Dateien für den Kurs
"Werkzeuge der empirischen Forschung"

Die angegebenen Variablennamen sind als Vorschläge zu verstehen.

1 Tibetische Schädel

Dateiname: tibetan.dat

Colonel L.A.Wadddell sammelte 32 Schädel in den südwestlichen und in den östlichen Bezirken Tibets. Der erste, Typ A, enthält 17 Schädel, die aus Gräbern in Sikkim und angrenzenden Gebieten Tibets stammen. Die übrigen 15 Schädel, Typ B, wurden auf einem Schlachtfeld im Lhasa-Gebiet gefunden, und man nimmt an, daß sie von Soldaten aus den Ostprovinzen stammen, die von den Kham bewohnt sind. Diese Schädel waren von besonderem Interesse, da man annahm, daß die Kham-Tibeter Überlebende eines bestimmten grundlegenden menschlichen Typs seien, die in keiner Beziehung zu den Mongolischen und Indischen Typen stehen, die in ihrer Umgebung leben.

Die Daten bestehen aus 5 Messungen an jedem Schädel. Die 5 Messungen geben folgende Werte an:

Var(1)=MAXLENGT:	größte Länge des Schädels
Var(2)=MAXWIDTH:	größte horizontale Breite des Schädels
Var(3)=HEIGHT:	Höhe des Schädels
Var(4)=UPFACE:	Obere Gesichtshöhe
Var(5)=FACEWIDT:	Gesichtsbreite,zwischen den äußersten Punkten der Backenknochen.

(Die Originalquelle gibt 45 weitere Messungen an jedem Schädel an.)

Dateistruktur: 32 Wertesätze, 5 Variablen

17 Wertesätze vom Typ A; 2 Leerzeilen; 15 Wertesätze vom Typ B.

MAXLENGT	MAXWIDTH	HEIGHT	UP-FACE	FACEWIDTH
Sp.1-5	Sp.9-13	Sp.17-21	Sp.25-29	Sp.33-37

2 Methadon-Behandlung Heroin-Süchtiger

Dateiname: heroin.dat

Die Daten sind die Zeiten, in Tagen, die Heroin-Abhängige in der Klinik verbringen. Es gibt 2 Kliniken und man nimmt an, daß die Kovariaten die Zeit, die die Süchtigen in der Klinik verbringen, beeinflussen.

Var(1)=ID:	Identifikationsnummer
Var(2)=CLINIC:	Kliniknummer (1;2)
Var(3)=STATUS:	Status, 0: zensiert; 1: aus der Klinik entlassen
Var(4)=TIME:	Aufenthaltsdauer in der Klinik (in Tagen)
Var(5)=PRISON:	im Gefängnis, 0: nein 1: ja
Var(6)=DOSE:	Dosis (in mg/Tag)

Datenstruktur: 238 Wertesätze, 6 Variablen

	ID	CLINIC	STATUS	TIME	PRISON	DOSE
Spaltennr.	1-3	9	17	25-27	33	41-43
	49-51	57	65	73-75	81	89-90

3 Sterblichkeit und Wasserhärte

Dateiname: water.dat

Diese Daten wurden zur Untersuchung von Umweltursachen für Krankheiten zusammengetragen. Sie geben die jährliche Sterblichkeitsrate von 100 000 Männern, gemittelt über die Jahre 1958-64 (Var(1)=MORTAL) und die Kalziumkonzentration (Anteile von 1 Mill.) im Trinkwasser (Var(2)=CALCIUM) an. Je höher die Kalziumkonzentration ist, je härter ist das Wasser. Für Städte, die wenigstens soweit im Norden wie Derby liegen, hat die Indikatorvariable NORTH den Wert 1, sonst 0. In welcher Beziehung stehen Sterblichkeit und Wasserhärte? Gibt es einen geografischen Faktor in der Beziehung?

Datenstruktur: 61 Wertesätze, 3 Variablen

MORTAL	CALCIUM	NORTH	MORTAL	CALCIUM	NORTH
Sp.1-4	Sp.9-11	Sp.13	Sp.18-21	Sp.26-28	Sp.30
	30 Zeilen			31 Zeilen	

4 Leistungen verschiedener Computer-CPUs

Dateiname: computer.dat

Die Daten sind Charakteristika, Leistungsmaßzahlen und relative Leistungsmaßzahlen von 209 CPUs. Die relativen Leistungsmaßzahlen sind relativ zu einem IBM370/158-3. Welche Faktoren beeinflussen Leistung und relative Leistung?

Dateistruktur: 209 Wertesätze; 8 Variablen

		Spalten
Var(1)=CYCLETIM:	Zykluszeit (ns)	1-3
Var(2)=MINMEMOR:	Min.Speicher (kb)	9-14
Var(3)=MAXMEMOR:	Max.Speicher (kb)	17-21
Var(4)=CACHE:	Cache-Größe (kb)	25-27
Var(5)=MINCHANN:	Min.Anzahl der Kanäle	33-34
Var(6)=MAXCHANN:	Max.Anzahl der Kanäle	41-43
Var(7)=RELPERF:	relative Leistung	49-52
Var(8)=ESTRELPF:	geschätzte relative Leistung	57-60

5 Darwins kreuzbefruchtete und selbstbefruchtete Planzen

Dateiname: darwin.dat

Diese Daten stammen aus Darwins Studie (1876) über Kreuz- und Selbstbefruchtung. Paare von Setzlingen desselben Alters, von denen der eine Partner durch Kreuzbefruchtung und der andere durch Selbstbefruchtung erzeugt wurde, sind untersucht worden. Sie wuchsen gleichzeitig auf, so daß die einzelnen Partner unter etwa gleichen Bedingungen aufgezogen wurden. Das Ziel war, die größere Vitalität kreuz-befruchteter Pflanzen zu demonstrieren. Die Daten sind die Endhöhen jeder Pflanze nach einem festen Zeitraum. Darwin konsultierte Galton zur Analyse dieser Daten.

