

**Error! Not a valid filename.**

Rechenzentrum

**Einführung in das  
Statistical Analysis System  
für Windows, Version 6.10**

# Einführung in das Statistical Analysis System

Frank Elsner  
Universität Osnabrück  
- Rechenzentrum -  
Albrechtstraße 28 (AVZ)  
D-49069 Osnabrück  
E-Mail: F.Elsner @rz.Uni-Osnabruock.DE

anonymous FTP Server: <ftp.rz.uni-osnabruock.de>  
Datei: [pub/reports/rz/sas\\_tutorial-1.5.ps.Z](ftp://ftp.rz.uni-osnabruock.de/pub/reports/rz/sas_tutorial-1.5.ps.Z)

Version des Handbuchs: 1.5  
Stand: 97/01/28

## Inhaltsverzeichnis

1. Einleitung .....	1
1.1. Zielsetzung .....	1
1.2. Typographische Konventionen .....	1
1.3. Abhängigkeiten .....	2
2. Grundlagen der Wahrscheinlichkeitsrechnung und Statistik .....	3
2.1. Überblick über die deskriptive Statistik .....	3
2.2. Stichprobe und Grundgesamtheit .....	3
2.3. Messung von Variablen .....	4
2.4. Kenngrößen von Stichproben .....	5
2.5. Zufallsexperiment und Wahrscheinlichkeit .....	6
2.6. Beziehungen zwischen Zufallsvariablen .....	8
2.7. Überblick über die mathematische Statistik .....	9
2.8. Ein einfaches Beispiel .....	11
2.9. Normalverteilung und Normalverteilungsannahme .....	11
3. Überblick über das Statistical Analysis System .....	13
3.1. Funktionsumfang des SAS Systems .....	13
3.2. Module des SAS Systems .....	13
3.3. Benutzerschnittstelle zum SAS System .....	14
3.4. SAS Programme .....	15
3.5. Syntax der SAS Programmiersprache .....	16
3.6. Elemente einer SAS Datei .....	17
3.7. Übungen .....	18
4. Einführendes Beispiel .....	20
4.1. Starten einer Sitzung .....	20
4.2. Ausführen eines SAS Programmes .....	20
4.3. Speichern eines Programmes und Beenden einer SAS Sitzung .....	21
4.4. Übungen .....	21
5. Bedienen der Benutzeroberfläche .....	22
5.1. Elemente der Benutzeroberfläche .....	22
5.2. Wichtige Menüpunkte .....	22
5.4. Interaktives Entwickeln von SAS Programmen .....	23
Sie können ein SAS Programm alternativ (wie im einführenden Beispiel) auch direkt im PROGRAM EDITOR Fenster erfassen.	
5.5. Zusammenfassung .....	23
5.6. Übungen .....	24
6. Einlesen von Datenwerten in eine SAS Datei .....	25
6.1. DATA - Listengesteuertes Einlesen .....	25
6.2. DATA - Spaltengesteuertes Einlesen .....	26
6.3. DATA - Einlesen von mehreren Eingabezeilen .....	27
6.4. DATA - Mehrfaches Einlesen von einer Eingabezeile .....	28
6.5. DATA - Einlesen aus einer externen Datei .....	29
6.6. PROC CONTENTS - Auflisten von Informationen .....	29
6.7. Zusammenfassung .....	30
6.8. Übungen .....	31
7. Sortieren und tabellarisches Darstellen einer SAS Datei .....	33
7.1. PROC PRINT - Auflisten einer SAS Datei .....	33
7.2. PROC FREQ - Berechnen von Häufigkeiten .....	34

7.3. PROC MEANS - Berechnen von Maßzahlen .....	35
7.4. Zusammenfassung .....	36
7.5. Übungen .....	37
8. Grafisches Darstellen einer SAS Datei .....	38
8.1. PROC GCHART - Erzeugen von Diagrammen.....	38
8.2. PROC GPLOT - Erzeugen von Streudiagrammen.....	39
Übungen .....	40
9. Durchführen elementarer statistischer Verfahren.....	41
9.1. PROC UNIVARIATE - Exploratives Analysieren von Daten.....	41
9.2. PROC MEANS - Berechnen eines Vertrauensbereiches .....	42
9.3. Statistischer Hintergrund.....	43
9.4. PROC TTEST - Vergleichen der Erwartungswerte von zwei Gruppen.....	45
9.5. Statistischer Hintergrund.....	45
9.6. PROC FREQ - Testen auf Unabhängigkeit (Chi-Quadrat Test).....	46
9.7. Statistischer Hintergrund.....	47
9.8. PROC CORR - Testen auf Korrelation .....	48
9.9. Statistischer Hintergrund.....	48
9.10. Zusammenfassung.....	49
9.11. Übungen .....	50
10. Durchführen fortgeschrittener statistischer Verfahren .....	50
10.1. PROC REG - Berechnen einer Regressionsgeraden .....	50
10.2. Statistischer Hintergrund.....	52
11. Bearbeiten der SAS Datei .....	54
11.1. DATA - Erzeugen neuer (abgeleiteter) Variablen .....	54
11.2. DATA - Bedingtes Einlesen von Beobachtungen.....	55
11.3. DATA - Definieren von genaueren Bezeichnungen für Variablen.....	56
11.4. DATA - Definieren von genaueren Bezeichnungen für Datenwerte .....	57
11.5. DATA - Einlesen von Datums- und Zeitformaten .....	58
11.6. DATA - Einlesen aus permanenten SAS Dateien.....	58
11.7. Zusammenfassung .....	59
11.8. Übungen .....	60
12. Anzeigen und Ändern von Voreinstellungen .....	62
12.1. TITLE, FOOTNOTE - Hinzufügen von Titel- und Fußzeilen .....	62
12.2. (G)OPTIONS - Setzen von Systemoptionen.....	62
12.3. HELP - Anzeigen von Informationen .....	63
12.4. HELP - Arbeiten mit SAS Beispielprogrammen .....	64
12.5. Zusammenfassung.....	64
12.6. Übungen .....	66
13. Hinweise zu den Übungen .....	67
14. Anhang A - Literaturhinweise und Online Informationen .....	68
15. Anhang B - Quantile von wichtigen Verteilungen.....	69
16. Index.....	1

## 1. Einleitung

In diesem Kapitel wird die Zielsetzung und der Aufbau des Handbuches beschrieben.

### 1.1. Zielsetzung

Dieses Handbuch wendet sich an Benutzer, die bereits grundlegende Kenntnisse über Datenverarbeitung und Statistik besitzen und nun das **Statistical Analysis System** als ein konkretes Statistik-Paket kennenlernen wollen. Programmierkenntnisse sind hilfreich, aber nicht unbedingt erforderlich.

Dieses Handbuch will und kann nicht die umfangreiche Literatur zum Statistical Analysis System ersetzen; es will vielmehr einen behutsamen Zugang zu diesem anspruchsvollen System vermitteln und die Abwicklung einfacher Projekte veranschaulichen. Hierzu werden einfache Beispielprogramme verwendet, aus denen vereinfachte Syntaxbeschreibungen der SAS Programmiersprache abgeleitet werden.

### 1.2. Typographische Konventionen

Folgende typographischen Konventionen werden verwendet.

<b>Fett</b>	kennzeichnet Kommandos, Prozeduren, wichtige Textpassagen oder erstmalig genannte Begriffe; z.B.: Ein <b>SAS Programm</b> ist ...
<i>Courier</i>	kennzeichnet Programmbeispiele sowie Namen von Dateien, Variablen und Programmen im Fließtext; z.B.: Die Datei <code>PLOT.EPS</code> ....
<i>Kursiv</i>	kennzeichnet variable Werte in SAS Programmen, die durch konkrete Werte zu ersetzen sind; z.B.: <code>PLOT <i>y1</i>*<i>x1</i>;</code> /* Platzhalter */ <code>PLOT <i>anzahl</i>*<i>jahr</i>;</code> /* Variablennamen */

Programmbeispiele sind durch eine Umrahmung gekennzeichnet.

Syntaxbeschreibungen sind durch eine Umrahmung und zusätzlich durch eine Schattierung gekennzeichnet.

Eckige Klammern [...] kennzeichnen Syntaxelemente, die nicht unbedingt eingegeben werden müssen (optionale Angaben).

`/* ... */` Blöcke enthalten Kommentare.

### 1.3. Abhängigkeiten

Die Beschreibung des SAS Systems erfolgt weitgehend systemunabhängig. Ausnahmen bilden im wesentlichen SAS Programme, in denen ein Bezug zu externen Dateinamen notwendig ist, da sich die Konventionen für Dateinamen z.B. zwischen DOS, UNIX und VM/CMS stark unterscheiden. Ferner gibt es Unterschiede bei den Benutzerschnittstellen, z.B. unterschiedliche Belegung der Funktionstasten oder Mausunterstützung für graphische Benutzeroberflächen wie Windows.

Dieses Handbuch bezieht sich auf **SAS für Windows, Version 6.10, englisch**.

## 2. Grundlagen der Wahrscheinlichkeitsrechnung und Statistik

In diesem Kapitel werden wichtige Begriffe der Wahrscheinlichkeitsrechnung und Statistik zusammengestellt.

### 2.1. Überblick über die deskriptive Statistik

Die **deskriptive Statistik** (beschreibende Statistik) befaßt sich mit der tabellarischen und graphischen Darstellung von Daten und der Zusammenfassung (Verdichtung, Aggregation) auf charakteristische Kenngrößen oder Maßzahlen (z.B. Lage- und Streumaße) und damit als für die mathematische Statistik, indem sie erste Hinweise für weitergehende Analysen liefert.

Der Untersuchungsgegenstand der deskriptiven Statistik sind **Beobachtungen** von zufälligen und nicht-zufälligen **Variablen (Merkmalen oder Eigenschaften)** von Objekten oder Personen. Zufällige Variablen können im Anschluß mit Verfahren der mathematischen Statistik analysiert werden, um z.B. gesicherte Aussagen über Zusammenhänge zwischen einzelnen Variablen ableiten zu können.

In der deskriptiven Statistik werden Variablen und Beziehungen zwischen Variablen u.a. mit folgenden Tabellen und graphischen Hilfsmitteln dargestellt:

- Häufigkeitstabelle (*frequency table*)
- Histogramm (*histogram*)
- Kreuztabelle (*cross tabulation*)
- Streudiagramm (*scatter plot*)
- Liniendiagramm (*line diagram*)

In der Regel werden von einem Objekt oder einer Person mehrere Variablen gleichzeitig beobachtet, so daß eine Beobachtung aus mehreren Variablen besteht. Z.B. könnten bei einer Person gleichzeitig die zufälligen Variablen Größe, Gewicht, Alter und Geschlecht beobachtet werden; d.h. eine Beobachtung  $x$  setzt sich aus vier Datenwerten zusammen:

$$\mathbf{x} = (\text{Größe}, \text{Gewicht}, \text{Alter}, \text{Geschlecht})$$

Die Beobachtung  $x$  wird dann als **vektoriell** oder **multivariat** bezeichnet. Im Gegensatz hierzu wird eine Beobachtung, die aus nur einer Variablen besteht, bzw. eine Analyse, die nur eine Variable berücksichtigt, als **univariat** bezeichnet. Ein häufiger Spezialfall sind **bivariate** Auswertungen, die sich auf 2 Variablen beziehen (z.B. Korrelation).

### 2.2. Stichprobe und Grundgesamtheit

Eine **Stichprobe** (*sample*) ist eine Auswahl von **Beobachtungen** (*observations*) aus einer **Grundgesamtheit** (*population*). In der Terminologie der Wahrscheinlichkeitsrechnung besteht eine Stichprobe aus Realisierungen von zufälligen, also nicht deterministisch vorhersagbaren Variablen (Zufallsvariablen).

Eine Stichprobe wird u. a. aus folgenden **Gründen** durchgeführt:

- Eine Gesamterhebung, d.h. die Untersuchung der Grundgesamtheit, ist unmöglich, zerstört die untersuchten Objekte, ist zu teuer oder dauert zu lange.
- Aus vorherigen Untersuchungen ist bekannt, daß sich das Verhalten der Grundgesamtheit zufriedenstellend über Beobachtungen von ausgewählten Repräsentanten beschreiben läßt.

Eine Stichprobe ist gekennzeichnet durch:

- Erhebungszeitraum (einmalig, periodisch)
- Grundgesamtheit (Umfang, Verteilungsannahme)
- Auswahlverfahren (einstufig, mehrstufig)
- Erhebungsverfahren (Interview, Fragebogen, Messung)
- Umfang (Anzahl der Beobachtungen und Anzahl der Variablen)
- Art und Anzahl der Variablen pro Beobachtung

**Beispiele** für Stichproben sind:

- repräsentative Umfragen vor Bundestagswahlen
- Mikrozensus
- zufälliges Entnehmen von Wasserproben
- Testen einzelner Blitzlichter (Hierbei wird das Testobjekt zerstört!)
- Markieren einzelner Vögel und Beobachten ihrer Lebensweise
- Befragen ausgesuchter Haushalte über ihre Fernsehgewohnheiten

Eine Stichprobe erhebt in der Regel den Anspruch, im Kleinen das Verhalten der Grundgesamtheit widerzuspiegeln (**repräsentativ**) zu sein. Hierzu müssen Auswahlmechanismen festgelegt werden, die eine repräsentative Auswahl der Stichprobe aus der Grundgesamtheit garantieren, dies erfolgt z.B. durch Einteilung der Grundgesamtheit in Klassen und zufällige Auswahl von Repräsentanten aus jeder Klasse (mehrstufige Zufallsauswahl).

### 2.3. Messung von Variablen

Bei den beobachteten Variablen handelt es sich um **klassifizierende** und/oder **analysierbare** Variablen:

Eine Variable ist **klassifizierend**, wenn Sie die Stichprobe in sinnvolle "Klassen" oder "Gruppen" einteilt. Beispiele sind die Variable "Geschlecht", die die Stichprobe in die Gruppen "Männer" und "Frauen" aufspaltet, oder die Variable "Nationalität", die die Stichprobe nach Staatsangehörigkeit untergliedert. Sie ist **rein klassifizierend**, wenn die möglichen Datenwerte keine numerischen Größen repräsentieren wie z.B. das Geschlecht.

Eine Variable ist **analysierbar**, wenn sie konkrete numerische Datenwerte wie z.B. Alter, Gewicht oder Einkommen repräsentiert.

Bei der **Messung** von Variablen sind folgende Wertebereiche (Skalen) für Variablen zu unterscheiden:

1. dichotome Messung
2. nominale Messung

3. ordinale Messung
4. Intervall-Messung
5. Verhältnis-Messung

Bei einer **dichotomen Messung** wird nur zwischen zwei möglichen Werten einer Variablen unterschieden, z.B. Wahr/Falsch oder Tot/Lebendig. Die Unterscheidung kann dabei auch willkürlich gewählt werden, z.B. (Älter als 18)/(Jünger als 18). Werte für dichotome Messungen werden häufig zur Abkürzung mit den Zahlen 0 und 1 kodiert, wobei die Zahlen selbst keine Bedeutung haben, sondern einfach nur einen von zwei möglichen Werten repräsentieren.

Bei einer **nominalen Messung** werden die möglichen Werte einer Variablen zur Abkürzung willkürlich auf Zahlen abgebildet. Die Zahl hat keine andere Bedeutung als daß sie stellvertretend für einen Wert steht. Z.B. können die Bundesländer mit Zahlen von 1 bis 16 durchnummeriert werden, wobei die Zahl 1 für das Bundesland Berlin steht, 2 für Brandenburg usw.

Bei einer **ordinalen Messung** gibt es eine auf- oder absteigende Ordnung zwischen den möglichen Werten. Es kommt dabei aber nur auf die Reihenfolge und nicht auf die Abstände zwischen den Werten an. Z.B. legen Schulnoten oder andere Bewertungen zwischen 1 und 6 eine Reihenfolge fest, aber die Abstände haben keine gleichbleibende Bedeutung.

Bei einer **Intervall-Messung** gibt es eine Ordnung und zusätzlich sind die Abstände zwischen den Werten von Bedeutung. Z.B. ist eine Temperatur von 70 C um 30 C höher als 40 C. Die Absolutwerte sind allerdings nicht von Bedeutung, da der Nullpunkt willkürlich gewählt ist.

Bei einer **Verhältnis-Messung** gibt es einen ausgezeichneten Nullpunkt und die Abstände zwischen den Werten sind von Bedeutung. Z.B. wird das Gewicht, die Größe oder das Einkommen durch eine Verhältnis-Messung bewertet.

### 2.4. Kenngrößen von Stichproben

Eine **Stichprobe**  $S = (x_1, \dots, x_n)$  setzt sich aus  $n$  **Beobachtungen** (andere Bezeichnung: **Fälle**)  $x_1, \dots, x_n$  zusammen.

Oftmals sind nicht die einzelnen Werte interessant, sondern Kenngrößen oder Maßzahlen, die einen Überblick über die gesamte Stichprobe vermitteln. Z.B. können Sie eine Stichprobe "verdichten", indem Sie nur den kleinsten, den größten Wert und den Mittelwert der Stichprobe betrachten. Mit jeder Verdichtung ist grundsätzlich ein Informationsverlust verbunden.

Wichtige **Maßzahlen** (andere Bezeichnungen: **Kennzahlen** oder **-größen, Lage- oder Streumaße**) der Stichprobe sind für jede numerische Variable folgendermaßen definiert:

Der **empirische Mittelwert** (*empirical mean*)  $\bar{x}$  der Stichprobe  $S = (x_1, \dots, x_n)$  ist definiert als:

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Die **empirische Varianz** (*emp. variance*)  $s^2$  der Stichprobe  $S = (x_1, \dots, x_n)$  ist definiert als:

$$s^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Die **empirische Standardabweichung** (*emp. standard deviation, stddev*)  $s$  der Stichprobe  $S$  ist definiert als Wurzel aus der Varianz:

$$s = \sqrt{s^2}$$

Die **geordnete Stichprobe** (*ordered sample*) enthält die nach aufsteigender Reihenfolge geordneten Werte  $x_1, \dots, x_n$ . Mit  $x_{(1)}$  wird der kleinste, mit  $x_{(n)}$  der größte Wert bezeichnet, mit  $x_{(2)}$  der zweitkleinste usw.:

$$S_{\text{sorted}} = (x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)})$$

Das **Minimum** und das **Maximum** der Stichprobe  $S$  sind definiert als:

$$\min(S) = \min(x_1, \dots, x_n) = x_{(1)}$$

$$\max(S) = \max(x_1, \dots, x_n) = x_{(n)}$$

Die **Spannweite** (*range*) der Stichprobe  $S$  ist definiert als:

$$\text{range}(S) = \max(S) - \min(S) = x_{(n)} - x_{(1)}$$

Der **empirische Median** (*emp. median*) der Stichprobe  $S$  ist definiert als der mittlere Wert in der geordneten Stichprobe (bzw. als das arithmetische Mittel der mittleren Werte für  $n$  gerade):

$$\text{med}(S) = x_{(n/2)}$$

Die **empirische Verteilungsfunktion** (*empirical distribution function*) der Stichprobe  $S$  ist definiert als:

$$F^-(x) = \frac{1}{n} (\text{Anzahl beobachtete Werte } x_i \leq x)$$

## 2.5. Zufallsexperiment und Wahrscheinlichkeit

Ein Vorgang oder Versuch, dessen Durchführung "zufällig" zu genau einem von mehreren möglichen Ergebnissen führt, wird als **Zufallsexperiment** oder **Zufallsvorgang** bezeichnet.

