussetzungen nicht erfüllt sind, dürfen Sie den Test nicht durchführen. Aber Statistik-Programmpakete überprüfen die Voraussetzungen nicht automatisch. Daher können Sie mit Programmpaketen in Statistik alles "beweisen", wenn Sie die Voraussetzungen nicht beachten. Bei dem Beispiel hier sieht sicherlich jeder, wo Fehler liegen. Im Allgemeinen ist es nicht so leicht, Fehler zu finden.

Fehler in den Voraussetzungen können und sollten aber auch von Nicht-Statistikern erkannt werden. Dazu ist es nicht notwendig, den theoretischen Hintergrund des benutzten Tests genau zu kennen. Sie sollten aber wissen, welche Voraussetzungen gebraucht werden, um den jeweiligen Test sinnvoll anzuwenden. Deshalb sind statistische Grundkenntnisse unbedingt erforderlich.

Grundkenntnisse in Statistik können beispielsweise durch das Studium der Bücher von Bortz (1993), Hartung (1995), Sachs (1992) oder Zöfel (1993) erworben werden. Bibliographische Angaben zu den Büchern sind im Anhang C zu finden.

1.2 SAS und andere Programme zur Datenanalyse

SAS hat einen modularen Aufbau und besteht aus einem Grundpaket (SAS Base bzw. SAS CORE) und zahlreichen Zusatzpaketen. Mit dem Statistikmodul STAT können fast alle Verfahren der Statistik (von Häufigkeitsauszählungen bis Regressions- und Varianzanalysen oder Faktorenanalysen) durchgeführt werden. Grafiken aller Art können mit dem Modul GRAPH erzeugt werden. Weiteres zum Aufbau von SAS folgt im nächsten und übernächsten Kapitel. Kommen wir nun zuerst zu den anderen Softwareprodukten zur Datenanalyse.

Es gibt eine Vielzahl anderer Programme für statistische Auswertungen, wie beispielsweise BMDP (Hersteller: BMDP), CSS (Hersteller: Statsoft), Danet-Statistik (DMB), Micro TSP (QMS), NCSS (Dr. Hintze /USA), PlotIT (ICS, S.P. Eisensmith), P-STAT (P-STAT), SPSS (SPSS GmbH), STASY-500 (PIC), Statgraphics (STSC), Systat (Systat, SPSS), Unscrambler (Camo).

Dies sind nur einige der vielen Statistikprogramme für Mikrocomputer. Die Qualität dieser Programme ist sehr unterschiedlich. Übersichten über die Leistungsfähigkeiten verschiedener Programme werden regelmäßig in Computerzeitschriften ver-

öffentlicht; doch sie veralten relativ schnell, da das Angebot an Software und an verschiedenen Versionen sehr stark wächst.

Aber nicht nur "reine" Statistikprogramme, sondern auch andere Software-Produkte können bei einer statistischen Auswertung sinnvoll angewandt werden. Die Software, mit deren Hilfe statistische Auswertungen durchgeführt oder unterstützt werden, kann grob in vier Gruppen eingeteilt werden:

- Tabellenkalkulationsprogramme,
- Datenbankprogramme,
- Grafikprogramme und die "eigentlichen"
- Statistikprogramme.

Es gibt eine Vielzahl sogenannter "integrierter" Software-Pakete, in denen mehrere der oben genannten Gruppen integriert sind. Oftmals können z.B. auch mit einem Tabellenkalkulationsprogramm Grafiken erstellt werden. Eine klare Einordnung eines Software-Produktes in eine der vier Gruppen lässt sich deshalb nicht immer durchführen. Trotzdem ist diese Gruppeneinteilung sinnvoll, um einen Überblick über die vorhandene Software, die sich auch für statistische Auswertungen eignet, zu geben.

Tabellenkalkulationsprogramme dienen zum mühelosen Erstellen von Tabellen, Berichten und Statistiken. Sie ermöglichen es, eine Vielzahl von aufeinander bezogenen Rechenvorgängen ablaufen zu lassen. Das kann sicherlich auch mit einem Taschenrechner bewältigt werden. Das Besondere an einem Tabellenkalkulationsprogramm besteht aber darin, dass es zwischen dem Rechenweg und den eingegebenen Zahlen unterscheidet. Der große Vorteil liegt darin, dass - sobald eine Zahl geändert wird - automatisch alle nachfolgenden Rechenschritte mit dem neuen Wert ausgeführt werden. Der Umgang mit einem Tabellenkalkulationsprogramm ist sehr einfach. Bezüglich statistischer Auswertungen können jedoch meist nur wenige Kenngrößen - wie beispielsweise Mittelwerte und Standardabweichungen - berechnet werden.

Datenbank-Software dient hauptsächlich dazu, Datenbestände zu erstellen, zu verwalten und geeignet auszugeben. Auch komplexe Datenstrukturen können bei solchen Programmen im Gegensatz zu Statistik-Paketen berücksichtigt werden.

Grafikprogramme dienen zwar hauptsächlich zum Erstellen von Zeichnungen. Allerdings lassen sich auch manchmal einige Statistiken (Prozentzahlen, Häufigkeiten oder lineare Regressionen) erzeugen. Für statistische Auswertungen sind diese Programme jedoch nur eingeschränkt tauglich.