Dateistruktur:
VAR(1)=PAIR, linksbündig Spalte: 1-2
VAR(2)=HEIGHT-K, kreuzbefruchtet
VAR(3)=HEIGHT-S, selbsbefruchtet

6 Die Entwicklung des t-Tests

Dateiname: ttest.dat

In dem Zeitschriftenartikel von William Sealy Gosset (1908) (Pseudonym: "Student"), in dem er den t-Test entwickelte, wird u.a. die folgende illustrative Datenmenge verwendet. Der Schlaf verschiedener Patienten wurde sowohl ohne Hypnose als auch unter Behandlung mit 2 verschiedenen Medikamenten (a) und (b) gemessen. Die mittlere Anzahl der zusätzlichen Schlafstunden, die durch die Medikamenteneinnahme erreicht wurde, wurde tabelliert.

Dateistruktur:
Zeile 1-10:
VAR(1)=nr 1-2
VAR(2)=a 9-12
VAR(3)=b 17-20
VAR(4)=ba 25-28
Zeile 11: Mittelwerte
Zeile 12: Standardabweichungen
Die Variable ba ist die Differenz b-a.
Beachten Sie: In Zeile 5, letzte Spalte ist ein Druckfehler. Dieser ist zu korrigieren.

7 Banknoten

Dateiname: banknote.dat

Diese Datei beschreibt jeweils 100 echte und falsche Banknoten. Ziel der Analyse ist es, festzustellen, ob die gegebenen Merkmale ausreichen, um echte und falsche Banknoten relativ einfach voneinander zu unterscheiden. An Hand einer gegebenen Banknote soll man mit großer Wahrscheinlichkeit feststellen, ob sie echt oder gefälscht ist.

Dateistruktur:
Zeile 1-100: echte Banknoten
Zeile 101-200: gefäschte Banknoten

		Spalte
VAR(1)=NUMMER:	laufende Nummer	1-3
VAR(2)=LAENGE:	Länge der Banknote	5-9
VAR(3)=LINKS:	Breite der Banknote, links gemessen	10-15
VAR(4)=RECHTS:	Breite der Banknote, rechts gemessen	16-20
VAR(5)=UNTEN:	untere Randbreite	22-25
VAR(6)=OBEN:	obere Randbreite	27-30
VAR(7)=DIAGONAL:	Länge der Bilddiagonalen	31-35

8 Toxaemia in der Schwangerschaft

Dateiname: toxaemia.dat

Die Daten wurden in Bradford, England, im Zeitraum von 1968 bis 1977 gesammelt und beziehen sich auf 13384 Frauen, die vor der Geburt ihres ersten Kindes stehen. Die Frauen wurden klassifiziert bzgl. sozialem Status (5 Kategorien, Skala: 1-5) und bzgl. der Anzahl der Zigaretten, die während der Schwangerschaft pro Tag geraucht wurden (Kategorisierung auf 3 Niveaus: 1: Nichtraucher, 2: 1-19 Zigaretten pro Tag, 3: $\geq$ 20 Zigaretten pro Tag) Die Daten bestehen aus den Angaben, ob die Frauen krankhafte Anzeichen (Bluthochdruck oder Proteinuria) während der Schwangerschaft hatten (1: ja, 0: nein).

Die Variablen sind:

VAR(1)=SOCIAL
VAR(2)=SMOKING
VAR(3)=HYPERTEN
VAR(4)=PROTEINU

9 Ägyptische Schädel

Dateiname: skull.dat

Messungen an insgesamt 120 männlichen ägyptischen Schädeln aus 4 Epochen wurden analysiert mit dem Ziel, herauszufinden, ob es zwischen den Epochen Unterschiede gibt.

Die Variablen sind:

VAR(1)=GR	Gruppierungsvariable, gibt die Epoche an:
	GR=1: 4000 v.u.Z.
	GR=2: 3300 v.u.Z.
	GR=3: 1850 v.u.Z.
	GR=4: 200 v.u.Z.
VAR(2)=MB	Maximum Breath
VAR(3)=BH	Basibrematic Height
VAR(4)=BL	Basialveolar Length
VAR(5)=NH	Nasal Height.

10 Bewertung beim Synchronschwimmen

Datei: synchro.dat

Die Datei gibt die Bewertung von 40 Wettkämpfern (Mannschaften) im Synchronschwimmen durch jeweils 5 Preisrichter an. Die Frage ist, ob alle 5 Preisrichter zuverlässig sind bzw. ob sie alle gleich bewerten.

Die Variablen sind:

VAR(1)=Wettkaempfer	Nr. des Wettkaempfers
VAR(2)-VAR(6)=R1-R5	Preisrichter1-Preisrichter5