Klassische Beispiele stammen aus der Welt der Spiele wie das Werfen eines Würfels, das Ziehen von Losen in einer Lotterie oder das Ziehen einer Spielkarte beim Poker. Ein anderes Beispiel ist die Auswahl einer Stichprobe aus einer Grundgesamtheit, bei der  $m$  Objekte "zufällig" aus der Grundgesamtheit mit  $n$  Objekten ausgewählt werden ( $m \leq n$ ).

Sei ein Zufallsexperiment mit der Ergebnismenge  $\Omega$  (Menge der möglichen Ergebnisse) gegeben. Die **Wahrscheinlichkeit P** ist eine Abbildung von  $\Omega$  in das Intervall  $[0,1]$ , die jedem **Ergebnis** eine positive Zahl  $p$  zuordnet.

Die Wahrscheinlichkeit  $P$  für die Ergebnisse oder Ereignisse eines Zufallsexperimentes wird nicht "bewiesen", sondern ihre Existenz wird als plausible Annahme (Axiom) vorausgesetzt. Intuitiv ist  $P(w_1)$  die "stabilisierte" relative Häufigkeit für das Ergebnis  $w_1$  für eine große Anzahl von Versuchen. Die Abbildung  $P$ , die Ergebnissen "Wahrscheinlichkeiten" zuweist, wird aufgrund von plausiblen Annahmen, Erfahrungswerten oder Schätzungen aufgestellt.

Bei einem Zufallsexperiment seien die Ergebnisse  $\Omega = \{w_1, \dots, w_n\}$  möglich. Jedem Ergebnis werde durch die Abbildung  $X$  eine reelle Zahl zugeordnet. Die Abbildung  $X$  heißt **Zufallsvariable**, die möglichen Werte von  $X$  ergeben den **Wertebereich** von  $X$ . Zufallsvariablen sind die beobachtbaren **Merkmale** oder **Eigenschaften** von Objekten oder Personen, die in einem Zufallsexperiment ausgewählt werden.

Bei einem Zufallsexperiment interessieren oft nicht die elementaren Ergebnisse, sondern eine vom Ergebnis  $w$  abgeleitete Variable  $X(w)$ . Bei der zufälligen Auswahl einer Person könnte das Ergebnis  $w$  der Name sein, während  $X(w)$  das Einkommen der Person bezeichnet. Ähnlich interessieren bei einem Angelwettbewerb in der Regel nicht die einzelnen gefangenen Fische, sondern die Anzahl oder das Gesamtgewicht aller gefangenen Fische bzw. der schwerste gefangene Fisch.

Die (Wahrscheinlichkeits-) **Verteilung**  $P^X$  einer Zufallsvariablen  $X$  wird über die Wahrscheinlichkeit der Urbilder in der Ergebnismenge  $\Omega$  definiert; d.h. die Summe aller Wahrscheinlichkeiten für Ergebnisse, die zu einem Wert  $k$  von  $X$  führen.. Die Verteilung von  $X$ ,  $P^X$ , gibt also Auskunft darüber, mit welcher Wahrscheinlichkeit die Zufallsvariable  $X$  einen bestimmten Wert  $k$  aus dem Wertebereich annimmt.

Beachten Sie, daß

- Sie nur **vor der Durchführung** des Zufallsexperiments Aussagen über die **möglichen Werte** und **deren Wahrscheinlichkeit** treffen können, während **nach der Durchführung** genau ein **realisierter Wert** zur Verfügung steht.

Dieser Sachverhalt wird durch folgende Notation verdeutlicht:

Zufallsvariablen werden mit Großbuchstaben ( $X, Y, Z, \dots$ ) bezeichnet, während beobachtete Werte mit Kleinbuchstaben ( $x, y, z, \dots$ ) bezeichnet werden:

<b>Zufallsexperiment</b>	
<b>vor der Durchführung:</b>	<b>nach der Durchführung:</b>
Wahrscheinlichkeiten für mögliche Ergebnisse	Realisierung eines Ergebnisses
Zufallsvariable $X$ , mögl. Werte $x_1, x_2, x_3, \dots$	realisierter Wert, z.B. $x_3$
$P(X=x_3) = p_3; 0 \leq p_3 \leq 1$	-

Beachten Sie, daß

- es also sehr wohl möglich, daß der Wert  $x_3$  mit der kleinsten Wahrscheinlichkeit  $p_3$  beobachtet wird.
- nur **bei häufiger Wiederholung** eines Zufallsexperiments zu erwarten ist, daß Werte mit großen Wahrscheinlichkeiten auch häufiger eintreten.

Die **Binomialverteilung**  $B(x;n,p)$  beschreibt z.B. die Verteilung einer Zufallsvariablen  $X$ , die die Anzahl der Treffer in  $n$  unabhängigen Versuchen aufsummiert, wobei die Wahrscheinlichkeit für einen Treffer  $p$  beträgt.

Die konkrete Verteilung der Binomialverteilung für die Anzahl der Treffer  $k$  in  $n$  Versuchen sieht folgendermaßen aus:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \dots, n$$

Sie ist eindeutig bestimmt ("parametrisiert") durch die Anzahl der Versuche  $n$  und die Wahrscheinlichkeit für einen Treffer  $p$ .

### 2.6. Beziehungen zwischen Zufallsvariablen

Der **Erwartungswert** einer Zufallsvariablen  $X$ ,  $E[X]$ , ist definiert als das mit der Wahrscheinlichkeit gewichtete Mittel der möglichen Werte für  $X$ ; der Erwartungswert wird oft mit  $\mu$  abgekürzt.

$$\mu = E[X] = k_1 P(X=k_1) + \dots + k_n P(X=k_n) = \sum_{i=1}^n k_i P(X=k_i)$$

Sei  $X$  der Spielgewinn bei einem Spiel, z.B. beim Roulette, bei dem Sie mit Wahrscheinlichkeit  $P(X=k_1)$  einen Gewinn von  $k_1$  DM machen, mit Wahrscheinlichkeit  $P(X=k_2)$  den Gewinn von  $k_2$  DM usw. Der Erwartungswert von  $X$  ist der mittlere Spielgewinn pro Spiel, wenn Sie das Spiel sehr häufig spielen. (Diese Aussage läßt sich mathematisch formulieren und heißt dann "Schwaches Gesetz der großen Zahlen".)

Die **Varianz** einer Zufallsvariablen  $X$  ist definiert als die mittlere quadratische Abweichung vom Erwartungswert:

$$\text{Var}(X) = E[(X - E[X])^2] = E[(X - \mu)^2] \quad (\text{siehe vorherige Definition})$$

Die Wurzel aus der Varianz wird als **Standardabweichung** bezeichnet:

$$\text{StdAbw}(X) = \sqrt{\text{Var}(X)}$$

Varianz und Standardabweichung werden oft mit  $\sigma^2$  bzw.  $\sigma_{XX}$  und  $\sigma$  abgekürzt.

Die Varianz charakterisiert die Abweichung (Streuung) der möglichen Werte einer Zufallsvariablen um den Erwartungswert, wobei jeder mögliche Wert mit seiner Wahrscheinlichkeit gewichtet wird. Eine Zu-

fallsvariable, deren möglichen Werte "dicht" gedrängt liegen, hat eine kleine Streuung, eine Zufallsvariable, deren mögliche Werte "weit" auseinander liegen, hat eine große Streuung.

Zwei Ereignisse A und B heißen **unabhängig**, falls gilt:

$$P(\text{"Sowohl A als auch B treten ein."}) = P(A) P(B)$$

Zwei Zufallsvariablen X und Y heißen **unabhängig**, falls für alle möglichen Werte k von X und m von Y gilt:

$$P(X=k, Y=m) = P(X=k) P(Y=m)$$

Unabhängige Ereignisse beeinflussen einander nicht; d.h. die Kenntnis, daß das Ereignis A eingetreten ist, hat keinen Einfluß darauf, mit welcher Wahrscheinlichkeit das Ereignis B eintreten wird. In der Terminologie der bedingten Wahrscheinlichkeit kommt das in folgender Formel zum Ausdruck:  $P(A|B) = P(A)$

Die **Kovarianz**  $\sigma_{XY}$  zwischen zwei Zufallsvariablen X und Y ist definiert als:

$$\sigma_{XY} = \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Die **Korrelation**  $\rho$  zwischen zwei Zufallsvariablen X und Y ist definiert als:

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

Die Korrelation  $\rho$  ist ein schwächeres Maß als die Unabhängigkeit, um die gegenseitige Beeinflussung von Zufallsvariablen zu beschreiben. Für  $\rho = 1$  und  $\rho = -1$  besteht ein **linearer Zusammenhang** zwischen X und Y; d.h. die Punkte (X, Y) liegen auf einer Geraden,  $P(Y = aX + b) = 1$ . Für kleines  $\rho$  besteht kein oder nur ein geringer linearer Zusammenhang.

## 2.7. Überblick über die mathematische Statistik

In der Mathematischen Statistik werden Verfahren entwickelt und angewendet, um anhand von Stichproben (Auswahl einer Teilmenge der Grundgesamtheit) Rückschlüsse auf die Grundgesamtheit ziehen zu können.

**Stichprobe** und **Grundgesamtheit** lassen sich durch folgende, korrespondierende Größen beschreiben:

Grundgesamtheit	Stichprobe
Anzahl Elemente m	Anzahl Beobachtungen n
Verteilung $P(X=k)$	Relative Häufigkeit <sup>1</sup> $h_n(k) = \frac{\#(x_i=k)}{n}$

<sup>1</sup> Das Zeichen # dient zur Abkürzung für Anzahl, z.B. #(Augenzahl=5) steht für: Anzahl der Würfe, bei denen die gewürfelte Augenzahl 5 beträgt.

## 2. Grundlagen der Wahrscheinlichkeitsrechnung und Statistik

Verteilungsfunktion $F(k)=P(X\leq k)$	Empirische Verteilungsfunktion von S $F^-(k)=\frac{\#(x_i\leq k)}{n}$
Erwartungswert $\mu=E[X]=k_1 P(X=k_1)+\dots$	Mittelwert von S $\bar{x}=k_1 \frac{\#(x_i=k_1)}{n}+\dots$
Varianz $\sigma^2=\sigma_{XX}=E[(X-\mu)^2]=(k_1-\mu)^2 P(X=k_1)+\dots$	Empirische Varianz von S $s^2=s_{XX}=(k_1-\bar{x})^2 \frac{\#(x_i=k_1)}{n-1}+\dots$
Standardabweichung $\sigma=\sqrt{\sigma^2}$	Empirische Standardabweichung von S $s=\sqrt{s^2}$
Median $m=F^{-1}(0.5)$	Empirischer Median von S $m^-=x(\text{middle})$
Kovarianz $\rho_{XY}=\text{Cov}(X,Y)=E[(X-E[X])(Y-E[Y])]$	Empirische Kovarianz $s_{xy}=(k_1-\bar{x})(l_1-\bar{y}) \frac{\#(x_i=k_1, y_i=l_1)}{n}+\dots$

Beachten Sie, daß

- die beobachteten **Maßzahlen der Stichprobe** nicht mit den entsprechenden **Maßzahlen der Grundgesamtheit** übereinstimmen.
- die Mathematische Statistik jedoch Verfahren zur Verfügung stellt, um auf Grundlage der Stichprobe **plausible Schätzungen** für die Grundgesamtheit abzugeben oder um **Tests** über bestimmte Aussagen durchzuführen.

Viele Verfahren der Mathematischen Statistik lassen sich auf folgende Fragestellung zurückführen:

Welche Aussage über den **unbekannten** wahren Wert der Grundgesamtheit kann aufgrund der Beobachtung des korrespondierenden **realisierten** (empirischen) Wertes der Stichprobe gemacht werden?

Entgegen einer weitverbreiteten Meinung bedeutet mathematische Statistik **nicht** (oder nur in sehr geringem Maße) Sammeln und tabellarisches Zusammenstellen (evtl. auch Manipulation?) von Unmengen an Zahlenmaterial [... es gibt die Notlüge, die gemeine Lüge und die Statistik ...], sondern die Entwicklung und Begründung von **Verfahren** zur Auswertung von zufallsabhängigen Beobachtungsdaten, mit denen sich "vernünftige" Entscheidungen bei ungewisser Sachlage treffen lassen.

Vernünftig heißt in diesem Zusammenhang, daß die Sicherheit mit der das Verfahren zu einer Entscheidung führt, vertrauenserweckend ist. Ein Verfahren hat eine **Sicherheit (Erfolgswahrscheinlichkeit, Signifikanz-Niveau)** von z.B. 0.95, wenn es im Mittel in 95 von 100 Durchführungen zu einer richtigen Entscheidung führt, und entsprechend eine **Irrtumswahrscheinlichkeit** von 0.05.

Bei konkreten Problemen liegen oft gewisse Kenntnisse hinsichtlich der "Rahmenbedingungen" eines Zufallsexperimentes vor (z.B. "n-malige Stichprobenentnahme mit Zurücklegen"), so daß die Menge aller zugelassenen Verteilungen auf eine Klasse von Verteilungen eingeschränkt werden kann.

In diesem Fall spricht man von einer **Verteilungsannahme**, d.h. der Auswahl einer Klasse von Verteilungen, in der sich die einzelnen Verteilungen nur durch unterschiedliche Kenngrößen oder Maßzahlen wie Lage- oder Streumaße (z.B. Erwartungswert, Varianz) unterscheiden. Die einfachere Aufgabe besteht in diesem Fall

nun darin, Aussagen über die unbekanntes Maßzahlen zu erhalten. Aus der anderen Bezeichnung **Parameter** für Kenngröße oder Maßzahl leitet sich der Begriff **Parametrische Statistik** für diesen Bereich von statistischen Fragestellungen ab. Viele der bekannten statistischen Verfahren gehen übrigens davon aus, daß die beobachteten Zufallsvariablen unabhängig sind und daß die Verteilung der Grundgesamtheit eine Normalverteilung mit unbekanntes Parametern  $\mu$  und  $\sigma$  ist (Normalverteilungsannahme).

### 2.8. Ein einfaches Beispiel

Die möglichen statistischen Fragestellungen sollen am folgenden einfachen Beispiel erläutert werden:

Beim 100-maligen Werfen eines Würfels mit den Augensummen  $x_1, \dots, x_{100}$  interessiere der Erwartungswert der gewürfelten Augenzahl. Bei einem "echten" Würfel berechnet sich der Erwartungswert  $\mu$  aus Symmetriegründen zu  $(1+2+3+4+5+6)/6=3.5$ , aber vielleicht ist der Würfel manipuliert?!

1. Welcher **Schätzwert**  $T(x_1, \dots, x_n)$  für den Parameter  $\mu$  kann aus der Stichprobe  $S=(x_1, \dots, x_n)$  abgeleitet werden?  
(*Punkt-Schätzung*)
2. Welcher **Schätzwert für ein Intervall**  $[a, b] = [CI_L(x_1, \dots, x_n), CI_R(x_1, \dots, x_n)]$ , das den Parameter  $\mu$  mit großer Sicherheit enthält, kann aus der Stichprobe  $S=(x_1, \dots, x_n)$  abgeleitet werden?  
(*Konfidenzbereichs-Schätzung*)
3. Wie kann aufgrund der Stichprobe  $S=(x_1, \dots, x_n)$  eine begründete Entscheidung gegeben werden, ob die **Hypothese** ' $\mu = 3.5$ ' angenommen oder abgelehnt werden soll? Wie groß sind die Fehler 1. Art und 2. Art (Annahme der Hypothese, obwohl sie falsch ist bzw. Ablehnung der Hypothese, obwohl sie wahr ist)?  
(*Hypothesen-Test*)

Beim Hypothesentest gibt es ein Dilemma besonderer Art - es können 2 verschiedene Typen von Fehlern auftreten. Die folgende Tabelle zeigt die möglichen Kombinationen:

Wahrheit/ Entscheidung	Hypothese ist wahr.	Hypothese ist falsch.
<b>Hypothese wird angenommen.</b>	Richtige Entscheidung	Falsche Entscheidung Fehler 2. Art $\beta$
<b>Hypothese wird abgelehnt.</b>	Falsche Entscheidung Fehler 1. Art $\alpha$	Richtige Entscheidung

Es ist i.d.R. kein statistisches Verfahren bekannt, mit dem beide Fehlerarten gleichzeitig minimiert werden können. Es ist allerdings häufig möglich, bei **vorgegebenem** Fehler 1. Art ein Verfahren mit minimalem Fehler 2. Art zu konstruieren (z.B. *Maximum Likelihood* Verfahren).

### 2.9. Normalverteilung und Normalverteilungsannahme

Die wichtigste Verteilung in der Mathematischen Statistik stellt die **Normalverteilung** mit folgender Dichte und Verteilungsfunktion dar:

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$N(x,\mu,\sigma^2) = \int_{-\infty}^x f(y)dy$$

Wegen der Glockenform von  $f(y)$  und ihrem Entdecker Gauß spricht man auch von der Gauß'schen Glockenkurve (siehe z.B. 10 DM Schein) oder der Gauß'schen Normalverteilung

Die **standardisierte** oder **normierte Normalverteilung**  $N(x;0,1)$  besitzt speziell den Erwartungswert  $\mu=0$  und die Varianz  $\sigma^2=1$  und ist tabelliert:

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$$

$$N(x,0,1) = \int_{-\infty}^x f(y)dy$$

Die Bedeutung der Normalverteilung beruht u.a. auf einem fundamentalen Satz der Mathematischen Statistik (**Zentraler Grenzwertsatz**), der besagt, daß der Mittelwert  $\bar{X}$  einer Stichprobe unter bestimmten Voraussetzungen "ungefähr" (approximativ) normalverteilt ist.

Die Zufallsvariablen  $X_1, \dots, X_n$  seien unabhängig und besitzen alle den selben Erwartungswert  $\mu$  und die selbe endliche Varianz  $\sigma^2$ . Hierbei sei  $\bar{X} = (1/n)(X_1 + \dots + X_n)$

$$\frac{\sqrt{n}}{\sigma} (\bar{X} - \mu) \sim N(x; 0, 1)$$

$$E\left[\frac{\sqrt{n}}{\sigma} (\bar{X} - \mu)\right] = 0, \quad \text{Var}\left[\frac{\sqrt{n}}{\sigma} (\bar{X} - \mu)\right] = 1$$

Der Zentrale Grenzwertsatz läßt sich folgendermaßen interpretieren:

Für eine große Zahl von Beobachtungen ist die Wahrscheinlichkeit dafür, für den normierten Mittelwert  $\bar{X}$  einen bestimmten Wert zu erhalten, genauso groß, als ob nur eine einzige, normalverteilte Zufallsvariable  $Z$  beobachtet würde. Der Zentrale Grenzwertsatz dient zur Begründung der sogenannten **Normalverteilungsannahme** bei vielen statistischen Verfahren.