1.3 Was leistet SAS?

SAS bietet mit seinen Modulen sehr viele Möglichkeiten in diversen Bereichen Daten auszuwerten. Das Grundpaket BASE bietet u.a. die Möglichkeit, Datenmanagement zu betreiben. Hiermit können SAS-DATASETS (so nennt SAS seine Datendateien) professionell und ähnlich wie mit einem Datenbanksystem verwaltet werden. Mit der zu BASE gehörenden Prozedur SQL können sogar SQL-Abfragen durchgeführt werden. Außerdem können mit dem BASE-Modul einfache statistische Auswertungen, wie Mittelwertsberechnungen bzw. t-Tests (Prozedur MEANS), Häufigkeitsauszählungen (Prozedur FREQ) oder auch einfache Plots (Prozedur PLOT) erstellt werden.

Interessant in Bezug auf statistische Auswertungen ist das SAS-Modul STAT. Mit ihm können die meisten bekannten Verfahren in der Statistik durchgeführt werden. Hierzu zählen univariate Statistiken (Prozedur UNIVARIATE), Regressionsanalysen (Prozedur REG) und Varianzanalysen (Prozeduren ANOVA bzw. GLM) bis hin zu kategoriellen Modellen (Prozedur CATMOD) oder Faktorenanalysen (Prozedur FACTOR) und viele mehr. Unter Verwendung der SAS-Programmierung können zu jeder Prozedur auch eine Vielzahl von Optionen verwendet werden, um zusätzliche Statistiken zu erhalten. Die in diesem Buch beschriebenen Verfahren beziehen sich zum größten Teil, bis auf wenige Ausnahmen, auf das Modul STAT.

Zur Erstellung von Grafiken dient das Modul GRAPH. Innerhalb dieses Moduls stellt SAS eine Reihe von Prozeduren zur Verfügung, mit denen u.a. Balken-, Kreis- und x-y-Diagramme erstellt werden können.

Weitere SAS-Module sind: Der ASSIST, der Standardauswertungen mit einem Menü ohne Programmierung ermöglicht; das Modul OR für Optimierungsprobleme im Bereich des Operations Research; das Modul QC für statistische Qualitätskontrollen; das Modul INSIGHT für interaktive Auswertungen; das Modul IML für Matrizenoperationen, sowie die Module ACCESS, CALC, CONNECT, CPE, ETS, FSP, IMS-DL, LAB, PH, TOOLKIT, TUTOR und mehr.

Damit SAS für Windows verwendet werden kann, müssen folgende Voraussetzungen erfüllt sein:

- Dos 5.0 oder höher
- Windows 3.1 oder höher
- 80386-Prozessor oder höher
- Mindestens 4 Megabyte Arbeitsspeicher (RAM)
  Zum sinnvollen Arbeiten mit SAS sind mindestens 8 Megabyte zu empfehlen.
- Festplatte mit mindestens 80 Megabyte bis zu 300 Megabyte freiem Plattenplatz, je nachdem wie viele SAS-Module Sie installieren.
  Mit 80 Megabyte können Sie nur das Grundmodul und STAT installieren.
- 3 1/2" Diskettenlaufwerk hoher Dichte oder CD-ROM-Laufwerk

- Grafikadapter mit Mindestauflösung von 640 × 480 (VGA)
- Microsoft Win32s. Dieses System wird von der SAS-Setup-Prozedur automatisch installiert.

Weitere Informationen zu SAS erhalten Sie beim SAS Institute in Heidelberg.

1.4 Der erste Einstieg in SAS

SAS bietet dem Benutzer mit dem ASSIST die Möglichkeit, eine Auswertung menügesteuert durchzuführen. In unserem Buch werden wir aber alle Verfahren über die SAS-Programmierung realisieren, da diese dem Benutzer erstens mit zusätzlichen Optionen viel mehr Variationsmöglichkeiten bietet und einiges aus unserem Buch mit dem ASSIST nicht zu realisieren wäre, wie beispielsweise ein großer Teil unseres Kapitels über das Datenmanagement. Zweitens benötigen Sie die Programmierung, falls Sie bestimmte Auswertungen standardisieren möchten. Weiterhin sind die Programme mit drei bis fünf Zeilen sehr kurz, so dass Sie diese bei einer Auswertung schnell eingeben können.