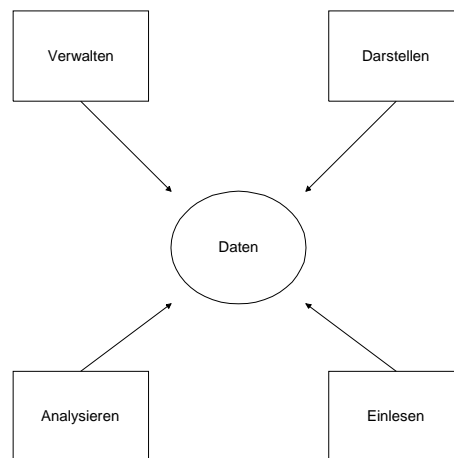
### 3. Überblick über das Statistical Analysis System

In diesem Kapitel wird der Funktionsumfang und die prinzipielle Arbeitsweise des Statistical Analysis Systems vorgestellt.

#### 3.1. Funktionsumfang des SAS Systems

Das **Statistical Analysis System** (im folgenden wie in der Originaldokumentation etwas inkonsequent mit **SAS System** abgekürzt<sup>2</sup>) ist ein umfassendes Software-System für das **Einlesen, Verwalten, Darstellen** und **Analysieren** von Daten aller Art.

Die folgende Skizze verdeutlicht die zentrale Rolle der Daten:



#### 3.2. Module des SAS Systems

Sie können das SAS System in den unterschiedlichsten Gebieten einsetzen, wobei Ihnen vorgefertigte **SAS Prozeduren** zur Verfügung stehen. Die Prozeduren eines Anwendungsgebietes sind in einem **SAS Modul** zusammengefaßt. Das SAS System besteht aus einem Basis-Modul (**SAS/Base**), das unbedingt notwendig ist, und weiteren, zusätzlich zu lizenzierenden Modulen.

<b>SAS/Stat</b>	<b>SAS/Graph</b>	<b>SAS/FSP</b>	...
<b>SAS/Base</b>			

In diesem Handbuch werden folgende SAS Module behandelt

- |    |                  |                        |
|----|------------------|------------------------|
| 1. | <b>SAS/Base</b>  | Basisfunktionalität    |
| 2. | <b>SAS/Stat</b>  | Statistische Verfahren |
| 3. | <b>SAS/Graph</b> | Graphik                |

Beispielsweise enthält das **Modul Base** u.a. folgende Prozeduren:

<sup>2</sup> Neuerdings soll SAS als Abkürzung für Strategic Analysis System stehen.

<b>PROC PRINT</b>	Ausgeben von Berichten
<b>PROC MEANS</b>	Berechnen einfacher Statistiken
<b>PROC FREQ</b>	Berechnen von Häufigkeiten
<b>PROC CHART</b>	Ausgeben eines Diagramms

Neben den zuvor genannten "bekanntem" Modulen existieren u.a. folgende weiteren SAS Module:

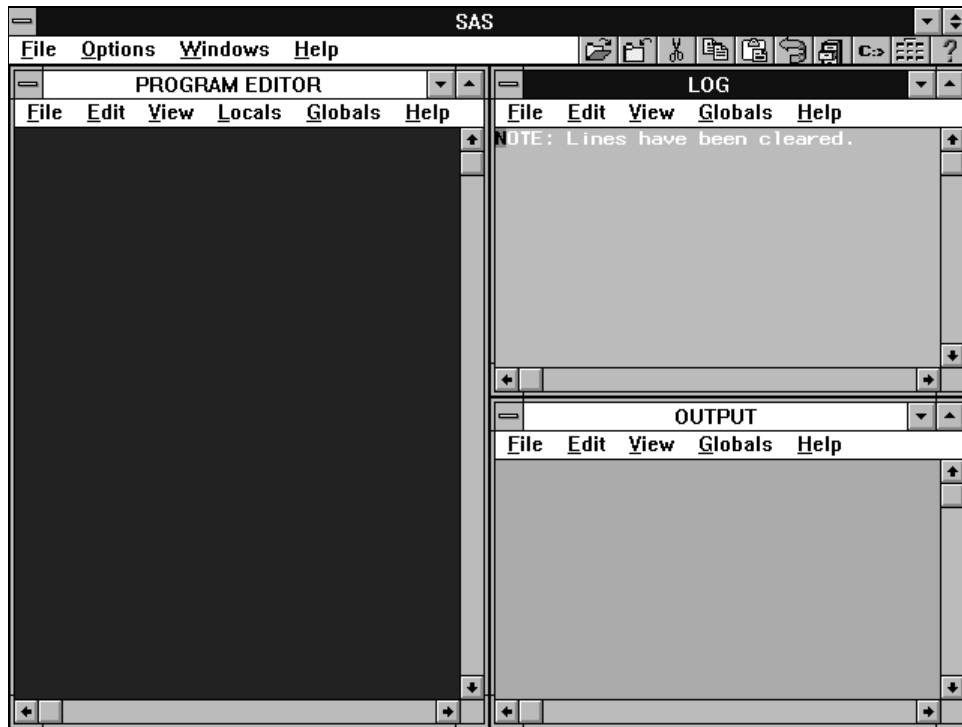
1.	<b>SAS/OR</b>	Operations Research
2.	<b>SAS/ETS</b>	Zeitreihenanalyse
3.	<b>SAS/QR</b>	Qualitätssicherung
4.	<b>SAS/Insight</b>	Explorative Datenanalyse
5.	<b>SAS/FSP</b>	Interaktive Dateneingabe, Masken
6.	<b>SAS/Spectraview</b>	Erweiterte Grafik
7.	<b>SAS/ASSIST</b>	Menügeführte Benutzerschnittstelle

### 3.3. Benutzerschnittstelle zum SAS System

In diesem Handbuch wird als Benutzerschnittstelle das **Display Manager System (DMS)** unter Windows betrachtet, das Ihnen als Kontrollzentrum (*Workbench*) für die Entwicklung und Ausführung von SAS Programmen und zum Anzeigen der Ergebnisse dient. Die weiteren für das SAS System verfügbaren Benutzerschnittstellen wie z.B. **SAS/ASSIST** werden nicht behandelt.

Sie kontrollieren das SAS System über das **Menüsystem** durch Auswählen ("Anklicken") entsprechender Menüpunkte oder alternativ mit **Kommandos**, die Sie in einer Kommandozeile (Command ==>) eingeben können<sup>3</sup>. Die folgende Abbildung zeigt das Display Manager System unter Microsoft Windows:

<sup>3</sup> Sie können zwischen den Eingabemöglichkeiten Menu, Kommando und Popup wechseln (Edit->Preference).



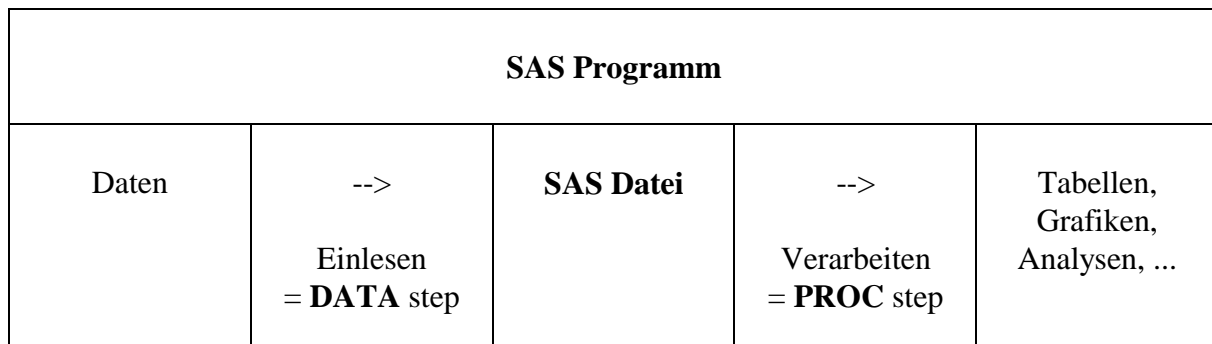
Das Display Manager System verwaltet mehrere **Fenster**, die zur Eingabe von SAS Programmen (**PROGRAM EDITOR Fenster**), zur Ausgabe von Text-Ergebnissen (**OUTPUT Fenster**), zur Ausgabe von Grafiken (**GRAPHICS Fenster**) und Meldungen des SAS Systems (**LOG Fenster**) dienen.

### 3.4. SAS Programme

Sie lösen ein Problem, indem Sie ein SAS Programm entwickeln und dieses SAS Programm vom SAS System ausführen lassen.

Jedes SAS Programm besteht aus einer Folge von **Schritten** (*steps*):

1. Jedes SAS Programm beginnt mit einem **Dateneingabeschritt** (*DATA step*), mit dem eine SAS Datei (*SAS data set*) erzeugt wird.
2. Es folgen ein oder mehrere **Verarbeitungsschritte** (*PROC steps*), die die SAS Datei verarbeiten.



Ein SAS Programm hat somit folgendes Aussehen (siehe auch folgenden Abschnitt):

```

/* Globale Anweisungen */
OPTIONS ...;
GOPTIONS ...;
...

/* Dateneingabeschritt */
DATA ...
...
RUN;

/* Verarbeitungsschritt */
PROC ...
...
RUN;

... weitere globale Anweisungen, DATA und/oder PROC Schritte

```

### 3.5. Syntax der SAS Programmiersprache

Ein SAS Programm besteht aus folgenden Teilen:

- **globale Anweisungen**                    (*global statements*)
- **Dateneingabeschritte**                (*DATA steps*)
- **Verarbeitungsschritte**                (*PROC steps*)

Globale Anweisungen beziehen sich auf **alle folgenden** Anweisungen und legen z.B. das graphische Ausgabegerät oder Kopf- und Fußzeilen fest.

Dateneingabe- oder Verarbeitungsschritte bestehen aus einer oder mehreren SAS Anweisungen, die die genaue Ausführung des Schrittes festlegen.

Eine SAS Anweisung wird durch ein **Schlüsselwort** eingeleitet (Ausnahme: Zuweisung). Die Schlüsselwörter **DATA** und **PROC** leiten einen Dateneingabeschritte bzw. einen Verarbeitungsschritt (Prozedur) ein. Jede SAS Anweisung wird mit einem **Semikolon** ';' beendet. In einer Anweisung können Sie die vom SAS System vorgenommenen Voreinstellungen durch **Argumente** und **Optionen** verändern.

Die folgende SAS Anweisung beschreibt z.B. einen Verarbeitungsschritt:

PROC	SORT	DATA=umsatz	/DESCENDING	;
------	------	-------------	-------------	---

Diese Anweisung schlüsselt sich folgendermaßen in elementare Bestandteile auf:

Bestandteil	Typ	Aufgabe
PROC	Schlüsselwort	leitet einen Verarbeitungsschritt ein.

SORT	Argument	wählt die Prozedur SORT aus.
DATA=umsatz	Argument	wählt die SAS Datei umsatz aus.
/DESCENDING	Option	sortiert in absteigender Reihenfolge (voreingestellt ist aufsteigende Reihenfolge).
;	abschließendes Semikolon	beendet die Anweisung.

Zusammengefaßt ergibt sich für eine SAS Anweisung folgende Struktur:

**Schlüsselwort Argumente [/Optionen] ;**

### 3.6. Elemente einer SAS Datei

Sie lesen in einem SAS Programm in einem Dateneingabeschritt Datenwerte in eine SAS Datei ein. Die Datenwerte müssen dabei maschinenlesbar in Form einer rechteckigen Tabelle vorliegen, wobei unterschiedliche Quellen (externe Textdateien, externe Tabellen aus Tabellenkalkulationsprogrammen und Datenbankprogrammen oder Datenwerte direkt im SAS Programm) möglich sind.

Name	Vorname	Alter	Gewicht
Meier	Werner	33	75
Müller	Sigrid	31	65

Anhand der vorstehenden einfachen Beispieltabelle `tab1` werden folgende wichtigen Begriffe erläutert:

- Datenwert            (*data value*)
- Beobachtung         (*observation, case*)
- Variable            (*variable*)
- Variablenname     (*variable name*)
- SAS Datei          (*SAS dataset, SAS file*)

Jede Zelle der Tabelle stellt für das SAS System einen **Datenwert** (*data value*) dar. Ein Datenwert ist für das SAS die kleinste Informationseinheit, die verarbeitet werden kann. Das Gewicht von Werner Meier (75) ist beispielsweise ein Datenwert:

Name	Vorname	Alter	Gewicht
Meier	Werner	33	75
Müller	Sigrid	31	65

Jede Zeile der Tabelle stellt für das SAS System eine **Beobachtung** (*observation*) dar. Eine Beobachtung setzt sich aus Informationen über ein Objekt oder eine Person zusammen. Die unterschiedlichen Informationen werden als **Variablen** (Eigenschaften oder Merkmale) bezeichnet.

Variablen können numerische Werte annehmen oder alphanumerische Werte (z.B. Name). Die Beobachtung für Werner Meier setzt sich aus beobachteten Werten von 4 Variablen zusammen:

Name	Vorname	Alter	Gewicht
Meier	Werner	33	75
Müller	Sigrid	31	65

In einem SAS Programm werden Variablen über **Variablennamen** bezeichnet, während sie in der Tabelle durch die Überschriften benannt werden:

Name	Vorname	Alter	Gewicht
Meier	Werner	33	75
Müller	Sigrid	31	65

Jede Spalte der Tabelle enthält alle beobachteten Werte für eine Variable. Ein fehlender Wert muß besonders gekennzeichnet werden. In der Spalte mit der Überschrift "Gewicht" sind alle beobachteten Werte für die Variable mit dem Namen "Gewicht" aufgelistet:

Name	Vorname	Alter	Gewicht
Meier	Werner	33	75
Müller	Sigrid	31	65

Das SAS System liest eine Tabelle, die maschinenlesbar vorliegt, im Dateneingabeschritt in eine **SAS Datei** (*SAS data set*) ein. Die SAS Datei beinhaltet alle Datenwerte und zusätzlich beschreibende Informationen wie z.B. die Variablennamen und die zugehörigen Datentypen :

Name	Vorname	Alter	Gewicht
alphanumerisch	alphanumerisch		
Meier	Werner	33	75
Müller	Sigrid	31	65

Zusammenfassend bestehen folgende Zuordnungen zwischen den Bezeichnungen:

Bezeichnungen in einer Tabelle	Bezeichnungen im SAS System
Zelle (Feldelement)	Datenwert
Zeile	Beobachtung oder Fall
Spalte	Variable
Überschrift	Variablenname
Tabelle	SAS Datei

### 3.7. Übungen

1. Ordnen Sie der folgenden Tabelle `tab2` die folgenden Begriffe zu:

Datenwert, Beobachtung, Variable und Variablenname

Name	Vorname	Fachbereich	Semester	ID
Mustermann	Marianne	FB 03	4	SAS001
Normalbuerger	Otto	FH FB E	2	SAS002

2. Welche Variablen in der vorherigen Tabelle sind numerisch (Zahlen), welche alphanumerisch (Zeichenketten)?

3. Versuchen Sie intuitiv, die folgende SAS Anweisung für die obige Tabelle zu interpretieren:

```
PROC SORT DATA=tab2; BY semester;
```

4. Nennen Sie Beispiele für Programme, in denen Daten in Tabellenform verarbeitet werden.

## 4. Einführendes Beispiel

In diesem Kapitel wird ein vollständiges SAS Programm entwickelt, ausgeführt und das Ergebnis analysiert.

### 4.1. Starten einer Sitzung

Starten Sie das SAS System durch Doppelklicken auf das Programm-Piktogramm:

Das SAS System meldet sich nun folgendermaßen:

Geben Sie im Texteingabebereich des **PROGRAM EDITOR Fensters** folgendes SAS Programm **SAMPLE1** ein, mit dem Sie eine SAS Datei `tab2` mit 4 Variablen und 2 Fällen (Beobachtungen) erzeugen und im **OUTPUT Fenster** ausdrucken. (Die genaue Bedeutung der verwendeten SAS Anweisungen wird erst in den folgenden Kapitel behandelt!)

```
PROGRAM EDITOR
Command ==>

00001 /* SAMPLE1 */
00002 DATA tab2;
00003     INPUT name $ vorname $ einkomm guthab;
00004 CARDS;
00005 Meyer   Werner   1400 3880
00006 Müller  Sigrid   2100 2400
00007 ;
00008 RUN;
00009
00010 PROC PRINT
00011     DATA=tab2;
00011 RUN;
```

### 4.2. Ausführen eines SAS Programmes

Wählen Sie *Locals -> Submit*, um das im PROGRAM EDITOR Fenster eingegebene SAS Programm **SAMPLE1** auszuführen.

Das SAS System führt nun das SAS Programm Schritt für Schritt aus und zeigt Ihnen im **OUTPUT Fenster** das Ergebnis der Prozedur **PRINT** an:

```
OUTPUT
Command ==>

      OBS    NAME          VORNAME          EINKOMM          GUTHAB
      1     Meyer          Werner           1400             3880
      2     Müller         Sigrid           2100             2400
```

Beachten Sie, daß

- der Dateneingabeschritt (*DATA step*) kein Ergebnis im OUTPUT Fenster liefert, sondern ausschließlich die SAS Datei `tab2` für den folgenden Verarbeitungsschritt erzeugt.
- das SAS System automatisch eine zusätzliche Variable `OBS` erzeugt hat, die die Beobachtungen fortlaufend numeriert.
- Sie im **LOG Fenster** informative Meldungen über die Ausführung der einzelnen Schritte angezeigt bekommen.

#### 4.3. Speichern eines Programmes und Beenden einer SAS Sitzung

Wechseln Sie über *Fenster -> Program Editor* in das Program Editor Fenster. Wählen Sie *Local-> Recall*, um das zuletzt bearbeitete SAS Programm wieder anzuzeigen:

Speichern Sie nun das SAS Programm über *File -> Save* in die Datei `SAMPLE1 . SAS`.

Verlassen Sie nun das SAS System über *File -> Exit SAS*.

#### 4.4. Übungen

1. Führen Sie die genannten Schritte an einem System aus (z.B. an einem PC mit SAS für DOS).
2. Erzeugen Sie einen Ausdruck Ihres SAS Programmes und der Ausgabe auf Papier.
3. Speichern Sie Ihr SAS Programm und die Ausgabe zur Sicherheit auf Diskette ab.

## 5. Bedienen der Benutzeroberfläche

In diesem Kapitel wird die Bedienung des **Display Manager Systems** (DMS) beschrieben.

### 5.1. Elemente der Benutzeroberfläche

Sie haben im vorherigen Kapitel bereits die grundsätzliche Funktionsweise und Bedienung des **Display Manager Systems** (DMS) kennengelernt. Wesentliche Elemente des DMS sind Eingabe- und Ausgabefenster, die eine interaktive Entwicklung von SAS Programmen ermöglichen.

Sie können u.a. in folgenden Fenstern arbeiten, wobei das PGM Fenster nur zur Eingabe und das OUTPUT und LOG Fenster nur zur Ausgabe dienen:

Name des Fensters	Funktion
PROGRAM EDITOR Fenster (kurz: PGM Fenster)	Erfassen, Speichern, Laden und Ausführen von SAS Programmen
OUTPUT Fenster	Ausgeben der Ergebnisse eines SAS Programms
LOG Fenster	Ausgeben der Meldungen beim Ausführen eines SAS Programms

Beim Start des SAS Systems sind die Fenster zunächst folgendermaßen angeordnet:

<b>OUTPUT</b> Command ==> <Hier erscheinen Ergebnisse des SAS Systems.>
<b>LOG</b> Command ==> <Hier erscheinen Meldungen des SAS Systems.>
<b>PROGRAM EDITOR</b> Command ==>  00001 <Hier steht die erste Programmzeile.> ...

Sie können die Größe und Anordnung der Fenster über die entsprechenden Menüpunkte verändern, z.B. über Fenster -> Nebeneinander alle Fenster nebeneinander anordnen.

### 5.2. Wichtige Menüpunkte

Sie können u.a. die folgenden Kommandos in der Kommandozeile eingeben:

**Einlesen und Speichern von Dateien und Beenden einer Sitzung:**

Menüpunkt	Bedeutung	
File->Save	schreibt den Inhalt des aktuellen Fensters in die Datei <i>fid</i> .	
File->Open	lädt den Inhalt der Datei <i>fid</i> in das PGM-Fenster.	
File->Exit SAS	beendet die SAS Sitzung.	

**Ausführen und Zurückholen von SAS Programmen:**

Kommando	Bedeutung	
<i>Locals-&gt;Recall</i> [nur im PGM Fenster]	holt das zuletzt ausgeführte SAS Programm in das PGM EDITOR Fenster zurück.	
<i>Locals-&gt;Submit</i> [nur im PGM Fenster]	führt das SAS Programm im PGM EDITOR Fenster aus.	

**Wechseln in ein Fenster:**

Kommando	Bedeutung	
?	wechselt in das HELP Menu System.	
Fenster-> Log	wechselt in das LOG Fenster.	
Fenster-> Output	wechselt in das OUTPUT Fenster.	
Fenster-> Program Editor	wechselt in das PROGRAM Editor Fenster.	

**5.4. Interaktives Entwickeln von SAS Programmen**

Sie können ein SAS Programm mit einem beliebigen Editor (z.B. **vi**, **KEDIT** oder **XEDIT**) außerhalb des SAS Systems erfassen und danach in das PROGRAM EDITOR Fenster laden. Anschließend können Sie es mit dem Kommando **SUBMIT** ausführen.

**Sie können ein SAS Programm alternativ (wie im einführenden Beispiel) auch direkt im PROGRAM EDITOR Fenster erfassen.**

**5.5. Zusammenfassung**

Sie können die Entwicklung von SAS Programmen vollständig im Display Manager System (DMS) durchführen.

Im einzelnen führen Sie folgende Arbeitsschritte zur Entwicklung eines SAS Programm im Display Manager System aus:

1. Erfassen oder Laden eines SAS Programms
2. Speichern (nur bei neuem oder verändertem SAS Programm)
3. Ausführen des SAS Programms
4. Kontrollieren der Ausgabe im OUTPUT und des Protokolls im LOG Fenster
5. Speichern der Ausgabe (falls gewünscht)
6. Zurückholen des SAS Programms und Korrigieren ... (erfahrungsgemäß)

## 5.6. Übungen

## 6. Einlesen von Datenwerten in eine SAS Datei

In diesem Kapitel werden einige SAS Programme zum Einlesen von Daten aus Textdateien vorgestellt. Sie unterscheiden sich nur in Hinblick auf das Format und die Quelle der einzulesenden Datenwerte.

Sie können prinzipiell Daten aus Textdateien einlesen, solange Sie den Namen und den Aufbau der Datei kennen, in der sich die Rohdaten befinden.

Die abschließenden Syntaxdiagramme sind **nicht** vollständig, sondern fassen lediglich die in den Beispielen verwendeten SAS Anweisungen zusammen und ergänzen sie um einige Argumente und Optionen.

### 6.1. DATA - Listengesteuertes Einlesen

In diesem Abschnitt wird zunächst der einfache Fall behandelt, bei dem die Datenwerte jeder Beobachtung in genau einer Eingabezeile stehen und direkt im SAS Programm enthalten sind. Die Trennung zwischen Programm und Daten erfolgt durch eine **CARDS** Anweisung.

Im folgenden Beispielprogramm **SAMPLE2** werden die Datenwerte aus der bekannten Tabelle **tab2** **listengesteuert** eingelesen; d.h. das SAS System liest die Datenwerte in Form einer Liste von Daten ein, wobei die einzelnen Listenelemente durch Leerzeichen voneinander getrennt sein müssen:

<pre>/* SAMPLE2 */  DATA tab2; INPUT     name \$     vorname \$     einkomm     guthab     ; CARDS ;  Meyer Werner 1400 3880 Müller Sigrid 2100 2400 ;  PROC PRINT;  RUN;</pre>	<p>Ein Kommentar wird durch /* ... */ eingeschlossen.</p> <p>Die SAS Datei erhält den Namen <b>tab2</b>.</p> <p>Die SAS Datei enthält Beobachtungen für 4 Variablen mit den Namen <b>name</b>, <b>vorname</b>, <b>einkomm</b> und <b>guthab</b>.</p> <p>Alphanumerische Variablen werden durch das Zeichen '\$' gekennzeichnet.</p> <p>Die SAS Anweisung <b>CARDS</b> bewirkt, daß das SAS System die folgenden Zeilen als <b>Beobachtungen</b> interpretiert.</p> <p>Beobachtung 1</p> <p>Beobachtung 2</p> <p>Die Liste der Beobachtungen wird durch ein Semikolon ';' in Spalte 1 beendet. Das SAS System erwartet ab hier wieder SAS Programm-Anweisungen.</p> <p>Die eingelesenen Datenwerte werden zur Kontrolle im OUTPUT Fenster ausgegeben.</p> <p>Die SAS Anweisung <b>RUN</b> startet die vorhergehenden Anweisungen.</p>
---	--

Beachten Sie, daß

- Variablenamen mit einem Buchstaben beginnen und ggf. auch Ziffern enthalten dürfen.
- Variablenamen und Namen von SAS Dateien höchstens 8 Zeichen lang sein dürfen.

- alphanumerische Werte (Zeichenketten) beim freien Eingabeformat nicht länger als 8 Zeichen sein dürfen und keine Leerzeichen enthalten dürfen (siehe nächsten Abschnitt).
- fehlende numerische Datenwerte mit einem Punkt (.) eingegeben werden müssen, fehlende Zeichenketten durch einen benutzerdefinierten fehlenden Wert wie z.B. kA (keine Angaben).
- zwischen zwei Datenwerten mindestens ein Leerzeichen stehen muß.

Führen das SAS Programm über *Locals* -> *Submit* aus.

Sie erhalten nun im LOG Fenster folgende informative Meldungen vom SAS System:

```
LOG
Command ==>

      23   data tab2;
      24   input name $ vorname $ einkomm guthab;
      25   cards;
      28   ;
NOTE: The data set WORK.tab2 has 2 observations and 4 variables. NOTE: The
DATA statement used 3.00 seconds.
      29   run;
      30   proc print data=tab2;
      31   run;
NOTE: The PROCEDURE PRINT used 2.00 seconds.
```

Die Syntax für den listengesteuerten Dateneingabeschritt lautet:

<pre><b>DATA</b> [<i>SAS_file</i>]; <b>INPUT</b> <i>var1</i> [\$] <i>var2</i> [\$] ...; <b>CARDS</b>; <input ;<="" line(s)="" pre=""/> </pre>	<pre>... Schlüsselwort: DATA ... \$ bezeichnet alphanumerische Variable ... Datenwerte folgen im Programm.</pre>
---	--

## 6.2. DATA - Spaltengesteuertes Einlesen

Sie müssen bei der **INPUT** Anweisung die genauen Spaltenpositionen der einzelnen Datenwerte angeben, wenn die Bedingungen für das freie Eingabeformat nicht erfüllt sind (z.B. bei einer Zeichenkette mit mehr als 8 Zeichen).

Im folgenden Beispielprogramm `SAMPLE3` mit der SAS Datei `tab2` werden die Datenwerte **spaltenpositioniert** eingelesen:

```
/* SAMPLE3 */
DATA tab2;
```

```

INPUT
    name      $      1-7
    vorname   $      8-13
    einkomm   16-19
    guthab    21-24;

CARDS ;
Meyer Werner 1400 3880
Müller Sigrid 2100 2400
;
/*
123456789012345678901234
*/
RUN ;

```

Die Position der zugehörigen Datenwerte in den Eingabezeilen wird durch Bereiche angegeben, z.B. steht die Variable name in Spalte 1 bis 7.

Beobachtung 1, spaltenpositioniert  
Beobachtung 2, spaltenpositioniert.

Diese Zeile dient nur zur Verdeutlichung der Spaltenpositionen.

Beachten Sie, daß

- die Datenwerte nach der **CARDS** Anweisung exakt in den verlangten Spaltenpositionen stehen müssen.
- für jede Variable der "längste" Datenwert bestimmt werden muß, damit genügend Platz reserviert werden kann.
- fehlende Datenwerte nicht besonders gekennzeichnet werden müssen, da Leerzeichen automatisch als fehlender Datenwert interpretiert werden.

Die Syntax für den Dateneingabeschritt mit spaltengesteuertem Eingabeformat lautet:

```

DATA [SAS_file];
INPUT    var1 [$] x1-x2      x1-x2: Spaltenbereich für var1
          var2 [$] y1-y2...;  y1-y2: Spaltenbereich für var2

CARDS;
input line(s)
;

```

### 6.3. DATA - Einlesen von mehreren Eingabezeilen

Sie können eine Beobachtung von mehreren Eingabezeilen einlesen, wenn eine Eingabezeile nicht ausreicht (z.B. bei sehr langen oder sehr vielen Zeichenketten oder zum Überlesen von Kommentaren).

Im folgenden Beispielprogramm **SAMPLE4** mit der SAS Datei **tab2** werden die Datenwerte einer Beobachtung von 2 Eingabezeilen eingelesen:

```

/* SAMPLE4 */
DATA tab2;
INPUT
    #1 name $      1-7
      vorname $    8-23
    #2 einkomm     1-4
      guthab       6-9;

```

Die Datenwerte für name und vorname stehen jeweils in der ersten Eingabezeile, die Datenwerte für einkomm und guthab jeweils in der zweiten.

```

CARDS;
Meyer Werner-Siegfried           Beobachtung 1, Eingabezeile 1
1400 3880                       Beobachtung 1, Eingabezeile 2
Müller Sigrid-Brigitte          Beobachtung 2, Eingabezeile 1
2100 2400                       Beobachtung 2, Eingabezeile 2
;
/*
123456789012345678901234       Diese Zeile dient nur zur Verdeutlichung der Spal-
*/
RUN;

```

Die Syntax für den Dateneingabeschritt mit mehreren Eingabezeilen pro Beobachtung lautet:

```

DATA [SAS_file];
    INPUT      #1 var_1 var_2 ...      Eingabezeile 1
                #2 var_7 var_8 ...    Eingabezeile 2
                ...;
    CARDS;
    input lines
    ;

```

#### 6.4. DATA - Mehrfaches Einlesen von einer Eingabezeile

Sie können im listengesteuerten Eingabeformat das SAS System mit der Formatsteuerung @@ anweisen, daß in **einer** Eingabezeile Datenwerte für **mehrere** Beobachtungen stehen.

Im folgenden Beispielprogramm SAMPLE5 bestehen die Beobachtungen aus Datenwerten für eine Variable gewicht mit mehreren Beobachtungen in einer Eingabezeile:

```

/* SAMPLE5 */
DATA tab2;
INPUT      gewicht                Die Formatsteuerung @@ bewirkt, daß mehrere Be-
                @@;                obachtungen von einer Zeile gelesen werden.
CARDS;
57 54 78 90 65 56 98             Beobachtungen 1-7
54 66 77 82                       Beobachtungen 8-11
;
RUN;

```

Die Syntax für den Dateneingabeschritt mit mehreren Beobachtungen pro Eingabezeile lautet:

```

DATA [SAS_file];
    INPUT      var1 [$]
                var2 [$]
                ... @@;           Ende der Variablenliste
    CARDS;
    input line(s)

```

### 6.5. DATA - Einlesen aus einer externen Datei

Sie können die Datenwerte aus einer externen Datei einlesen; d.h. die Daten stehen nicht mehr im Programm. In diesem Fall müssen Sie die **CARDS** Anweisung durch eine **INFILE** Anweisung ersetzen.

Im folgenden Beispielprogramm SAMPLE6 werden die Datenwerte aus der externen Datei INPUT.DAT eingelesen (Dateiname gültig unter DOS):

```

/* SAMPLE6 */
FILENAME DATAFILE 'INPUT.DAT';

DATA tab3;
INFILE DATAFILE PAD MISSOVER4;

INPUT
    name $          1-7
    vorname $       8-23
    einkomm         1-4
    guthab          6-9;
RUN;

```

Der interne Dateiname DATAFILE wird mit dem externen Dateinamen INPUT.DAT (unter DOS) verknüpft.

Die SAS Anweisung **INFILE** bewirkt, daß das SAS System die Datenwerte aus der Datei DATAFILE einliest.

Der interne Name DATAFILE ist über eine **FILENAME** Anweisung mit einer externen Datei INPUT.DAT verknüpft.

Die Syntax für den Dateneingabeschritt mit Einlesen der Datenwerte aus einer externen Datei lautet:

```

FILENAME datafile 'filename';      filename ist systemabhängig!
DATA [SAS_file];
INFILE datafile
[PAD] [MISSOVER] [FLOWOVER];
INPUT ...;

```

Beachten Sie, daß

- die Datenwerte in der externen Datei exakt der Beschreibung in der **INPUT** Anweisung entsprechen müssen.
- der vom SAS System verwendete (interne) Name *datafile* durch eine vorangehende **FILENAME** Anweisung mit einem externen Dateinamen *filename* verknüpft werden muß.

### 6.6. PROC CONTENTS - Auflisten von Informationen

<sup>4</sup> Die Option PAD bewirkt, daß fehlende Zeichen aufgefüllt werden (padding). Die Option MISSOVER bewirkt, daß fehlende Variablen nicht von der nächsten Eingabezeile gelesen werden, sondern als fehlend eingetragen werden.

Die Prozedur **CONTENTS** gibt eine **Beschreibung** einer SAS Datei aus. (Verwechseln Sie diese Prozedur nicht mit der Prozedur **PRINT**, die den **Inhalt** ausgibt!)

Im folgenden Beispielprogramm **SAMPLE7** werden Informationen über die SAS Datei **tab4** ausgegeben:

```
/* SAMP7 * /
DATA tab4;
...;

PROC CONTENTS DATA=tab4;          PROC CONTENTS liefert Informationen
                                   über die SAS Datei tab4.
RUN;
```

Die Prozedur **CONTENTS** erzeugt folgende Ausgabe im **OUTPUT** Fenster:

OUTPUT					
Command ==>					
<b>CONTENTS PROCEDURE</b>					
Data Set Name:	WORK.tab4	Type:			
Observations:	2			Record Len:	36
Variables:	4				
Label:					
-----Alphabetic List of Variables and Attributes-----					
#	Variable	Type	Len	Pos	Label
3	EINKOMM	Num	8	20	
4	GUTHAB	Num	8	28	
1	NAME	Char	8	4	
2	VORNAME	Char	8	12	

Beachten Sie, daß

- es zwei numerische und zwei alphanumerische Variablen mit jeweils der Länge 8 gibt.
- die SAS Datei 2 Beobachtungen enthält.

### 6.7. Zusammenfassung

In diesem Kapitel sind folgende Varianten des Dateneingabeschrittes behandelt worden:

numerische Datenwerte (Zahlen)	alphanumerische Datenwerte (Zeichenketten)
<b>DATA ...;</b> <b>INPUT var1;</b> /* Zahl */ <b>CARDS;</b> 123 ;	<b>DATA ...;</b> <b>INPUT var1 \$;</b> /* Zeichenkette */ <b>CARDS;</b> Mueller ;

<b>listengesteuerte Eingabe</b>	<b>spaltengesteuerte Eingabe</b>
<b>DATA ...;</b> <b>INPUT</b> <i>var1</i> [\$] <i>var2</i> [\$] ...; <b>CARDS;</b> <i>input line(s)</i> ; ;	<b>DATA ...;</b> <b>INPUT</b> <i>var1</i> [\$] 1-10 <i>var2</i> [\$] 11-20 ...; <b>CARDS;</b> <i>input line(s)</i> ; ;
<b>mehrere Beobachtungen pro Eingabezeile</b>	<b>mehrere Eingabezeilen pro Beobachtung</b>
<b>DATA ...;</b> <b>INPUT</b> <i>var1</i> [\$] <i>var2</i> [\$] ... @@; <b>CARDS;</b> <i>input line(s)</i> ; ;	<b>DATA ...;</b> <b>INPUT</b> #1 <i>var1</i> [\$] ... #2 <i>varx</i> [\$] ...; <b>CARDS;</b> <i>input line(s)</i> ; ;
<b>Datenwerte innerhalb des SAS Programms</b>	<b>Datenwerte aus einer externen Datei</b>
<b>DATA ...;</b> <b>INPUT</b> <i>var1</i> [\$] ...; <b>CARDS;</b> <i>input line(s)</i> ; ;	<b>FILENAME</b> DATAFILE '...'; <b>DATA ...;</b> <b>INFILE</b> DATAFILE; <b>INPUT</b> <i>var1</i> [\$] ...; ;

## 6.8. Übungen

1. Schreiben Sie ein SAS Programm `SAMPLE8`, mit dem Beobachtungen auf Grundlage des folgenden Fragebogens eingelesen und ausgegeben werden können:

Umfrage zur Qualität des Mensaeessens	Kodierung
1. Alter in Jahren:  _ _	_ _
2. Geschlecht: o 1=männlich o 2=weiblich	_
3. Status: o 1=Student o 2=Wiss. Mitarbeiter o 3=Dozent o 4=sonstige	_
4. Die Qualitaet des Essens ist: o 1=sehr gut o 2=gut o 3=befriedigend	_
4. Der Preis für das Essens ist: o 1=niedrig o 2=angemessen o 3=ueberteuert	_

**Hinweis:**

Füllen Sie zunächst einige Fragebögen (fiktiv) aus und verwenden Sie dann im Programm **SAMPLE9** einen Dateneingabeschritt **DATA** mit den Variablen **alter**, **sex**, **status**, **qual** und **preis** und die Prozedur **PRINT** zur Kontrollausgabe der erzeugten SAS Datei.

2. Bauen Sie in Ihr SAS Programm absichtlich einen Fehler ein, indem Sie z.B. ein Semikolon löschen, und untersuchen Sie die vom SAS System im **LOG Fenster** angezeigten Meldungen.
3. Lesen Sie im SAS Programm **SAMPLE10** die folgenden Datenwerte mit einem geeignetem Dateneingabeschritt ein:  
(Treten beim listengesteuerten Einlesen Probleme auf ?)

Deutschland Berlin Großbritannien London Niederlande Den Haag Belgien Brüssel  
Dänemark Kopenhagen Frankreich Paris

4. Lesen Sie nun im SAS Programm **SAMPLE11** die selben Datenwerte aus einer externen Datei **EUROPA.DAT** ein.
5. Fügen Sie Kommentare in Ihre Programme ein.

**Hinweis:**

Kommentare werden durch die Zeichen **/\*** eingeleitet und durch **\*/** beendet und können sich auch über mehrere Zeilen erstrecken, z.B:

```
/*
Analyse von Bodenproben
Zeitraum: ...
*/
```

## 7. Sortieren und tabellarisches Darstellen einer SAS Datei

In diesem Kapitel werden elementare Prozeduren zum Sortieren und tabellarischen Darstellen einer SAS Datei beschrieben. Die einfache bzw. erweiterte Syntaxbeschreibung am Ende eines Abschnittes enthalten nur eine kleine Auswahl der möglichen Optionen und Argumente.