Wenn Sie nun SAS starten, erhalten Sie den folgenden Anfangsbildschirm:

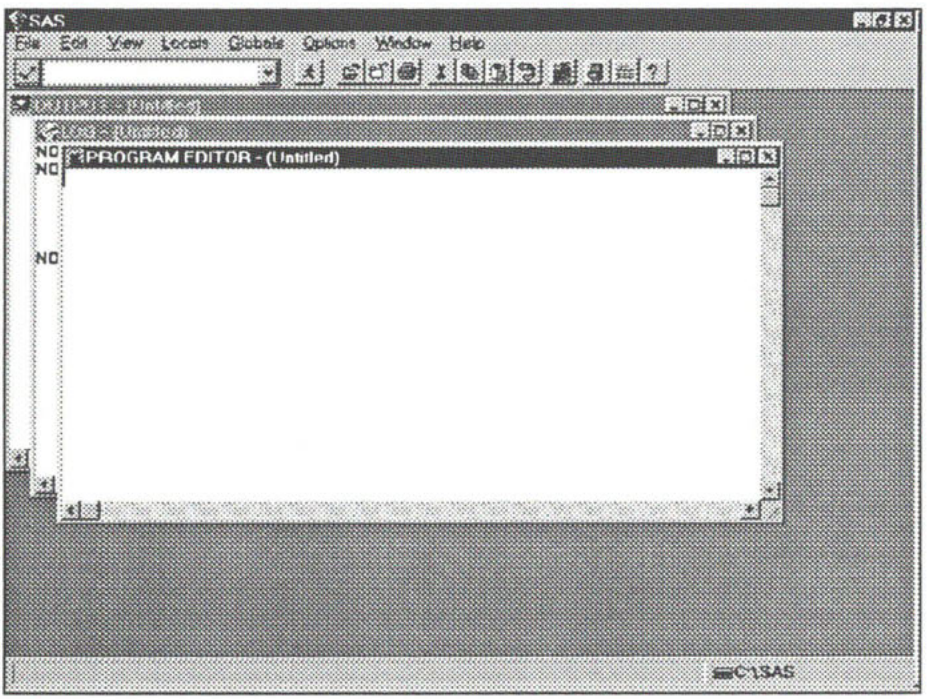

Abb. 1-1

Wie Sie sehen können, ist dieser in drei Fenster geteilt. Das erste Fenster ist der PROGRAM-Editor. In diesem werden die Programme eingegeben. Um ein Programm zu starten, müssen Sie die Taste <F8> drücken oder auf die Ikone in der Leiste mit dem Männchen klicken.

Das nächste Fenster ist das LOG-Fenster. In diesem werden von SAS Kommentare nach jeder Programmausführung geschrieben. U.a. werden hier die Programmzeilen wiederholt. Falls Sie ein SAS-Programm ausführen lassen und ein Fehler auftritt, wird dies in diesem Fenster mit einem Kommentar angezeigt. Fehlermeldungen erscheinen in Rot. Warnungen werden in der Farbe Grün ausgegeben. Einige Tippfehler kann SAS auch erkennen und interpretiert die Anweisung korrekt. In diesem Fall gibt SAS eine Warnung aus.

Im dritten Fenster, dem OUTPUT-Fenster, wird die jeweilige Ergebnisausgabe angezeigt. Sollte ein Prozedur fehlerfrei sein, öffnet SAS automatisch das OUTPUT-Fenster, falls die entsprechende Prozedur eine Ausgabe liefert.

Zwischen den Fenstern können Sie wechseln, indem Sie entweder mit der Maustaste in das entsprechende Fenster wechseln oder unter dem Menüpunkt WINDOW das entsprechende Fenster wählen. Eine weitere Möglichkeit zum Wechseln zwischen den Fenstern bieten die Funktionstasten. In der Regel gelangen Sie mit der Taste <F5> in den PROGRAM-Editor, mit der Taste <F6> in das LOG-Fenster und mit der Taste <F7> in das OUTPUT-Fenster.

Um die Tastenbelegung der Funktionstasten zu erfahren oder zu verändern, wählen Sie unter dem Menüpunkt HELP den Punkt KEYS. Sie erhalten dann das folgende Fenster:

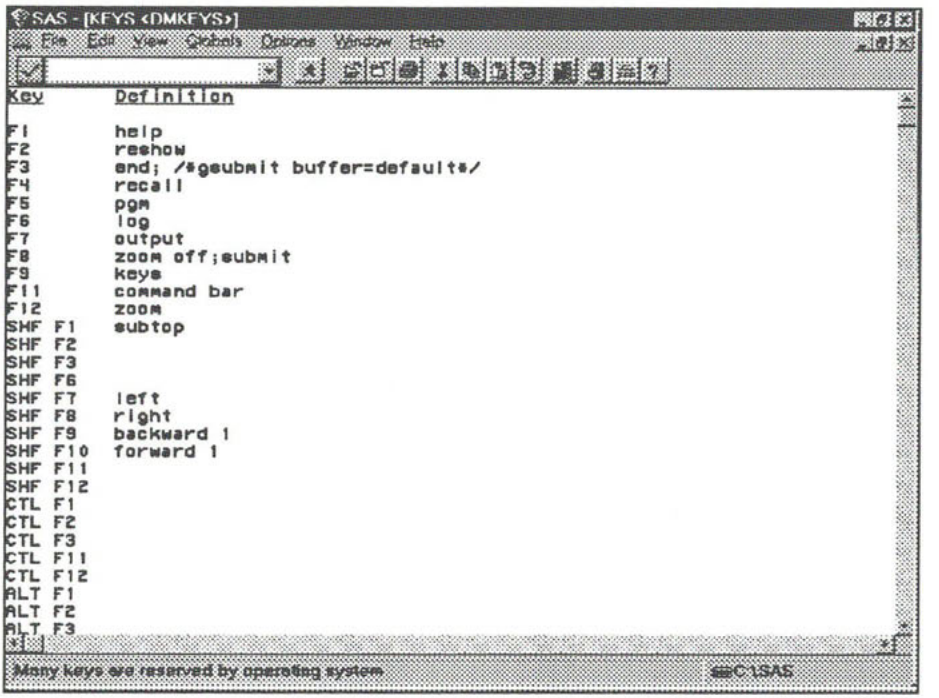

Abb. 1-2

Sie können hier nun selbst Funktionstasten mit Anweisungen belegen oder die bestehenden Anweisungen umbelegen. Wie Sie aus Abb. 1-2 ersehen, ist mit der Taste <F1> das Hilfemenü aufzurufen. Eine weitere wichtige Taste ist die Taste <F8>. Mit ihr können Sie ein SAS-Programm, das im PROGRAM-Editor steht, starten. Nach dem Start ist das SAS-Programm nicht mehr sichtbar, der PROGRAM-Editor ist also leer. Mit der Taste <F4> (RECALL) ist es möglich, das Programm wieder in den PROGRAM-Editor zu laden.

Mit den Tasten <F5> bis <F7> wechseln Sie, wie bereits erwähnt, zwischen den Fenstern. Mit der Taste <F12> (ZOOM) vergrößern Sie das Fenster, in dem Sie sich befinden.

Dieses Fenster können Sie, wie in Windows üblich, durch Klicken in die obere linke Ecke (bei Windows 3.1 bzw. 3.11 Doppelklicken) oder mit der Taste <F8> schließen.

Alle Programme, die Sie in SAS schreiben und speichern, werden von SAS im ASCII-Format gespeichert. Ebenso jede Ergebnisausgabe, die Sie speichern.

Speichern können Sie jeweils das aktuelle Fenster mit dem Menüpunkt FILE, dann SAVE AS (oder beim 2. Speichern SAVE) und WRITE TO FILE ... .

Wenn Sie in SAS eigene Dateien (DATASETS) speichern und verwalten wollen, müssen Sie nach jedem Programmstart dem System mitteilen, in welchem Verzeichnis diese stehen. SAS legt dies mit sogenannten LIBREFS (Referenzen für Libraries) fest. Vom System werden automatisch jeweils die Referenzen SASUSER und WORK (für temporäre Dateien) festgelegt. Die SAS-DATASETS, die sich in dem zu SASUSER gehörenden Verzeichnis befinden, können Sie sich ansehen, falls Sie in der linken Zeile, auf der Symbolleiste, das Wort LIB eingeben. SAS öffnet danach ein Fenster und zeigt Ihnen die LIBREFS an. Wenn Sie z.B. auf SASUSER doppelklicken, zeigt SAS Ihnen alle SAS Dateien im zu SASUSER gehörenden Pfad.

Weiterhin gibt es eine SAS-Prozedur mit dem Namen DATASETS, mit der Sie sich im LOG-Fenster alle SAS-DATASETS in einem Verzeichnis ansehen können. Wir wollen uns nun alle SAS-DATASETS im zur LIBREF SASUSER gehörenden Verzeichnis ansehen. Geben Sie hierzu das folgende Programm im PROGRAM-Editor ein und drücken Sie die Taste <F8> oder die Ikone mit dem rennenden Männchen in der Symbolleiste:

```
PROC DATASETS LIBRARY = SASUSER;
RUN;
```

Im LOG-Fenster sehen Sie nun alle SAS-Dateien in dem entsprechenden Verzeichnis. Wie Sie sehen, endet jede SAS-Anweisung mit einem Semikolon. Zwischen Klein- und Großschreibung macht SAS keinen Unterschied.

Falls Sie einen Tippfehler beim Eingeben des obigen Programms gemacht haben, erscheint ein Hinweis im LOG-Fenster. Sie können dann in den PROGRAM-Editor wechseln, mit der Taste <F4> das Programm zurückladen und den Fehler korrigieren.

Um Ihre eigenen Dateien in einem extra Verzeichnis speichern zu können, empfehlen wir Ihnen, auf Ihrer Platte (mit Hilfe des Dateimanagers von Windows) ein eigenes Verzeichnis anzulegen. Legen Sie zum Beispiel das Verzeichnis C:\DATA an. Im nächsten Schritt weisen Sie in SAS diesem Verzeichnis einen LIBREF zu. Geben Sie hierzu die folgende Programmzeile ein und starten Sie es mit <F8>:

```
LIBNAME DISK 'C:\DATA';
```

Diese Programmzeile müssen Sie nach jedem Neustart von SAS eingeben und starten, da diese selbstdefinierten LIBREFS nur temporär angelegt werden. Mit dem LIB-Kommando können Sie sich vergewissern, ob die LIBREF DISK existiert.

Sobald Sie jetzt vor irgendeinem Datensatz 'DISK.' angeben, weiß das SAS-System, dass es auf das oben definierte Verzeichnis zugreifen soll. Statt des Namens DISK könnten Sie auch ein beliebiges, nicht von SAS reserviertes, maximal achtstelliges Wort verwenden.

Mit einigen SAS Kommandos können Sie auch die Ergebnisausgabe (den Output) formatieren. Z.B. können Sie mit dem folgenden Kommando (dieses müssen Sie in der ersten Zeile Ihres Programm in den PROGRAM-Editor schreiben) die Zeilen- und Spaltenzahl definieren:

```
OPTIONS LINESIZE = a  PAGESIZE = b;
```

Hierbei müssen Sie für a eine Zahl zwischen 64 und 256 und für b eine Zahl zwischen 15 und 32767 einsetzten. Mit LINESIZE geben Sie an, wie viele Zeichen in einer Zeile der Ergebnisausgabe maximal stehen dürfen. Mit PAGESIZE wird die maximale Anzahl von Zeilen pro Seite festgelegt.

Mit der OPTIONS-Anweisung können Sie eine Reihe weiterer Parameter definieren. SAS zählt u.a. die Seitenzahl im OUTPUT-Fenster automatisch mit. Wenn Sie in der obigen Anweisung die Option PAGENO = 1 einfügen, wird die Seitennummer auf 1 gesetzt. Die Ergebnisausgabe beginnt dann wieder mit der Seitenzahl 1. SAS löscht das OUTPUT-Fenster sowie das LOG-Fenster nämlich nicht automatisch. Die Ergebnisausgabe, die Sie innerhalb einer SAS-Sitzung erzeugen, wird jeweils nach dem Vorhergehenden im OUTPUT-Fenster hinzugefügt. Falls Sie das OUTPUT- oder das LOG-Fenster löschen wollen, müssen Sie nur mit der rechten Maustaste in das entsprechende Fenster klicken, und Sie erhalten ein Menü:

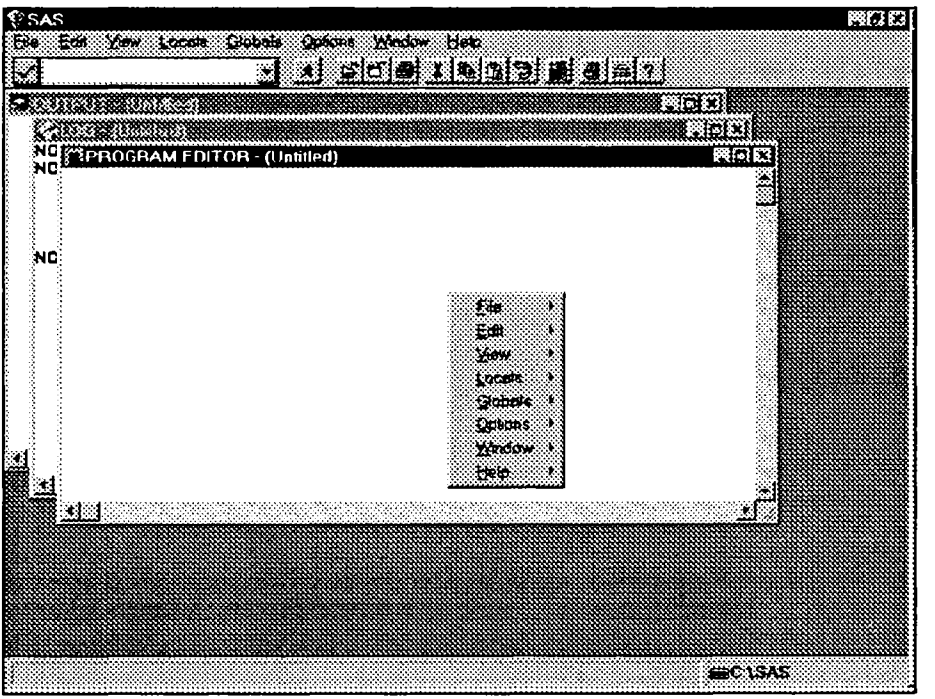

Abb. 1-3

Wenn Sie hier unter dem Menüpunkt 'Edit' 'Clear Text' auswählen, löscht SAS das komplette Fenster. Den Punkt 'Edit' finden Sie außerdem in der oberen Menüleiste. Die folgenden beiden Programmzeilen, die Sie zu Beginn eines Programms eingeben müssen, bieten Ihnen auch die Möglichkeit, das OUTPUT- bzw. LOG-Fenster zu löschen:

```
DM 'CLEAR OUT';
DM 'CLEAR LOG';
```

DM steht hier für Display Manager.

Innerhalb eines SAS-Programms können Sie auch Kommentarzeilen einfügen, die von SAS unbeachtet bleiben. Hier bietet SAS die folgenden beiden Möglichkeiten:

```
/* KOMMENTAR */

* KOMMENTAR;
```

Ein Kommentar kann, wie auch andere SAS-Befehle, über mehrere Zeilen hinweg verlaufen. SAS-Befehle enden jeweils, wie die zweite Kommentierungsmöglichkeit, mit einem Semikolon. Die erste Kommentierungsmöglichkeit endet mit '*/', so dass Sie mit dieser sogar mehrere Programmzeilen (eine Programmzeile ist der jeweilige Programmcode, der mit einem Semikolon endet) 'auskommentieren' können.