Achtung: Aus Platzgründen enthalten SAS Beispielprogramme ab diesem Kapitel keine abschließende **RUN** Anweisungen mehr. Fügen Sie Ihren Beispielprogrammen mindestens eine abschließende **RUN** Anweisung hinzu.

### 7.1. PROC PRINT - Auflisten einer SAS Datei

Die Prozedur **PRINT** dient zum Ausgeben aller Beobachtungen einer SAS Datei im OUTPUT Fenster.

Im folgenden Beispielprogramm SAMPLE12 gibt die Prozedur **PRINT** die SAS Datei tab5 aus:

```

/* SAMPLE 12 */
DATA tab5;
    INPUT sex $ alter groesse @@;
CARDS;
m 26 175 m 27 190 f 26 178 m 26 179 f 26 180
;
PROC SORT DATA=tab5;
    BY sex; /* Sortierung nach der Gruppenvariablen sex */
PROC PRINT DATA=tab5;
    VAR alter groesse;
    BY sex; /* Ausgabe getrennt nach Gruppen */

```

Beachten Sie, daß

- die **VAR** Anweisung die Reihenfolge der auszugebenden Variablen festlegt.
- die **BY** Anweisung festlegt, daß die Ausgabe sortiert nach Gruppen (hier getrennt nach Männern und Frauen) erfolgen soll und daß die **BY** Anweisung nur Gruppen enthalten darf, die vorher durch die Prozedur **SORT** sortiert wurden.

Das Programm SAMPLE12 erzeugt folgende Ausgabe im OUTPUT Fenster:

```

OUTPUT
Command ==>
----- SEX=f -----
      OBS   ALTER   GROESSE
      1     26      178
      2     26      180
----- SEX=m -----
      OBS   ALTER   GROESSE
      3     26      175
      4     26      179
      5     27      190

```

Die Syntax für die Prozedur **PRINT** lautet:

<b>PROC PRINT;</b>	gibt die aktuelle SAS Datei mit allen Variablen aus. *
<b>PROC PRINT DATA=SAS_file</b>	... gibt die SAS Datei ... aus.
<b>VAR var1 ..</b>	... nur die Variablen .
<b>BY varx ...;</b>	... gruppiert nach ...
<b>SUM vary ...;</b>	... Summation für ...

### 7.2. PROC FREQ - Berechnen von Häufigkeiten

Die Prozedur **FREQ** dient zum Berechnen von Häufigkeiten für einzelne Variablen und zum Berechnen von Häufigkeiten in Kreuztabellen.

Im folgenden Beispielprogramm **SAMPLE13** mit der SAS Datei **tab6** gibt die Prozedur **FREQ** Häufigkeiten für die Variablen **sex**, **alter** und **groesse** und eine Kreuztabelle der Häufigkeiten für die Variablen **sex** und **alter** aus:

```

/* SAMPLE 13 */
DATA tab7;
    INPUT sex $ alter groesse @@;
CARDS;
m 26 175 m 27 190 f 26 178 m 26 179 f 26 180
;
PROC FREQ DATA=tab7;
    TABLES sex alter groesse sex*alter;

```

Das Programm **SAMPLE13** erzeugt folgende Ausgabe (gekürzt):

```

OUTPUT
Command ==>>

```

<b>SEX</b>	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	2	40.0	2	40.0
m	3	60.0	5	100.0

TABLE OF SEX BY ALTER			
SEX ALTER			
Frequency			
Percent			
Row Pct			
Col Pct	26	27	Total
f	2	0	2
	40.00	0.00	40.00
	100.00	0.00	
	50.00	0.00	
m	2	1	3
	40.00	20.00	60.00
	66.67	33.33	
	50.00	100.00	
Total	4	1	5
	80.00	20.00	100.00

Beachten Sie, daß

- die Prozedur in der Kreuztabelle für jede Zelle zusätzlich zur Häufigkeit und zum Prozentanteil auch den Prozentanteil der Zelle bzgl. der Zeile und der Spalte berechnet.
- für die Kombination (f, 26) die Häufigkeit 2 gezählt wird, d.h. es gibt 2 Beobachtungen mit 26-jährigen Frauen.

Die Syntax für die Prozedur **FREQ** lautet:

**PROC FREQ;**

**TABLES** var1 ...;

**PROC FREQ DATA=SAS\_file;**

**TABLES** var1 ... var1\*var2 ...

**BY** varz ...

### 7.3. PROC MEANS - Berechnen von Maßzahlen

Die Prozedur **MEANS** berechnet u.a. Mittelwert, Standardabweichung, größten und kleinsten Wert, Summe über alle Werte und Spannweite von numerischen Variablen.

Im folgenden Beispielprogramm **SAMPLE14** mit der SAS Datei **tab8** berechnet **PROC MEANS** Maßzahlen für die Variablen **alter** und **gewicht**:

```

/* SAMPLE14 */
DATA tab8;
    INPUT alter gewicht @@;
CARDS;
20 80 23 65 18 60 21 56 . 60 19 .
;
PROC MEANS;

```

Das Programm **SAMPLE14** liefert im im **OUTPUT**-Fenster folgende Ausgabe:

OUTPUT

Command ==>						
N_Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
6	ALTER	5	18.00	23.00	20.20	1.92
	GEWICHT	5	56.00	80.00	64.20	9.39

Beachten Sie, daß

- fehlende Werte bei den Berechnungen **nicht** berücksichtigt werden.

Die Syntax für die Prozedur **MEANS** lautet:

**PROC MEANS;** berechnet Standardstatistiken für alle numerischen Variablen.

Die Prozedur **UNIVARIATE** liefert die selben Ergebnisse wie die Prozedur **MEANS** und ggf. zusätzliche Maßzahlen:

**PROC UNIVARIATE;** berechnet Standardstatistiken für alle numerischen Variablen (größerer Funktionsumfang).

#### 7.4. Zusammenfassung

In diesem Kapitel sind folgende SAS Prozeduren behandelt worden:

Prozedur	Funktion	OUTPUT Fenster
<b>PROC FREQ;</b> <b>TABLES ...;</b>	berechnet Häufigkeiten.	Häufigkeitstabellen und Kreuztabellen
<b>PROC MEANS;</b>	berechnet Maßzahlen.	Maßzahlen wie Mittelwert, Summe und kleinster Wert
<b>PROC PRINT;</b>	gibt eine SAS Datei aus.	tabellierte SAS Datei
<b>PROC SORT;</b> <b>BY var1 ...;</b>	sortiert eine SAS Datei nach einer oder mehreren Variablen.	keine Ausgabe
<b>PROC UNIVARIATE;</b>	berechnet Maßzahlen, Quantile und Test-Statistiken.	diverse Maßzahlen und Plots

In diesem Kapitel sind ferner folgende SAS Anweisungen behandelt worden:

Anweisung	Funktion	Bemerkungen
<b>BY varx ...;</b>	definiert Variable(n) für Gruppeneinteilung.	Vorheriges Sortieren durch <b>SORT</b> ist notwendig.

<b>VAR</b> <i>var1</i> ...;	definiert Variable(n) zur Bearbeitung.	Voreinstellung ist zumeist, daß alle Variablen bearbeitet werden.
<b>CLASS</b>	definiert Variablen für Gruppeneinteilung.	vergleichbar zu <b>BY</b> , aber: keine vorherige Sortierung notwendig!

### 7.5. Übungen

1. Führen Sie im Programm `SAMPLE15` die Berechnungen der Maßzahlen und Häufigkeiten für die Variablen der SAS Datei durch, die auf Grundlage der Mensa-Umfrage erstellt wurde (**PROC MEANS**). Ergänzen Sie ggfs. einige weitere (fiktive) Beobachtungen.
2. Führen Sie nun Programm `SAMPLE16` die gleichen Berechnungen getrennt nach Status und Geschlecht durch. Sortieren Sie hierzu zunächst nach den Variablen `status` und `sex` und verwenden Sie in der Prozedur **MEANS** die **BY** Anweisung mit den Variablen `status` und `sex`.
3. Führen Sie zusätzlich eine Kreuztabulation für `alter` und `qual` durch. Berechnen Sie Häufigkeiten für `preis` und `qual` getrennt nach männlichen und weiblichen Besuchern.
4. Verwenden Sie nun im Programm `SAMPLE16` die Optionen **MAXDEC=3** (Anzahl Dezimalstellen) und **MISSING** (fehlende Werte als weitere Gruppe) und vergleichen Sie mit den bisherigen Ergebnissen (**PROC MEANS MAXDEC=3 MISSING;**).
7. Verwenden Sie in einem Beispielprogramm zur **MEANS** Prozedur die **CLASS** Anweisung statt der **BY** Anweisung. Ist eine vorherige Sortierung notwendig?

## 8. Grafisches Darstellen einer SAS Datei

In diesem Kapitel werden elementare Prozeduren zum grafischen Darstellen einer SAS Datei beschrieben. Die einfache bzw. erweiterte Syntaxbeschreibung am Ende eines Abschnittes enthalten nur eine Auswahl der möglichen Optionen und Argumente.

### 8.1. PROC GCHART - Erzeugen von Diagrammen

Die Prozedur **GCHART** (bzw. **CHART** zum Testen) dient zur graphischen Darstellung von Datenwerten in Diagrammen (Balken, Säulen, Kreise bzw. Sterne).

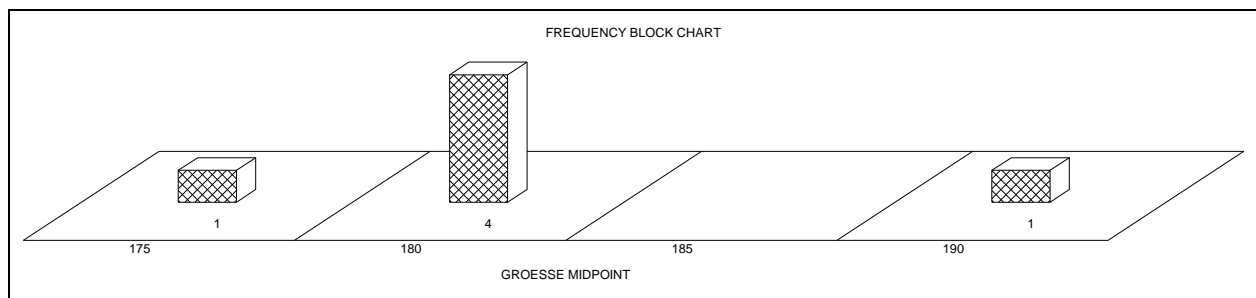
Im folgenden Beispielprogramm **SAMPLE17** mit der SAS Datei **tab9** stellt die Prozedur **GCHART** für die Variable **groesse** Häufigkeiten in Form von 3D-Säulen dar:

Sie können über eine vorausgehende **GOPTIONS** Anweisung festlegen, daß das folgende Diagramm in eine Graphik-Datei gespeichert wird, wobei Sie unter einer Vielzahl von Grafikformaten auswählen können (z.B. bietet sich für den Import nach **Word für Windows** **PSEPSF** oder **CGMMWWA** an). Hierzu sind folgende Änderungen notwendig:

```
FILENAME plotout 'sas1.cgm';
GOPTIONS gsfmode=replace gsfname=plotout device=cgmmwwa;
```

```
/* SAMPLE 17 */
DATA tab9;
    INPUT sex $ alter groesse @@;
CARDS;
m 26 175 m 27 190 f 26 178 m 26 179 f 26 180 f 28 178
;
RUN;
PROC GCHART DATA=tab9;
    BLOCK groesse; /* Blockdiagramm fuer die Variable groesse */
```

Das Programm **SAMPLE17** erzeugt folgende graphische Ausgabe<sup>5</sup>:



Die Syntax für die Prozedur **GCHART** für ein vertikales Balkendiagramm lautet:

```
PROC GCHART;
VBAR;
```

<sup>5</sup> Die Abbildung wurde als CGM Graphikdatei in dieses Dokument eingefügt.

**PROC GCHART DATA=SAS\_file;**

**VBAR** var1 ...;

**/MIDPOINTS=** mp1 mp2 ...;

[bzw. **/LEVELS=** n]

[bzw. **/DISCRETE**];

... nur die Variablen ...

Intervall-Mittelpunkte ...

Anzahl Balken =  $n$

Balken für jede Ausprägung (jeden unterschiedlichen Wert)

Beachten Sie, daß

- Sie für jeden Wert eine Säule erzeugen können (Option **/DISCRETE**)
- Sie Datenwerte aus einem Intervall zusammenzufassen können (Option **/MIDPOINTS**) und daß die Zahlen  $mp1$ ,  $mp2$ , ... die Mittelpunkte der Intervalle angeben, für die Balken erzeugt werden sollen .
- Sie die Anzahl der Säulen vorzugeben können (Option **/LEVELS**) und daß die Zahl  $n$  die Anzahl der Balken angibt.
- sich die Optionen **/DISCRETE**, **/MIDPOINTS** und **/LEVELS** ausschließen.
- die **VBAR** Anweisung für einen vertikalen Balken (*vertical bar*) durch eine der folgende Anweisungen ersetzt werden kann:

<b>HBAR</b>	horizontaler Balken
<b>BLOCK</b>	3D-Quader
<b>PIE</b>	Tortendiagramm

## 8.2. PROC GPLOT - Erzeugen von Streudiagrammen

Die Prozedur **GPLOT** (bzw. **PLOT** zum Testen) dient zur graphischen Darstellung von Datenpunkten (jeweils zwei Datenwerte einer Beobachtung) in einem **Streudiagramm** oder **X-Y-Plot**.

Im folgenden Beispielprogramm **SAMPLE18** mit der SAS Datei `tab10` trägt die Prozedur **GPLOT** die Variablen `jahr` und `anzahl` (Jahr 19xx und Bevölkerung in Deutschland in Tausend) gegeneinander auf:

```

/* SAMPLE 18 */
DATA tab10;
    INPUT jahr anzahl @@;
CARDS;
50 68377 55 70326 60 72674 65 74963 70 77709 75 78996
;
SYMBOL1 INTERPOL=spline;          /* Kurve durch Datenpunkte */
PROC GPLOT DATA=tab10;
    PLOT anzahl*jahr;

```

Das Programm **SAMPLE18** erzeugt einen X-Y-Plot der Beobachtungspunkte mit einer Interpolation zwischen den Beobachtungspunkten.

Die Syntax für die Prozedur **GPLOT** lautet:

<b>SYMOL1 INTERPOL=spline;</b>	glatte Kurve durch Datenpunkte
<b>PROC GPLOT [DATA=SAS_file];</b>	
<b>PLOT y1*x1 y2*x2 ..</b>	y-x-Plots für ...
<b>[/OVERLAY];</b>	Option Überlagerung

Beachten Sie, daß

- in der **PLOT** Anweisung die erste Variable *y1* in vertikaler (Y) und die zweite Variable *x1* in horizontaler (X) Richtung aufgetragen wird.
- Sie mit der Option **OVERLAY** mehrere X-Y-Plots übereinander ausgeben können (*Overlay*).

<b>PROC GCHART;</b> <b>VBAR ... ;</b>	gibt Diagramme aus.	Balken-, Säulen-, Stern- und Torten-Diagramme
<b>PROC GPLOT;</b> <b>PLOT y1*x1 ...;</b>	gibt X-Y-Plots aus.	X-Y- Plots

## Übungen

5. Erstellen Sie im Programm **SAMPLE19** mit der Prozedur **GCHART** ein Diagramm für die Variable **qual**. Experimentieren Sie mit den Optionen **MIDPOINTS** und **LEVELS** und den Argumenten **BLOCK**, **HBAR**, **VBAR**, **PIE** und **STAR**.
6. Verwenden Sie nun im Programm **SAMPLE20** die **SUBGROUP** Option in der Form **SUBGROUP=sex**, um innerhalb einer Säule nach Männern und Frauen zu differenzieren. Verwenden Sie alternativ die **GROUP** Option in der Form **GROUP=sex** und vergleichen Sie die Ergebnisse. (Hinweis: ... **VBAR qual /subgroup=sex;**)

## 9. Durchführen elementarer statistischer Verfahren

In diesem Kapitel werden einige grundlegende Verfahren der mathematischen Statistik behandelt.

### 9.1. PROC UNIVARIATE - Exploratives Analysieren von Daten

Die Prozedur **UNIVARIATE** berechnet neben den Kenngrößen oder Maßzahlen, die auch von der Prozedur **MEANS** ermittelt werden, weiterhin Momente, Extremwerte, Quantile, Anzahl der fehlenden Werte, sowie univariate Test-Statistiken und die zugehörigen Irrtumswahrscheinlichkeiten.

Sie können u.a. die folgenden Maßzahlen (Lagemaße, Kenngrößen) berechnen:

Option	Bedeutung
<b>NOBS</b>	Anzahl Beobachtungen
<b>N</b>	Anzahl gültiger Werte
<b>NMISS</b>	Anzahl fehlender Werte (NOBS - N)
<b>MIN</b>	kleinster Wert
<b>MAX</b>	größter Wert
<b>MEAN</b>	arithmetischer Mittelwert $\bar{x}$
<b>STD</b>	empirische Standardabweichung
<b>SUM</b>	Summe
<b>RANGE</b>	Spannweite (MAX-MIN)

Die Prozedur **UNIVARIATE** bietet darüberhinaus mit der Option **PLOT** die Möglichkeit, Daten mit **Stem-and-Leaf-Plots** und **Box-and-Whisker-Plots** zu visualisieren, und mit der Option **NORMAL** die Möglichkeit, die empirische Verteilungsfunktion mit einer Normalverteilung zu vergleichen.

Das Programm **SAMPLE21** bearbeitet die SAS Datei `tab11` mit der Prozedur **UNIVARIATE**:

```

/* SAMPLE21 */
DATA tab11;
    INPUT gewicht @@;
CARDS;
65 73 65 80 81 75 54 76 99 102 55 45 67 88
;
PROC UNIVARIATE DATA=tab11 PLOT NORMAL; /* Maßzahlen ? */

```

Das Programm **SAMPLE21** erzeugt u.a. folgende Ausgaben für die Variable `gewicht`:

```

Variable=GEWICHT

           Moments
           N           14  Sum Wgts           14

```

Mean	73.21429	Sum	1025
Std Dev	16.40909	Variance	269.2582
Skewness	0.164458	Kurtosis	-0.34173
USS	78545	CSS	3500.357
CV	22.41242	Std Mean	4.385514
T:Mean=0	16.69457	Prob> T	0.0001
Sgn Rank	52.5	Prob> S	0.0001
Num ^= 0	14		
W:Normal	0.975841	Prob<W	0.9087

Ferner erzeugt das Programm SAMPLE21 u.a. folgende Plots (Stem-and-Leaf-Plot und Box-and-Whisker-Plot)<sup>6</sup>:

Variable=GEWICHT			
Stem	Leaf	#	Boxplot
10	2	1	<pre>   +-----+(75 %) *---+---*(50 %) +-----+(25 %)   </pre>
9	9	1	
8	018	3	
7	356	3	
6	557	3	
5	45	2	
4	5	1	
-----+			
Multiply Stem.Leaf by 10**+1			

Beachten Sie, daß

- Sie im **Stem-and-Leaf-Plot** den ersten beobachteten Datenwert **65** z.B. mit Stamm **6**\*10 und Blatt **5** angezeigt bekommen, entsprechend **102** mit Stamm **10**\*10 und Blatt **2**.
- Sie im **Box-and-Whisker-Plot** den Median und 25%- und 75%-Quantile ablesen können; im Beispiel liegen zwischen 60 und 90 50% aller Beobachtungen.

Die Syntax für die Prozedur **UNIVARIATE** lautet:

```

PROC UNIVARIATE [DATA=SAS_file]
  [NORMAL] [PLOT];
VAR var1 ...;
BY varx ...;

```

## 9.2. PROC MEANS - Berechnen eines Vertrauensbereiches

Die Prozedur **MEANS** kann zur Berechnung von **Vertrauensbereichen** oder **Konfidenzintervallen** verwendet werden, die den tatsächlichen Erwartungswert mit einer vorgegebenen Sicherheit enthalten.

Im folgenden Beispielprogramm SAMPLE22 mit der SAS Datei tab12 soll für eine (fiktive) Stichprobe von n=48 Beobachtungen von Brötchengewichten gewicht [in Gramm] ein 95%-Konfidenzintervall ( $\alpha=5\%$ ) für den unbekanntem Erwartungswert  $\mu$  berechnet werden:

```

/* SAMPLE22 */

```

<sup>6</sup> Leider nicht in hochauflösender Graphik!

```

DATA tab12;      /* 48 Beobachtungen */
INPUT gewicht @@;
CARDS;
50 55 60 34 57 39 80 76 55 37 76 63
51 45 64 32 47 49 30 56 65 39 72 73
50 65 70 44 58 39 80 76 55 37 46 63
50 55 80 34 57 69 30 46 55 37 76 63
;
PROC MEANS DATA=tab12;
DATA tab13;
      t=tinv(0.975,47);          /* 0.975 t-Quantil, df=(n-1)=47 */
PROC PRINT DATA=tab13;

```

Das Programm SAMPLE22 liefert alle notwendigen Werte zur Berechnung eines 95%-Konfidenzintervalls:

N Obs	N	Mean	Std Dev
48	48	55.0000000	14.7056220
	OBS	T	
	1	2.01174	

N = 48  
 T = 2.01  
 STD = 14.70  
 MEAN = 55

Das Konfidenzintervall  $CI=[a,b]$  für den unbekanntem Erwartungswert  $\mu$  besitzt folgende Endpunkte:

$$CI = [MEAN - T*STD/\sqrt{N}, MEAN + T*STD/\sqrt{N}]$$

Das derartig berechnete Konfidenzintervall (*Confidence Interval*) CI enthält den unbekanntem Parameter  $\mu$  mit einer Sicherheit von 95%.

Sie können natürlich bei Ihrer Durchführung des Zufallsexperimentes (Auswahl einer Stichprobe) zu den 5 von 100 Fällen gehören, in der das Verfahren ein Konfidenzintervall liefert, das den wahren Parameter  $\mu$  **nicht** enthält. Bei einer Schätzung aufgrund einer Stichprobe, also zufälliger Auswahl von Beobachtungen, bleibt ein Risiko, das Sie nur mit einer Gesamterhebung (Stichprobe = Grundgesamtheit) ausschalten können. Bezogen auf das Beispiel müßten Sie **alle** hergestellten Brötchen wiegen, um den Erwartungswert (der Brötchenproduktion eines Tages) exakt, d.h. mit Irrtumswahrscheinlichkeit 0, zu bestimmen.

### 9.3. Statistischer Hintergrund

Sei  $t(x;n-1)$  die Verteilungsfunktion der Student'schen t-Verteilung mit  $(n-1)$  Freiheitsgraden und  $\bar{X}$  der arithmetische Mittelwert der Beobachtungen. Dann gilt in Analogie zum zentralen Grenzwertsatz:

$$(1) \quad \frac{\sqrt{n}}{s} (\bar{X} - \mu) \sim t(x;n-1)$$

$$(1a) \quad P(-z \leq \frac{\sqrt{n}}{s} (\bar{X} - \mu) \leq z) = t(z,n-1) - t(-z;n-1)$$

Die Formel (1a) läßt sich nach dem Erwartungswert  $\mu$  umstellen:

$$(2) \quad P( \bar{X} - zs/\sqrt{n} \leq \mu \leq \bar{X} + zs/\sqrt{n} ) = t(z;n-1) - t(-z;n-1)$$

Sei im folgenden die Irrtumswahrscheinlichkeit  $\alpha$  für das Verfahren folgendermaßen (von Ihnen kraft eigener Willkür oder bestimmter Vorgaben) festgelegt:

$$(3) \quad \alpha = 0.05$$

Damit der Erwartungswert  $\mu$  mit einer Wahrscheinlichkeit von  $(1-\alpha)$  im Konfidenzintervall liegt, muß die rechte Seite von (2) den Wert  $(1-\alpha)$  ergeben:

$$(4) \quad t(z;n-1) - t(-z;n-1) = 1 - \alpha = 0.95$$

Nach einigen Umformungen läßt sich der Wert von  $z$  bestimmen als  $z=t^{-1}(1-\alpha/2;n-1)$ :

$$(5) \quad \begin{aligned} t(z;n-1) - t(-z;n-1) &= t(z;n-1) - (1 - t(z;n-1)) \\ 2 t(z;n-1) - 1 &= 1 - \alpha \\ t(z;n-1) &= 1 - \alpha/2 \\ z &= t^{-1}(1-\alpha/2;n-1) \end{aligned} \quad (1-\alpha/2)\text{-Quantil der } t\text{-Verteilung mit } (n-1) \text{ df}$$

Beachten Sie, daß Sie das Signifikanz-Niveau  $\alpha$  festlegen. Sie "erkaufen" sich eine **größere Sicherheit** Ihrer Schätzung (kleines  $\alpha$ ) durch ein **längeres Konfidenzintervall** und umgekehrt erhalten Sie bei kleinerer Sicherheit ein kürzeres Konfidenzintervall. Sie können das Intervall natürlich auch verkleinern, indem Sie die Stichprobe vergrößern ( $\sqrt{n}$  im Nenner). Die Ableitung eines Konfidenzintervalls bei **bekanntem** Varianz  $\sigma^2$  erfolgt analog, wobei  $t(x;n-1)$  durch  $N(x;0,1)$  und  $s$  durch  $\sigma$  zu ersetzen ist. Grundlage für die Berechnung des Konfidenzintervalls ist in diesem Fall der zentrale Grenzwertsatz.

#### 9.4. PROC TTEST - Vergleichen der Erwartungswerte von zwei Gruppen

Die Prozedur **TTEST** dient zur Untersuchung der **Hypothese H**, ob für zwei Gruppen die Erwartungswerte  $\mu_1$  und  $\mu_2$  gleich sind. (Die **Alternative A** besagt entsprechend, daß die Erwartungswerte verschieden sind.)

Im folgenden Beispielprogramm **SAMPLE23** mit der SAS Datei **tab14** werden bei zwei Gruppen (Variable **sex**: Frauen und Männer) Körperfettanteile **fett** gemessen und miteinander verglichen:

```

/* SAMP23 */
DATA tab14;
    INPUT sex $ fett @@;
CARDS;
m 13.3 f 22   m 19.0 f 26 m 20 f 16 m 8   f 12 m 18 f 21.7
m 22   f 23.2 m 20   f 21 m 31 f 28 m 21 f 30 m 12 f 23
m 16   m 12   m 24
;
PROC TTEST DATA=tab14;
    CLASS sex;           /* Einteilung in zwei Gruppen */
    VAR fett;           /* untersuchte Variable */

```

Das Programm **SAMPLE23** liefert folgende Ausgabe:

```

Variable: FETT

```

SEX	N	Mean	Std Dev	Std Error	Minimum	Maximum
f	10	22.29	5.31965955	1.68222406	12.00000000	30.00000000
m	13	18.17	6.03243371	1.67309608	8.00000000	31.00000000

	T	DF	Prob> T
Unequal	1.7336	20.5	<b>0.0980</b>
Equal	1.7042	21.0	<b>0.1031</b>

For H0: Variances are equal, F' = 1.29    DF = (12,9)    Prob>F' = 0.7182

Beachten Sie, daß

- aufgrund der berechneten Werte für die Testgröße T (**PROB>|T|**) die Hypothese H nur mit der sehr hohen Irrtumswahrscheinlichkeit von 0.0980 (ca 10%) angenommen werden könnte. Die Hypothese sollte deshalb abgelehnt werden, d.h. der beobachtete Unterschied läßt sich nicht zufällig erklären, sondern ist statistisch signifikant.
- die Prozedur **TTEST** nur unter bestimmten Voraussetzungen sinnvolle Ergebnisse liefert (Unabhängigkeit, gleiche Varianzen, Normalverteilungsannahme).
- Sie einen nicht-parametrischen Test mit der Prozedur **NPARIWAY** durchführen können (**Wilcoxon** oder **Mann-Whitney U-Test**).

#### 9.5. Statistischer Hintergrund

Unter den oben genannten Voraussetzungen gilt:

$$(1) \quad \frac{\sqrt{n}}{\sigma} [(X - \mu_1) + (Y - \mu_2)] \sim N(0,2)$$

Durch Umstellen von (1) läßt sich in ähnlicher Form wie bei der Berechnung von Konfidenzintervallen ein Annahmebereich für die Hypothese  $\mu_1 = \mu_2$  berechnen.

### 9.6. PROC FREQ - Testen auf Unabhängigkeit (Chi-Quadrat Test)

Die Prozedur **FREQ** kann zum Testen auf Unabhängigkeit in einer Kreuztabelle verwendet werden. Die **Hypothese H** beim **Test auf Unabhängigkeit (Chi-Quadrat Test)** lautet, daß die Variablen unabhängig sind, die Alternative, daß Sie nicht unabhängig sind.

Im folgenden Beispielprogramm **SAMPLE24** mit der SAS Datei `tab15` werden folgende Beobachtungen der Variablen **X** und **Y** ausgewertet<sup>7</sup>:

	S (Schwarz)	W (Weiß)	Summe Zeile
T (Todesurteil)	17	19	36
Z (Zuchthaus)	149	141	190
Summe Spalte	166	160	326

```

/* SAMP24 */
DATA tab15;
    INPUT x $ y $ anzahl;          /* x: Strafe, y: Hautfarbe */
    CARDS;
T S 17
T W 19
Z S 149
Z W 160
;
PROC FREQ DATA=tab15;
    WEIGHT anzahl;
    TABLES x*y /CHISQ EXPECTED;

```

Das Programm **SAMPLE24** liefert folgende Ausgabe (gekürzt):

<sup>7</sup> Eine amerikanische Untersuchung analysierte den Effekt der Rassenzugehörigkeit **Y** (schwarz/weiß) auf das Urteil bei Mördern **X** (Todesstrafe/Zuchthaus).

TABLE OF X BY Y			
X [Strafe]	Y [Hautfarbe]		# nachträglich ergänzt
Frequency			
Expected	S	W	Total
T	17 18.331	19 17.669	36
Z	149 147.67	141 142.33	290
Total	166	160	326

STATISTICS FOR TABLE OF X BY Y			
Statistic	DF	Value	Prob
Chi-Square	1	0.221	0.638
Fisher's Exact Test (Left)			0.384
(Right)			0.741
(2-Tail)			0.725
Sample Size =	326		

Beachten Sie, daß

- die **tatsächlich beobachteten** Häufigkeiten sehr nahe bei den unter der Nullhypothese  $H$  (X und Y sind unabhängig!) **erwarteten** Häufigkeiten liegen.
- aufgrund des berechneten Wertes für den Chi-Quadrat-Tests die Hypothese  $H$  **nicht** abgelehnt werden sollte. Die Testgröße T liefert einen Wert, für den die zugehörige Irrtumswahrscheinlichkeit bei 0.638 liegt ( $P(T > 0.221) = 0.638$ ); d.h. die Irrtumswahrscheinlichkeit  $\alpha$ , die Hypothese abzulehnen, obwohl sie wahr ist, liegt bei 63% und nicht bei 1% oder 5%.

### 9.7. Statistischer Hintergrund

Es besteht Unabhängigkeit zwischen den Variablen X und Y, falls folgende Beziehungen erfüllt sind (hier wird der einfache Fall vorausgesetzt, daß X und Y jeweils nur 2 Werte annehmen können, X:  $x_1$  und  $x_2$ , Y:  $y_1$  und  $y_2$ ):

$$(1) \quad \begin{aligned} P(x_1, y_1) - P(x_1) P(y_1) &= 0, & P(x_2, y_1) - P(x_2) P(y_1) &= 0 \\ P(x_2, y_1) - P(x_2) P(y_1) &= 0 & P(x_2, y_2) - P(x_2) P(y_2) &= 0 \end{aligned}$$

Hierfür lassen sich leicht (durch Ersetzen von Wahrscheinlichkeiten mit Häufigkeiten) die analogen Beziehungen für die empirischen (beobachteten) Datenwerte aufstellen.

In der Prozedur **FREQ** werden die erwarteten Produkte (Option: **EXPECTED**) und die Testgröße T für einen Test auf Unabhängigkeit berechnet. T wird berechnet, indem die Quadrate der Abweichungen, geeignet normiert, summiert werden. Der beobachtete Wert der Testgröße T sollte "klein" sein, um die Hypothese der Unabhängigkeit zu rechtfertigen.

Die Testgröße  $T$  besitzt eine **Chi-Quadrat-Verteilung** mit  $(r-1)(c-1)$  Freiheitsgraden ( $r$ : Anzahl Zeilen,  $c$ : Anzahl Spalten). Die kritischen Werte der Chi-Quadrat-Verteilung sind im SAS System berechenbar (siehe Anhang).

### 9.8. PROC CORR - Testen auf Korrelation

Die Prozedur **CORR** berechnet Korrelationskoeffizienten numerischer Variablen. Die Korrelation zwischen 2 Variablen ist ein Maß für ihren linearen (!) Zusammenhang.

Im folgenden Beispielprogramm `SAMPLE25` mit der SAS Datei `tab16` wird der Zusammenhang zwischen den Variablen `jahr` und `anzahl` (Bevölkerung in Deutschland in Tausend) untersucht. Die Untersuchung beruht auf einer Stichprobe aus den Jahren 1950 bis 1985 in Abständen von 5 Jahren (Quelle: Statistisches Jahrbuch 1991):

```

/* SAMPLE25 */
DATA tab16;
    INPUT jahr anzahl @@;
CARDS;
50 68377 55 70326 60 72674 65 75647
70 77709 75 78697 80 78275 85 77619
;
PROC CORR COV PEARSON;
    VAR jahr anzahl;

```

Das Programm `SAMPLE25` liefert u.a. folgende Ausgabe:

```

OUTPUT
Command ==>
          CORRELATION ANALYSIS

Pearson Correlation Coefficients /
Prob > |R| under Ho: Rho=0 / N = 8

          JAHR              ANZAHL

          JAHR              1.00000          0.91605 /* emp. Korr. */
                   0.0              0.0014 /* alpha */

          ANZAHL            0.91605          1.00000
                   0.0014            0.0

```

Beachten Sie, daß

- aufgrund des berechneten Wertes für die Testgröße  $\hat{\rho}=0,91605$  ein signifikanter linearer Zusammenhang zwischen `jahr` und `anzahl` vermutet werden kann. Die Hypothese, daß eine Korrelation besteht, sollte deshalb angenommen und weiterhin untersucht werden (siehe **PROC REG**).

### 9.9. Statistischer Hintergrund

Die Korrelation zweier Variablen berechnet sich über Erwartungswert und Varianz der beteiligten Variablen.

$$(1) \quad \text{Cov}(X,Y) = E[XY] - E[X] E[Y]$$

$$(2) \quad \rho = \text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\sigma_{XY}}{\sqrt{(\sigma_{YY} \sigma_{XX})}}$$

Beim Test der Unkorreliertheit von zwei Variablen X und Y, d.h.  $\text{Cov}(X,Y)=0$ , auf Grundlage einer Stichprobe werden analog zum Test auf Unabhängigkeit die entsprechenden empirischen Werte verwendet, d.h. der empirische Korrelationskoeffizient  $\rho^{\wedge}$  wird berechnet:

$$(3) \quad \rho^{\wedge} = \frac{s_{XY}}{\sqrt{s_{YY} s_{XX}}}$$

Für Beobachtungspunkte, die "ungefähr" auf einer steigenden Geraden liegen, ergibt der empirische Korrelationskoeffizient  $\rho^{\wedge}$  einen Wert "ungefähr" bei 1 liegt, für solche auf einer fallenden Geraden einen Wert "ungefähr" bei (-1) und z.B. für stark streuende einen Wert bei 0.<sup>8</sup>

Die Verteilung von  $\rho^{\wedge}$  kann unter bestimmten Voraussetzungen approximativ bestimmt werden:

$$(4) \quad \frac{\sqrt{n-2} \rho^{\wedge}}{\sqrt{1-\rho^{\wedge 2}}} \sim t(n-2) \quad \text{Student t-Verteilung mit (n-2) Freiheitsgraden}$$

### 9.10. Zusammenfassung

In diesem Kapitel sind folgende statistische Verfahren behandelt worden:

Verfahren	SAS Prozedur	Verteilungsannahme
Test auf Unabhängigkeit	<b>PROC FREQ</b>	Chi-Quadrat-Verteilung
Test auf Korrelation	<b>PROC CORR</b>	t-Verteilung
Konfidenzintervall für den Erwartungswert	<b>PROC MEANS</b>	t-Verteilung
Lineare Regression und Varianzanalyse	<b>PROC REG</b>	F-Verteilung
t-Test für den Vergleich der Erfahrungswerte zweier Gruppen	<b>PROC TTEST</b>	Normalverteilung

<sup>8</sup> Es kann jedoch sehr wohl ein nicht-linearer Zusammenhang vorliegen, z.B. könnte approximativ gelten:  $Y=\sin(X)$

### 9.11. Übungen

1. Berechnen Sie im Programm SAMPLE26 für folgende (fiktive) Stichprobe ein Konfidenzintervall für den Erwartungswert zum Niveau  $\alpha=0.01$ . Benutzen Sie einen Taschenrechner oder eine Überschlagsrechnung, um die Formel auszuwerten. (Hinweis:  $t=tinv(1-\alpha/2, n-1)$ )

100 110 120 140 130 120 130 120 120 120 130 100 200 110 120 130

2. Verwenden Sie nun im Programm SAMPLE27 das Niveau  $\alpha=0.1$  und vergleichen Sie die Lage und Länge der Konfidenzintervalle zum Niveau 0.01 und 0.1.
3. Die Prozedur **REG** bietet die Möglichkeit, eine erweiterte Modellgleichung zu verwenden. Bei der Modellgleichung  $Y=a+bX+cX^2 + Z$  werden die Beobachtungswerte quadratisch (durch ein Polynom 2. Grades) approximiert. Berechnen Sie im Programm SAMPLE28 für die Bevölkerungsdaten im Dateneingabeschritt zusätzlich `jahr_2` als `jahr*jahr` und ändern Sie die Modellgleichung:

```
DATA ...;
  jahr_2=jahr*jahr;          /* jahr_2 wird spaeter benoetigt */
PROC REG ...;
  MODEL anzahl=jahr jahr_2; /* neues Modell */
```

4. Beurteilen Sie anhand der `MODEL_F` und `R_SQUARE` Werte, welches Modell den Zusammenhang zwischen `anzahl` und `jahr` besser "erklärt". War dieses Ergebnis zu erwarten?
5. Erzeugen Sie nun im Programm SAMPLE29 einen Plot mit den Beobachtungswerten und der Regressions-Parabel. Wo schmiegt sich der Konfidenzgürtel enger an die Parabel und wie ist das zu erklären?