Möchten Sie, dass in der Ergebnisausgabe auf jeder Seite eine Überschrift erscheint, müssen Sie im PROGRAM-Editor die folgende Anweisung schreiben:

```
TITLE  'ÜBERSCHRIFT';
```

In der SAS-Programmierung wird zwischen zwei Arten unterschieden. Die erste besteht aus sogenannten DATASTEPS. Hiermit können SAS-DATASETS (SAS-Datenfiles) gespeichert oder verändert werden. Sie beginnen immer mit der DATA-Anweisung. Z.B. speichern die folgenden Anweisungen einen SAS-DATASET mit einer Datenzeile und einer Variablen X, die den Wert 5 hat, in der temporären Datei mit dem Namen DATEIX:

```
DATA DATEIX;
X = 5;
OUTPUT;
RUN;
```

Hätten wir anstelle DATEIX die Zeile DISK.DATEIX oder SASUSER.DATEIX geschrieben, so hätte SAS den DATASET DATEIX in dem entsprechenden Verzeichnis permanent gespeichert. Dieser Datensatz wird nun aber nur temporär im sogenannten WORK gespeichert und ist bei der nächsten SAS-Sitzung (nach dem nächsten Programmstart) wieder verschwunden. Während einer SAS-Sitzung stehen die temporären Dateien im Unterverzeichnis SASWORK des SAS-Verzeichnisses. Anstatt nur DATEIX zu schreiben, können Sie auch WORK.DATEIX schreiben, um Dateien temporär zu speichern, was natürlich umständlicher ist.

Die zweite Art besteht aus den sogenannten PROC-STEPS. SAS bietet hier für jede „Aufgabe" (z.B. Berechnung von univariaten Statistiken mit der Prozedur UNIVARIATE oder die Sortierung eines DATASETS mit der Prozedur SORT) eine Prozedur an. Eine der grundlegenden Prozeduren in SAS ist die Prozedur PRINT. Die Prozedur PRINT gibt einen SAS-DATASET im OUTPUT-Fenster aus (hier den temporären SAS-DATASET mit dem Namen DATEIX):

```
PROC PRINT DATA = DATEIX;
RUN;
```

Am obigen Beispiel sehen Sie den Aufbau der Prozeduren. Sie beginnen immer mit PROC und dem Namen der Prozedur. Danach folgt mit DATA = der Name des

DATASETS, den SAS beim Ausführen der Prozedur verwenden soll, und danach das Ende der Prozedur, die RUN- oder auch QUIT-Anweisung. Dazwischen können oft noch eine Reihe von Anweisungen und Optionen angegeben werden.

Wenn Sie zwischen der PROC- und der RUN-Anweisung die Zeile

```
VAR ...;
```

einfügen, können Sie eine Reihe von Variablen (mit Leerzeichen) getrennt angeben, die ausgegeben werden sollen. Falls Sie die Ausgabe gruppieren wollen (nach einer Variablen trennen, z.B. Männer und Frauen getrennt auflisten), können Sie dies tun, indem Sie die Zeile

```
BY ...;
```

einfügen. Hierbei ist zu beachten, dass Sie den DATASET zuvor nach der unter BY angegebenen Variablen sortieren müssen. Wie Sie dies tun können, wird im Kapitel über das Datenmanagement beschrieben.

Wenn Sie die Zeile

```
SUM ...;
```

einfügen, werden zusätzlich unter der oder den Variablen, die in der SUM-Anweisung aufgelistet (getrennt wieder mit Leerzeichen) sind, die Summe über alle Werte der Variablen im OUTPUT-Fenster ausgegeben.

Falls Sie die Ergebnisse aller Prozeduren direkt in eine Datei schreiben wollen, können Sie dies mit der Prozedur PRINTTO tun. Das folgende SAS-Programm, steuert die Ergebnisausgabe direkt in eine Datei mit dem Namen OUTPUT.DAT:

```
PROC PRINTTO PRINT = 'C:\DATA\OUTPUT.DAT';
RUN;
```

1.5 Überblick über die wichtigsten Prozeduren in SAS

Die wichtigsten Prozeduren, die wir im Folgenden (mit einigen Optionen) kurz beschreiben, gehören alle zu den Modulen BASE, STAT oder GRAPH:

ANOVA:
Für Varianzanalysen (von einfachen bis zu multivariaten mit Messwiederholungen).

APPEND:
Zum Hinzufügen von Variablen von einem DATASET zu einem anderen.

CATMOD:
Modelle der kategoriellen Datenanalyse (lineare, loglineare Modelle und logistische Regressionsanalysen).

CLUSTER:
Für Clusteranalysen.

COMPARE:
Zum Vergleich zweier DATASETS.

CORR:
Zum Berechnen von Kovarianzen und Korrelationskoeffizienten, sowie für Tests auf Korrelation.

DATASETS:
Zum Löschen, Anzeigen und Umbenennen von DATASETS.

DISCRIM:
Für Diskriminanzanalysen.

FACTOR:
Für Faktorenanalysen.