## 10. Durchführen fortgeschrittener statistischer Verfahren

### 10.1. PROC REG - Berechnen einer Regressionsgeraden

Die Prozedur **REG** dient zur genaueren Untersuchung eines linearen Zusammenhangs zwischen zwei oder mehreren Variablen.

Bei der linearen Regression wird die Hypothese aufgestellt, daß sich die Variablen Y und X in der Form  $Y=mX+b+Z$  darstellen lassen. Dabei sind **m** und **b** feste, aber unbekannte Parameter und **Z** ist ein "zufälliger Fehler", z.B. ein Meßfehler oder ein Störeffekt.

In der Terminologie der linearen Regression heißen:

Bezeichnung	Bedeutung
Y	abhängige oder erklärte Variable (beobachtete Zufallsvariable)
X	unabhängige oder erklärende Variable (beobachtete Zufallsvar.)
b	Schnittpunkt der Regressionsgeraden mit der horizontalen Achse

## 10. Durchführen fortgeschrittener statistischer Verfahren

m	Steigung der Regressiongeraden oder Koeffizient von X
Z	Residuum oder zufälliger Fehler (nicht gemessen, sondern postuliert)
$Y=mX+b+Z$	Modellgleichung oder lineares Modell
$y=mx+b$	Gleichung der Regressionsgeraden

Im folgenden Beispielprogramm SAMPLE30 mit der SAS Datei tab17 wird der lineare Zusammenhang zwischen den Variablen jahr und anzahl (Bevölkerung in Deutschland in Tausend) mittels einer linearen Regression untersucht:

```

/* SAMPLE30 */
DATA tab17;
    INPUT jahr anzahl @@;
CARDS;
50 68377 55 70326 60 72674 65 75647
70 77709 75 78697 80 78275 85 77619
;
PROC REG DATA=tab17;
    MODEL anzahl=jahr; /* anzahl erklärt durch jahr */

```

Das Programm SAMPLE30 liefert u.a. folgende Ausgabe:

```

OUTPUT
Command ===>
                Parameter Estimates
Variable  DF      Parameter      Standard      T for H0:
           Estimate      Error      Parameter=0  Prob>|T|
INTERCEP  1          54895    3629.6431651   15.124  0.0001
JAHR      1          296.595238  53.01433166    5.595  0.0014

```

```

b^ = INTERCEP = 54895          y-Abschnitt
m^ = YEAR      = 296.59        Steigung

```

Das Beispielprogramm SAMPLE31 mit der SAS Datei tab18 soll nun Plots ausgeben, in denen die Beobachtungspunkte und die durch sie gelegte Regressionsgerade dargestellt werden:

```

/* SAMPLE31 */
DATA tab18;
...;
PROC REG DATA=tab18;
    MODEL anzahl=jahr;
    OUTPUT out=tab19
           p=erw u95=oben l95=unten;
SYMBOL1 INTERPOL=spline;
PROC GPLOT DATA=tab19;
    PLOT erw*jahr anzahl*jahr /OVERLAY;

```

In den Variablen oben und unten der SAS Datei tab19 hat das SAS System 95%-Konfidenzintervalle für jeden Punkt berechnet, die Sie in einem weiteren Plot als "Konfidenzgürtel" darstellen können:

```

SYMBOL2 INTERPOL=spline;
SYMBOL3 INTERPOL=spline;

```

```
PROC GPLOT DATA=tab19;
    PLOT erw*jahr unten*jahr oben*jahr anzahl*jahr /OVERLAY;
```

## 10.2. Statistischer Hintergrund

Die Prozedur **REG** berechnet ausgehend von der Stichprobe  $S=((x_1, y_1), \dots, (x_n, y_n))$  **Schätzwerte**  $m^\wedge$  und  $b^\wedge$  für eine Gerade, die "möglichst dicht" bei den Beobachtungspunkten liegt. Die Minimaleigenschaft "möglichst dicht" wird dabei durch die Summe der Abstandsquadrate (*Sum of Squared Errors*, SSE) ausgedrückt:

$$(1) \quad \text{SSE} = [y_1 - (m^\wedge x_1 + b^\wedge)]^2 + [(y_2 - (m^\wedge x_2 + b^\wedge))]^2 + \dots + [y_n - (m^\wedge x_n + b^\wedge)]^2 \\ = \text{Minimum über alle Kombinationen von } (m, b)$$

Die berechnete Regressionsgerade  $y = m^\wedge x + b^\wedge$  hat **unter allen Geraden** den "kleinsten" Abstand zu den beobachteten Punkten. Der verbleibende Abstand zur Geraden (SSE) drückt das Verhalten von Y aus, daß sich nicht durch das Modell  $Y = m^\wedge X + b^\wedge$  erklären läßt, sondern vom Fehler Z abhängt.

Die Schätzwerte  $b^\wedge$  (Achsenabschnitt, *intercept*) und  $m^\wedge$  (Koeffizient vor der unabhängigen Variablen) werden folgendermaßen berechnet:

$$(2) \quad m^\wedge = \frac{s_{XY}}{s_{XX}}$$

$$(3) \quad b^\wedge = \bar{y} - \bar{x} \cdot \frac{s_{XY}}{s_{XX}}$$

Im unmittelbaren Zusammenhang mit der linearen Regression steht die **Varianzanalyse des linearen Modells**, mit der die Güte der Modellgleichung  $Y = mX + b$  überprüft wird. Die empirische Varianz  $s_{YY}$  der abhängigen Variablen Y läßt sich in zwei Summanden, nämlich die empirische Varianz des Fehlers (SSE) und die empirische Varianz des Modells (SSM), aufspalten:

$$(4) \quad s_{YY} = \text{SSE} + \text{SSM}$$

Die **Varianz des Fehlers (SSE)** beschreibt die Abweichung der Beobachtungspunkte von der Regressionsgeraden:

$$(5) \quad \text{SSE} = [y_1 - (m^\wedge x_1 + b^\wedge)]^2 + [(y_2 - (m^\wedge x_2 + b^\wedge))]^2 + \dots + [y_n - (m^\wedge x_n + b^\wedge)]^2$$

Die **Varianz des Modells (SSM)** beschreibt die Abweichung des Mittelwertes  $\bar{y}$  von der Regressionsgeraden:

$$(6) \quad \text{SSM} = [\bar{y} - (m^\wedge x_1 + b^\wedge)]^2 + [\bar{y} - (m^\wedge x_2 + b^\wedge)]^2 + \dots + [\bar{y} - (m^\wedge x_n + b^\wedge)]^2$$

Der gewichtete Quotient **F** aus SSM und SSE ist ein Maß, wie groß die Varianz des Modells im Vergleich zur Varianz des Fehlers ist; d.h. wie "gut" das Modell die Varianz der abhängigen Variablen erklärt:

$$(7) \quad F = \frac{SSM/1}{SSE/(n-2)} \sim F(x;1,n-2) \quad \text{Fisher-Verteilung mit } (1,n-2) \text{ Freiheitsgraden}$$

Eine ähnliche Größe  $R^2$  beschreibt den gewichteten Quotienten aus SSM und  $s_{YY}$ :

$$(8) \quad R^2 = \frac{SSM/1}{s_{YY}/(n-1)}$$

Für "großes"  $F$  oder  $R^2$  "nahe bei 1" erklärt das lineare Modell  $Y=mX+b$  einen Großteil der gesamten empirischen Varianz von  $Y$ , während der Fehler  $Z$  nur unwesentlich beiträgt. Im anderen Fall erklärt das lineare Modell den Zusammenhang zwischen  $Y$  und  $X$  nicht ausreichend. In diesem Fall sollte ein anderes Modell verwendet werden, z.B. durch Hinzunahme weiterer erklärender Variablen oder durch Verwenden eines anderen funktionalen Zusammenhangs wie z.B die Modellgleichung  $Y=aX^2+bX+Z$ .

## 11. Bearbeiten der SAS Datei

In diesem Kapitel werden weitere Möglichkeiten der Eingabe und Bearbeitung von Daten erläutert.

### 11.1. DATA - Erzeugen neuer (abgeleiteter) Variablen

Sie können in einem Dateneingabeschritt (*DATA step*) neue Variablen aus bekannten Variablen erzeugen oder Variablen mit einer Zuweisung neu definieren.

Im Beispielprogramm SAMPLE32 mit der SAS Datei tab20 werden z.B. die Variable `jahr` um 1900 vergrößert und die Variablen `kaltmiet`, `nebenk` und `heizk` zu einer neuen Variablen `miete` addiert:

```
/* SAMPLE32 */
DATA tab20;
    INPUT jahr kaltmiet nebenk heizk @@;
    jahr = jahr + 1900;
    miete = kaltmiet + nebenk + heizk;
CARDS;
90 800 100 100 91 900 200 150 92 1000 200 150
PROC PRINT;
```

Das Programm SAMPLE32 liefert folgende Ausgabe:

```
OUTPUT
Command ==>

      OBS      JAHR  KALTMIENT      NEBENK      HEIZK      MIETE
      ---  ---
      1      1990      800      100      100      1000
      2      1991      900      200      150      1250
      3      1992     1000      200      150      1350
```

Beachten Sie, daß

- das SAS System eine neue Variable `miete` erzeugt hat und die Variable `jahr` mit neuen Datenwerten berechnet hat.
- die Erzeugung und Zuweisung für alle Beobachtungen gleichermaßen durchgeführt werden.

Sie können in Zuweisungen u.a. folgende **Bausteine** verwenden:

Baustein	Typen	Beispiele
Variablen	Variablen in <b>INPUT</b> Liste bzw. vorher definierte Variablen	flaeche=a*b preis=flaeche*3.5
Konstanten	Zahlen Zeichenketten	status='OK'
arithmetische Operatoren	Addition: + Subtraktion: - Multiplikation: * Division: / Potenzierung: **	jahr=jahr+1900 diff=a-b mult=a*b div=a/b y=x**2+4
SAS Funktionen	Wurzel: sqrt(x) Exp.funktion: exp(x)	y=sqrt(x)
logische und Vergleichsoperatoren	Gleichheit: = Ungleichheit: ^= Kleiner: < Kleiner oder gleich: <= Größer: > Größer oder gleich: >= log. Und: AND log. Oder: OR log. Verneinung: NOT	<b>IF</b> (a=b) <b>THEN</b> c=...;  <b>IF</b> (a=b AND c=d) <b>THEN</b> ...;
sonstige Operatoren	Verkettung:    Minimum: <> Maximum: ><	name=vorname    nachname MINAB=a<>b MAXAB=a><b

## 11.2. DATA - Bedingtes Einlesen von Beobachtungen

Sie können das Einlesen einer Beobachtung in eine SAS Datei davon abhängig machen, ob die Beobachtung eine bestimmte Bedingung erfüllt.

Im folgenden Beispielprogramm SAMPLE33 mit der SAS Datei tab21 werden nur Beobachtungen berücksichtigt, deren Datenwerte für die Variable `alter` zwischen 10 und 30 liegen:

```

/* SAMPLE33 */
DATA tab21;
    INPUT alter @@;
    IF (alter < 10 OR alter > 30) THEN DELETE;
CARDS;
5 10 15 20 25 30 35 40
;
PROC PRINT;

```

Das Programm SAMPLE33 liefert folgende Ausgabe:

```

OUTPUT
Command ==>

      OBS      ALTER
      1         10
      2         15
      3         20
      4         25
      5         30

```

Sie können in ähnlicher Weise die Zuweisung eines Datenwertes an eine Variable über eine Bedingung steuern, die von anderen Datenwerten der Beobachtung abhängt.

Im folgenden Beispielprogramm SAMPLE34 mit der SAS Datei tab22 wird die Variable `status` über die Datenwerte der Variablen `alter` definiert. Die Variable `status` repräsentiert eine Gruppeneinteilung:

```

/* SAMPLE34 */
DATA tab22;
  INPUT alter @@;
  IF (alter < 10 )                THEN status='< 10 ' ;
  IF (alter >= 10 AND alter < 20) THEN status='10-19' ;
  IF (alter >= 29 AND alter < 30) THEN status='20-39' ;
  IF (alter > 30)                 THEN status='> 30 ' ;
CARDS;
5 10 15 20 25 30 35 40
;
PROC PRINT;

```

Das Programm SAMPLE34 liefert folgende Ausgabe (gekürzt):

```

OUTPUT
Command ==>

      OBS      ALTER      STATUS
      1         5         < 10
      2        10        10-19
      ...
      7        35         > 30
      8        40         > 30

```

### 11.3. DATA - Definieren von genaueren Bezeichnungen für Variablen

Im SAS System sind Namen von Variablen auf nur 8 Zeichen Länge begrenzt. Sie können aber mit der Anweisung **LABEL** einen Namen durch eine aussagekräftigere Bezeichnung versehen (maximal 40 Zeichen), die z.B. in der Prozedur **PRINT** ausgegeben wird.

Im folgenden Beispielprogramm SAMPLE35 mit der SAS Datei tab23 wird eine genauere Bezeichnung für die Variable `alter` definiert und ausgegeben:

```

/* SAMPLE35 */
DATA tab23;
    INPUT alter @@;
    LABEL alter 'Alter der Schueler';
CARDS;
5 6 6 7 5 6 6 5 7 7 5 6 7 6 5 6 6 7 5 6 7
;
PROC PRINT DATA=tab23 LABEL;

```

Beachten Sie, daß

- nun in der **PRINT** Anweisung zusätzlich die Anweisung **LABEL** notwendig ist.

#### 11.4. DATA - Definieren von genaueren Bezeichnungen für Datenwerte

Sie können mit der Prozedur **FORMAT** auf ähnliche Weise Datenwerten eine aussagekräftigere Bezeichnung zuweisen.

Im folgenden Beispielprogramm **SAMPLE36** mit der SAS Datei `tab24` wird in der Prozedur **FORMAT** unter der Formatbezeichnung `frageb` eine Zuordnung zwischen Zahlen und zugehörigem Text durchgeführt. Zur Unterscheidung von Variablen wird eine Formatbezeichnung in einer **FORMAT** Anweisung mit einem Punkt beendet. Die Datenwerte der Variablen `antwort` werden in der Prozedur **PRINT** unter Kontrolle der Formatbezeichnung `frageb` ausgegeben:

```

/* SAMPLE36 */
PROC FORMAT;           /* Definition von frageb */
    VALUE frageb       0='Keine Meinung'
                      1='Nicht zutreffend'
                      2='Zutreffend'
                      3='Sehr zutreffend';

DATA tab24;
    INPUT antwort @@;
    FORMAT antwort frageb.;           /* Formatierung durch frage. */
CARDS;
0 1 2 3
;
PROC PRINT DATA=tab24;

```

Das Programm **SAMPLE36** erzeugt folgende Ausgabe:

```

OUTPUT
Command ==>>

```

OBS	ANTWORT
1	keine Meinung
2	Nicht zutreffend
3	Zutreffend
4	Sehr zutreffend

### 11.5. DATA - Einlesen von Datums- und Zeitformaten

Sie können in einem Dateneingabeschritt auch Datenwerte einlesen, die in einem Datumsformat wie z.B. TTMMJJ vorliegen (30DEC94). Das folgende Beispielprogramm SAMPLE37 mit der SAS Datei tab25 bearbeitet Geburtsdaten und verwendet das bereits im SAS System definierte Format DATE7, das allerdings englische Abkürzungen voraussetzt:

```

/* SAMPLE37 */
DATA tab26;
    INPUT datum1 date7; /* TTMMJJ, englische Bez. */
    datum2=today(); /* aktuelles Datum */
    diff_t=datum2-datum1 /* Differenz in Tagen */
    diff_w=intchk('week', datum1, datum2);
                                /* Differenz in Wochen */
    diff_m=intchk('month', datum1, datum2);
                                /* Differenz in Monaten */
    diff_j=intchk('year', datum1, datum2);
                                /* Differenz in Jahren */

CARDS;
01jan60
14may75
17dec80
;
PROC PRINT;
    FORMAT datum1 datum2 date7.;
    VAR datum1 datum2 diff diff_w diff_m diff_j;

```

Das Programm SAMPLE37 erzeugt folgende Ausgabe:

OBS	DATUM1	DATUM2	DIFF_T	DIFF_W	DIFF_M	DIFF_J
1	01JAN60	30MAY94	12568	1796	412	34
2	14MAY75	30MAY94	6956	994	228	19
3	17DEC80	30MAY94	4912	702	161	14

### 11.6. DATA - Einlesen aus permanenten SAS Dateien

Die in einem Dateneingabeschritt erzeugte SAS Datei wird nach Beendigung der SAS Sitzung automatisch gelöscht, so daß sie immer wieder ( ggfs. zeitaufwendig) neu erzeugt werden muß.

Sie können eine SAS Datei permanent in einer externen SAS Datei abspeichern, indem Sie einen zweistufigen Namen der Form *Verzeichnis.Dateiname* verwenden.

Im folgenden Beispielprogramm SAMPLE38 mit der SAS Datei tab27 wird eine externe SAS Datei C:\SAS\DATASETS\TAB27.SSD (Dateiendung SSD unter DOS) erzeugt:

```

/* SAMPLE38 */
LIBNAME ds 'C:\SAS\DATASETS';
DATA ds.tab27; /* erzeugt permanente SAS Datei */
    INPUT alter @@;
CARDS;
5 6 6 7 5 6 6 5 7 7 5 6 7 6 5 6 6 7 5 6 7
;

```

Sie können nun in folgenden Sitzungen auf diese permanent gespeichert SAS Datei mit einer **SET** Anweisung zugreifen<sup>9</sup>:

```
/* SAMPLE38 */
LIBNAME ds 'C:\SAS\DATASETS'
DATA tab1;                               /* lädt permanente SAS Datei */
    SET ds.tab27;

PROC PRINT DATA=tab1;
```

Beachten Sie, daß

- Sie das aktuelle Verzeichnis mit der Abkürzung '.' angeben können, z.B.:  
LIBNAME datasets '.';

### 11.7. Zusammenfassung

In diesem Kapitel sind folgende SAS Prozeduren und Anweisungen behandelt worden:

Prozedur/Anweisung	Aufgabe	Bemerkung
<b>DATA</b> <b>IF ... THEN</b> <b>DELETE;</b> <b>IF ... THEN ...;</b>	liest Beobachtungen nur ein, wenn bestimmte Bedingungen erfüllt sind.	siehe Bausteine
<b>DATA</b> <b>SET lib1.SAS_file;</b>	verwendet eine permanente SAS Datei.	Externe Dateien werden über zweistufige Namen bezeichnet (siehe <b>LIBNAME</b> ).
<b>DATA</b> <b>LABEL var1='...' ...;</b>	definiert genauere Bezeichnungen für Variablenamen.	
<b>LIBNAME lib1 'lib2';</b>	verknüpft einen SAS Namen für ein Verzeichnis <i>lib1</i> mit einem externen Namen <i>lib2</i> .	Externe Verzeichnisnamen <i>lib2</i> sind betriebssystemabhängig.
<b>PROC FORMAT;</b> <b>VALUE</b> <i>fb value1='...';</i>	definiert eine Formatbeschreibung <i>fb</i> .	
<b>PROC PRINT;</b> <b>FORMAT var1 fb1.</b> <i>var2 fb2. ...;</i>	verwendet für die Ausgabe einer Variablen <i>var1</i> eine Formatbeschreibung <i>fb1</i> .	Zur Unterscheidung zwischen Variablenamen und Formaten ist ein Punkt '.' am Ende des Formates notwendig.

<sup>9</sup> Die SET Anweisung wird auch in einem anderen Zusammenhang benötigt, nämlich beim Zusammenfügen mehrerer SAS Dateien (Merge).

<b>Zuweisung:</b> <i>a=expression;</i>	definiert eine neue Variable <i>a</i> oder verändert eine eingelesene Variable.	siehe Bausteine
---	---	-----------------

Bislang sind folgende Typen von Dateien behandelt worden:

Typ	Endung	Bedeutung
Plot- oder Graphikdatei	.TIF, .PS, CGM, EPS, ...	enthält Graphikinformationen, kann in andere Programme importiert werden bzw. zum Drucker/Plotter geschickt werden.
Rohdatendatei	.DAT (zum Beispiel)	enthält Rohdaten, die eingelesen werden sollen.
SAS Datei	.SSD (DOS)	enthält eine SAS Datei (temporär oder permanent).
SAS Programm	.SAS	enthält ein SAS Programm.

## 11.8. Übungen

- Bei einer statistischen Erhebung werden Datenwerte für Unternehmen, Umsatz und Gewinn erhoben, wobei die numerischen Datenwerte in DM gemessen werden. Stellen Sie im SAS Programm SAMPLE39 die Datenwerte sowohl in DM als auch in Dollar dar (zur Vereinfachung gelte die Umrechnung 1:1.50).

Unternehmen	Umsatz	Gewinn
A	1.000.000	300.000
B	2.000.000	200.000
C	3.000.000	100.000

- Berechnen Sie nun im SAS Programm SAMPLE40 für die Beobachtungen zusätzlich den Quotienten aus Gewinn und Umsatz.
- Bei einer Erhebung sollen nur Personen berücksichtigt werden, deren Alter zwischen 20 und 39 liegt **und** deren Einkommen größer als 4000 ist. Schreiben Sie im Programm SAMPLE41 einen entsprechenden Dateneingabeschritt für folgende Beobachtungen von `alter` und `einkomm`:

19 4500 20 3900 39 4000 20 3999 20 4001 19.5 5000

Hinweis: **if not(a and b) then ...** entspricht **if (not a) or (not b) then ...**

- Teilen Sie die folgenden Beobachtungen im Programm SAMPLE42 in Klassen ein:

0, 1, 5, 10, 20, 22, 17

Verwenden Sie hierzu eine neue Variable `klasse`, die folgendermaßen definiert wird:  
Für  $1 \leq x \leq 10$  wird `klasse=1` gesetzt, für  $11 \leq x \leq 20$  auf 2, sonst auf 3.

5. Definieren Sie über die Prozedur **FORMAT** eine Formatbezeichnung für die Variable `klasse`, wobei 1 durch '1<= x <= 10' ersetzt wird usw.
6. Versuchen Sie, eine Eingabemaske für die SAS Datei aus Übung 1 über das **SAS/FSP Modul** zu erstellen (Hinweis: **PROC FSEDIT NEW=tabxy;**).
7. Lesen Sie die folgenden Geburtstage ein und berechnen Sie die Differenz zwischen dem ältesten und dem jüngsten Menschen in Tagen bzw. in Jahren:  
  
14. Januar 1934, 23. April 1988, 27. Juni 1980, 15. Juni 1981, 14. Juni 1981
8. Verwenden Sie die Prozedur **FSBROWSE** zum Betrachten einer SAS Datei (Hinweis: **PROC FSBROWSE DATA=...**;) )

## 12. Anzeigen und Ändern von Voreinstellungen

In diesem Kapitel werden einige Prozeduren zum Anzeigen und Ändern von systemspezifischen Einstellungen behandelt. Ferner wird das Abspeichern von Graphiken erläutert.

### 12.1. TITLE, FOOTNOTE - Hinzufügen von Titel- und Fußzeilen

Die SAS Anweisungen **TITLE** und **FOOTNOTE** versehen jede Seite im OUTPUT Fenster mit Titelzeilen und Fußnoten.

Im folgenden Beispielprogramm **SAMPLE43** mit der SAS Datei `tab28` werden 2 Titelzeilen in die Ausgabe eingefügt:

```
/* SAMPLE43 */
DATA tab28;
    INPUT alter gewicht @@;
CARDS;
20 80 23 65 18 60 21 56 . 60 19 .
;
TITLE 'Untersuchung von Alter und Gewicht';
TITLE1 '(c) Universitaet Osnabrueck, Sept. 92';
PROC MEANS;
```

Die Syntax für die **TITLE** und **FOOTNOTE** Anweisungen lautet:

```
TITLE '...';           /* Titelzeile */
TITLEn '...';        /* n.te Titelzeile, z.B. TITLE2 '(c) Univ. Osnabrueck' */

FOOTNOTE;           /* Fußnote */
FOOTNOTEn '...'
```

### 12.2. (G)OPTIONS - Setzen von Systemoptionen

Die Anweisungen **OPTIONS** und **GOPTIONS** dienen zur benutzerspezifischen Einstellung von Parametern (*options*) wie z.B. Seitenlänge, Datumsangabe, Farbe und Symboltyp.

Sie können sich die voreingestellten Parameter zunächst mit folgenden SAS Anweisungen anzeigen lassen<sup>10</sup>:

```
PROC OPTIONS;
PROC GOPTIONS;
```

<sup>10</sup> Einige Optionen sind betriebssystemabhängig. Konsultieren Sie ggf. das systemspezifische Handbuch oder die Online Hilfe, um die verfügbaren Optionen und ihre Bedeutung herauszufinden.

Im folgenden Beispielprogramm SAMPLE44 mit der SAS Datei tab28 wird die Ausgabe der Prozedur **GPLOT** in eine Datei PLOT.PS (unter DOS) geleitet und zwar im Format PS (PostScript):

```

/* SAMPLE 44 */
DATA tab28;
...
FILENAME PLOTOUT 'PLOT1.EPS'; /* stellt Verbindung zwischen
                               SAS und DOS Dateinamen her. */
GOPTIONS    DEVICE=PSEPSF      /* Grafikformat: Encaps. Postscript */
            GSFMODE=REPLACE    /* Ueberschreibemodus */
            GSFNAME=PLOTOUT;   /* SAS Dateiname, siehe FILENAME */
PROC GPLOT DATA=tab28;
...;
FILENAME PLOTOUT 'PLOT2.EPS'; /* neuer Dateiname */
PROC GPLOT ...                 /* nächste Abbildung */

```

Die vom SAS Programm erzeugte Graphikdatei PLOT1.EPS können Sie z.B. mit dem Programm **Word für Windows** importieren.

Anstelle von PSEPSF können Sie z.B. auch folgende Graphikformate verwenden:

CGMMWWA	für CGM Format, geeignet speziell für <b>Word für Windows</b>
PS	für PostScript Format, kann direkt gedruckt werden
FX85	für Epson FX-85
HPDJ*	für HP DeskJet

Die Prozedur **GDEVICE** liefert eine Auflistung aller möglichen Formate:

```
PROC GDEVICE;
```

Die Syntax für die **OPTIONS** und **GOPTIONS** Anweisungen lautet:

```

OPTIONS option1=value1 option2=value2 ...;
GOPTIONS goption1=gvalue1 goption2=gvalue2 ...;

```

### 12.3. HELP - Anzeigen von Informationen

Sie können sich zu den genannten und weiteren Prozeduren und Anweisungen genauere Informationen direkt im Display Manager System anzeigen lassen. Hierzu existiert ein menügeführtes Hilfesystem.

Wechseln Sie mit dem Kommando **HELP** in das HELP Menu System, um sich z.B. Informationen über die Prozedur **MEANS** anzeigen zu lassen:

```

PROGRAM EDITOR
COMMAND ==> HELP
HELP
Select Option ==> _

SAS SYSTEM HELP      Scroll down for more selections

  Examples: Select Option ==> append
              (help on PROC APPEND)

Base procedures
APPEND                CORR                FREQ                STANDARD
CALENDAR              CPORT              MEANS              SUMMARY
...

```

Wählen Sie nun das Stichwort **MEANS** aus:

```

HELP
Select Option ==> MEANS

```

```

HELP
Command ==>

PROC MEANS produces simple univariate
descriptive statistics for numeric variables.

PROC MEANS options;
  VAR variables;
...

```

Beachten Sie, daß

- Sie mit den DMS Kommandos **TOP**, **FORWARD**, **BACKWARD** und **BOTTOM** (oder den entsprechenden Funktionstasten) im angezeigten Text blättern können.
- Sie mit dem DMS Kommando **END** jeweils in das vorhergehende Menü gelangen.

#### 12.4. HELP - Arbeiten mit SAS Beispielprogrammen

Sie können als weitere Informationsquelle die mit dem SAS System ausgelieferten **Beispielprogramme** auflisten, testen und an Ihre Aufgabenstellung anpassen<sup>11</sup>.

#### 12.5. Zusammenfassung

In diesem Kapitel sind folgende SAS Prozeduren behandelt worden:

Prozedur	Aufgabe	OUTPUT Fenster
<b>PROC (G)OPTIONS;</b>	zeigt Systemparameter an.	Option=Wert
<b>PROC GDEVICE;</b>	listet Graphikformate auf.	Liste der Formate

<sup>11</sup> Erkundigen Sie sich bei einem lokalen Experten, wo die Beispielsprogramme installiert worden sind.

In diesem Kapitel sind ferner folgende SAS Anweisungen behandelt worden:

Anweisung	Aufgabe	Bemerkungen
<b>FOOTNOTE(n) '...';</b>	erzeugt Fußnoten.	
<b>GOPTIONS</b> <i>keyw=value ...;</i>	setzt Optionen für Graphikausgabe.	Graphikoptionen sind systemabhängig.
<b>OPTIONS</b> <i>keyw=value ...;</i>	setzt Optionen.	Optionen sind systemabhängig.
<b>TITLE(n) '...';</b>	erzeugt Titelzeilen.	
<b>FILENAME</b> <i>fn1 'fn2';</i>	verknüpft <i>fn1</i> mit externen Dateinamen <i>fn2</i> .	Externe Dateinamen sind systemabhängig.

## 12.6. Übungen

1. Lassen Sie sich die aktuellen Optionen anzeigen (**PROC OPTIONS**) und verändern Sie z.B. die Seitenlänge (**PAGELength**).
2. Fügen Sie Ihren bisherigen Programmen aussagekräftige Titelzeilen und Fußnoten hinzu (**TITLE** und **FOOTNOTE**).
3. Experimentieren Sie mit Beispielprogrammen für die bisher behandelten Prozeduren.
4. Listen Sie die verfügbaren Graphikformate auf und suchen Sie z.B. alle Formate für den HP DeskJet (**PROC GDEVICE**, Kommando: **FIND HP**).
5. Erzeugen Sie eine Graphik im PostScript Format (**PS**) und drucken Sie die Graphik auf einem PostScript Drucker aus.
6. Erzeugen Sie eine Graphik im **CGMMWWA** Format und binden Sie die Graphik in ein **Word für Windows** Dokument ein.
7. Rufen Sie den SAS Graphik-Editor auf und bearbeiten Sie eine Graphik (z.B. unter SAS für Windows möglich).

## 13.Hinweise zu den Übungen

### Allgemeine Hinweise:

- Speichern Sie Ihre Daten, Programme, Ausgaben und Graphiken von Zeit zu Zeit ab (z.B. auf eine Diskette).
- Trainieren Sie die Handhabung des DMS, bevor Sie umfangreiche SAS Programme schreiben.
- Versuchen Sie, die Programmierübungen zunächst selbständig zu lösen.

### Hinweise zum Programmieren:

- **KISS** - Keep it simple stupid!
- Nutzen Sie "sprechende" Namen für Variablennamen und SAS Dateien (maximale Länge: 8 Zeichen, 1. Zeichen: Buchstabe, sonst Buchstaben, Ziffern und Unterstrich '\_').
- Verwenden Sie ggfs. zusätzlich Variablen- und/oder Werte-Label zur besseren Lesbarkeit der Ausgaben.
- Schließen Sie jede Anweisung mit einem Semikolon ';' ab.
- Fügen Sie Kommentare hinzu:  
`/* comment_text */`
- Fügen Sie Leerzeichen und Leerzeilen ein und rücken Sie strukturiert ein.
- Beenden Sie jeden Schritt mit einer **RUN** Anweisung.
- Überprüfen Sie sorgfältig die Warnungen und Fehlermeldungen des SAS Systems im **LOG** Fenster.
- Schreiben Sie Ihre Programme durchgängig in einer Sprache (deutsch oder englisch) und verwenden Sie dieselbe Sprache für Kommentare.
- Überprüfen Sie Ergebnisse kritisch. Verwenden Sie zum Testen kleine, überschaubare Datensätze oder solche, für die Sie bereits die erwarteten Ergebnisse kennen.

## **14. Anhang A - Literaturhinweise und Online Informationen**

Sie können sich weitere Informationen aus folgenden Quellen beschaffen:

- **SAS Handbücher (Original-Dokumentation)**
- **SAS Sekundärliteratur**
- **SAS Online Help System**
- **SAS Online Tutorial**
- **SAS Beispielprogramme**

Bitte wenden Sie sich an den für Sie zuständigen SAS Betreuer, wenn Sie die Dokumentation zum SAS System sichten oder ein spezielles Problem mit dem SAS System lösen wollen.

## 15.Anhang B - Quantile von wichtigen Verteilungen

Sie können im SAS System folgendermaßen eine Tabelle der Quantile ("kritischen Werte") für den 2-seitigen t-Test zum Niveau  $\alpha$  (im Programm: alpha) erstellen:

```
DATA;
  PUT 'Quantile der t-Verteilung';
  PUT 'Niveau (alpha)   Freih.grade (df)   krit. Wert (t)';
  DO alpha=0.01 TO 0.05 BY 0.01;
    DO df= 1 TO 10 BY 1,
          10 TO 100 BY 10,
          100 TO 1000 BY 100;
      t=tinv(1.0 - alpha/2.0,df);
      PUT a df t;
    END;
  END;
```

### Quantile der t-Verteilung

```
Niveau (alpha)   Freih.grade (df)   krit. Wert (t)
0.01              1                63.656741163
...
```

Sie können sich analog Quantile ("kritische Werte") weiterer Verteilungen berechnen:

Verteilung	SAS Funktion	Parameter	Beispiel
Normalverteilung	PROBIT(p)	p: Wahrscheinlichkeit	n1=probit(0.75);
Chi-Quadrat-Verteilung	CINV(p,df)	p: Wahrscheinlichkeit df: Freiheitsgrade	q1=cinv(.95,5);
Fisher-Verteilung	FINV(p,ndf,ddf)	p: Wahrscheinlichkeit ndf: Freiheitsgrade im Nenner, ddf: Freiheitsgrade im Zähler	f1=finv(.975,1,57);
t-Verteilung	TINV(p,df)	p: Wahrscheinlichkeit df: Freiheitsgrade	t1=tinv(.075,47)

**16.Index**

Abspeichern von Graphiken, 62  
 alphanumerische Werte, 26

**Alternative**, 45

**analysierbar**, 4

Anweisungen, 16

Arbeitsschritte, 24

**Argumente**, 16

arithmetischer Mittelwert, 43

Bedingung, 55

Beispielprogramme, 1

Benutzerschnittstelle, 14

**Beobachtung**, 3, 17

Bezeichnung für Datenwerte, 57

Bezeichnung für Variablen, 56

**Binomialverteilung**, 8

Blatt, 42

Box-and-Whisker-Plot, 42

Chi-Quadrat-Test, 47

Chi-Quadrat-Verteilung, 48

Dateitypen, 60

**Dateneingabeschritt**, 15

Datenwert, 17

Datumsformat, 58

**Deskriptive Statistik**, 3

Diagramme, 38

Display Manager System, 15, 22

Einstellungen, 62

empirische Verteilungsfunktion, 41

**Erwartungswert**, 8, 42

externe Datei, 29

externe Dateinamen, 2

fehlende Datenwerte, 26

**Fenster**, 15, 22

Formatbezeichnung, 57

Formatsteuerung, 28

Funktionsumfang, 13

Fußnote, 62

Gesamterhebung, 4, 43

Graphikdatei, 63

**Grundgesamtheit**, 3, 9

Gruppen, 33

Häufigkeiten, 34

Hilfesystem, 63

**Hypothese**, 11, 45, 46, 48, 50

interaktive Entwicklung, 22

**Irrtumswahrscheinlichkeit**, 10, 44, 47

Kenngößen, 5, 41

**klassifizierend**, 4

kleinste Informationseinheit, 17

**Kommando**, 14, 22

Kommandozeile, 14, 22

Konfidenzgürtel, 51

Konfidenzintervall, 42, 51

Kontrollzentrum, 14

**Korrelation**, 9, 48

**Kovarianz**, 9

Kreuztabelle, 34

lineare Regression, 50

lineares Modell, 51

**listengesteuert**, 25

**LOG Fenster**, 15

Maßzahlen, 5, 41

mathematische Statistik, 10

Median, 42

Menüpunkte, 14

Meßfehler, 50

Mittelwert, 35

Modellgleichung, 51

neue Variablen, 54

**Normalverteilung**, 12, 41

**Normalverteilungsannahme**, 12

**Optionen**, 16

**OUTPUT Fenster**, 15

*Overlay*, 40

**Parametrische Statistik**, 11

prinzipielle Arbeitsweise, 13

**PROGRAM EDITOR Fenster**, 15

Quantile, 69

- Rohdaten, 25
- SAS Anweisung, 16
- SAS Datei, 15, 17, 25
- SAS Modul**, 13
- SAS Programm, 15, 20
- SAS Prozeduren**, 13
- Schätzung, 43
- Schätzwert**, 11
- Schätzwerte**, 52
- Spaltenpositionen, 27
- spaltenpositioniert**, 26
- Spannweite, 35
- Stamm, 42
- Standardabweichung**, 8, 35
- Statistical Analysis System**, 1
- statistische Verfahren, 49
- Stem-and-Leaf-Plot, 42
- Stichprobe**, 3, 5, 9, 43
- Summe, 35
- Syntaxbeschreibungen, 1
- Tabelle, 17
- Test auf Unabhängigkeit, 46
- Test auf Unabhängigkeit, 47
- Test auf Unkorreliertheit, 49
- Testgröße, 47
- Titelzeile, 62
- typographische Konventionen, 1
- unabhängig**, 9
- Variable, 17
- Variablenname, 17
- Variablennamen, 25
- Varianz**, 8
- Varianz des Fehlers**, 52
- Varianz des Modells**, 52
- Verarbeitungsschritte**, 15
- Verteilung**, 7
- Verteilungsannahme**, 10
- Wahrscheinlichkeit**, 7
- X-Y-Plot, 39
- Zeichenketten, 26
- Zelle, 17
- zentraler Grenzwertsatz**, 12
- Zielsetzung, 1
- zufälliger Fehler, 50
- Zufallsexperiment**, 6
- Zufallsvariable**, 7
- Zufallsvorgang**, 6
- Zuweisung, 54
- zweistufiger Name, 58