FREQ:
Für Häufigkeitsauszählungen, Kontingenztafeln bzw. Kreuztabellen, Korrelationskoeffizienten für kategorielle Variablen und Chi-Quadrat-Tests auf Unabhängigkeit und exakter Fisher-Test.

GCHART:
Für Histogramme, Balkendiagramme und Kreisdiagramme.

GPLOT:
Für x-y-Diagramme mit Regressionen und Interpolationen.

G3D:
Für 3-dimensionale Diagramme (Scatterplots und Surfaceplots).

GLM:
Für allgemeine lineare Modelle und Hypothesen.

MEANS:
Zur Berechnung von univariaten Statistiken (Mittelwerte, empirische Varianzen, Maximum,...), sowie zur Durchführung des t-Tests.

NLIN:
Für nichtlineare Regressionsanalysen.

NPAR1WAY:
Zur Durchführung parameterfreier Verfahren (Vergleich von Stichproben).

PLOT:
Für einfache Plots im OUTPUT-Fenster.

PRINT:
Zur Ausgabe von SAS-DATASETS im OUTPUT-Fenster.

PRINTTO:
Zur Festlegung, ob die Ausgabe auf dem Bildschirm oder in eine Datei erfolgen soll.

PROBIT:
Für Probitanalysen.

RANK:
Zur Berechnung von Rangsummen u.a. zur Durchführung parameterfreier Verfahren.

REG:
Zur Durchführung einfacher und multipler linearer Regressionsanalysen.

SORT:
Zum Sortieren eines SAS-DATASETS nach einer oder mehreren Variablen.

TABULATE:
Zur Erstellung von Kontingenztafeln bzw. Kreuztabellen.

TREE:
Zur Ausgabe von Dendrogrammen bei Clusteranalysen und Baumdiagrammen.

TTEST:
Zur Durchführung von klassischen t-Tests (Vergleich zweier Stichproben).

UNIVARIATE:
Zur Berechnung univariater Statistiken (u.a. auch Häufigkeitsauszählungen).

2 Planung einer empirischen Studie am Beispiel

2.1 Problemstellung und Konzeption des Fragebogens

Bevor ein Fragebogen konzipiert wird, sollte geklärt werden, welche Statistiken benötigt werden bzw. welche Informationen aus den Daten benötigt werden. Somit wird sichergestellt, dass nicht eine Menge von Fragen gestellt werden, die man bei der Auswertung nicht mehr benötigt. Außerdem wird ein nicht zu umfangreicher Fragebogen eher ausgefüllt.

Ein weiteres Problem sind Fragen mit freien Antworten. Hierauf sollte weitgehend verzichtet werden, denn diese können später so gut wie nicht statistisch ausgewertet werden. Allerdings bieten freie Antwortmöglichkeiten den Personen, die einen Fragebogen ausfüllen, die Möglichkeit, eigene Ideen einzubringen.

Mit dem in Abb. 2-1 angegebenen Fragebogen wurde eine Umfrage bei 21 Teilnehmern einer Vorlesung durchgeführt.

Bei der Augenfarbe haben wir die Antworten "blau", "grün", "braun" und "grau" zugelassen. Bei diesen Antwortmöglichkeiten sind Mischfarben nicht vorgesehen. Diese "schlechte" Konzeption wird dazu führen, dass die Frage nach der Augenfarbe nicht immer beantwortet wird. Dieser Fehler wurde trotzdem in diesem "Musterbeispiel" nicht beseitigt, um zu zeigen, welche Auswirkungen bereits eine schlechte Gestaltung des Fragebogens auf die spätere statistische Auswertung haben kann.

Bitte kreuzen Sie die zutreffenden Antworten an:

1. Sind Sie Studentin / Student ? ja ( ) 1
   nein ( ) 2

2. Wenn ja, in welchem Semester ? _____

3. Zu welchem Fachbereich gehören Sie ? _____

4. Ihr Geschlecht ? männlich ( ) 1
   weiblich ( ) 2

5. Ihr Gewicht (in kg) ? _____

6. Ihre Größe (in cm) ? _____

7. Ihr Alter ? _____

8. Rauchen Sie ? ja ( ) 1
   nein ( ) 2

9. Rauch(t)en Ihr Vater oder Ihre Mutter oder beide ? ja ( ) 1
   nein ( ) 2

10. Glauben Sie, dass Rauchen Lungenkrebs verursacht ? ja ( ) 1
    nein ( ) 2

11. Ihre Augenfarbe ? blau ( ) 1
    grün ( ) 2
    braun ( ) 3
    grau ( ) 4

12. Welche Zeitung lesen Sie regelmäßig ?

	ja	nein
12.1 Frankfurter Allgemeine Zeitung (FAZ)	( ) 1	( ) 2
12.2 Frankfurter Rundschau	( ) 1	( ) 2
12.3 Die Zeit	( ) 1	( ) 2
12.4 Die Welt	( ) 1	( ) 2
12.5 Süddeutsche Zeitung	( ) 1	( ) 2

13. Zum Schluss geben Sie bitte noch eine achtstellige Zahl an: _ _ _ _ _ _ _ _

Abb. 2-1

2.2 Festlegung der Variablen und Kodierung

Zunächst muss ein sogenannter Kodeplan erstellt werden. Hierbei wird jeder Frage eine Variable zugewiesen, in der später die jeweilige Antwort gespeichert wird. Um jederzeit anhand des Variablennamens die Frage erkennen zu können, empfiehlt es sich, jede Variable mit der Nummer der Frage zu kennzeichnen (z.B. V5 für Frage Nr. 5).

Bei der Festlegung des Variablennamens ist außerdem das Folgende zu beachten:

(1) Am Anfang eines Variablennamens darf niemals eine Zahl stehen.

(2) Ein Variablennamen darf nicht mehr als 8 Zeichen beinhalten.

(3) Es dürfen keine Sonderzeichen wie "=" oder "!" verwendet werden.

(4) Als Variablennamen darf nur ein zusammenhängendes Wort verwendet werden.

Sind die Variablennamen festgelegt, müssen die Antworten kodiert werden. Bei den Fragen, bei denen die Studierenden eine Zahl als Antwort eintragen können, wie z.B. bei der Frage nach der Größe oder nach dem Alter, ist dies nicht nötig.

Betrachten wir nun die erste Frage. Die Antwort auf diese Frage wird später in der Variable, die wir mit V1 bezeichnen, gespeichert. Diese Frage hat nun zwei mögliche Antworten, nämlich die Antwort "ja" oder "nein". Da der Wert einer Variablen auch ein Wort sein darf, könnten wir nun "ja" und "nein" als Wert der Variablen zulassen. Um aber die Eingabe der Daten zu vereinfachen, kodieren wir die Antwort "ja" mit "1" und die Antwort "nein" mit "2". Somit hat die Variable V1 die zwei möglichen Ausprägungen 1 und 2. Abb. 2-2 zeigt den kompletten Kodierungsplan für unseren Fragebogen.

Falls die Daten später nicht direkt in SAS, sondern mit einem anderen System erfasst werden, sollte ein Wert für fehlende Antworten definiert werden. Dies sollte ein Wert sein, der sich von der Kodierung der anderen Antworten unterscheidet. Z.B. könnte der Wert "99" oder auch "-1" für fehlende Antworten verwendet werden. Diese fehlenden Werte (Missing Values) können später auch umgewandelt werden. Das könnte z.B. mit der Rekodierung (Kapitel 3.3) getan werden. Ansonsten können Sie bei der Eingabe in SAS jeweils einen Punkt eingeben. SAS lässt dann diese Werte bei Auswertungen unberücksichtigt.

Frage	Variablenname	Kodierung der Antworten
1	V1	ja = 1; nein = 2
2	V2	
3	V3	
4	V4	männlich = 1; weiblich = 2
5	V5	
6	V6	
7	V7	
8	V8	ja = 1; nein = 2
9	V9	ja = 1; nein = 2
10	V10	ja = 1; nein = 2
11	V11	blau = 1; grün = 2; braun = 3; grau = 4
12.1	V12_1	ja = 1; nein = 2
12.2	V12_2	ja = 1; nein = 2
12.3	V12_3	ja = 1; nein = 2
12.4	V12_4	ja = 1; nein = 2
12.5	V12_5	ja = 1; nein = 2
13	V13	

Abb. 2-2

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12_1	V12_2	V12_3	V12_4	V12_5	V13
1	6	16	2	63	169	24	1	1	2	4	1	2		1	1	12345678
1	4	18	2	59	168	25	2	2	1	3	2	2	2	2	2	98730032
1	7	17	1	71	189	30	2	1	1	1	2	1	2	2	2	14847464
1	10	6	2	80	176	27	1	1	2	4	1	2	1	1		46202825
1	5	16	2	56	164	25	2	1	1	2						45873210
1	9	14	1	73	177	24	2	2	2	2	1	2	2	1	1	43672382
2		16	1	77	178	21	1	2	2	1	1	1	1	1	1	87204353
2		16	1	82	181	23	1	1	1		1	1	2	1	2	26252424
1	8	17	2	54	159	23	1	2	2	3	2	1	2	2	2	12345678
				74	171	27	2	1	1		2	1	1	1	1	23456789
1	8	15	2	67	170	21	2	2	1		1	1	2	2	2	99999999
1	14	16	1	75	175	28	2	2	1		1	1	1	1	1	34612854
2		14	2	49	161	31	2	2	1		1	1	2	1	1	28473028
2		16	2	61	169	26	2	1	1	4	1	1		1	1	45439282
1	18	16	1	85	179	27	2	2		3	1	2	1	2	1	30289262
1	7	25	1	75	182	24	1	2	2		1	1	1	1	2	26438226
1	7	18	1	63	165	24	2	1	1		1	1	1	1	2	27389282
1	6	16	2	53	177	23	2	1	1		2	1	2	2	2	11111111
1	8	7	1	70	180	22	2	1	2	1	2	1	1	1	1	56776556
1	8	7	1	72	180	25	2	1	2		2	1	2	1	2	00700700
2		7	1	70	175	45	1	2	2	3	2	1	2	1		00700700

Abb. 2-3 Daten: FRAGEBOG