



Hallo,

kurze Anmerkung: Diese Scripte stammen von 1999. Ich kann leider dazu

keine Fragen mehr beantworten! :-)

Euch trotzdem viel Erfolg!

Dorthe

dorthe@luebbert.net

EDV I,II,III

© Dorthe Lübbert, Dorthe.Luebbert@ruhr-uni-bochum.de

Dieser Text kann frei weitergegeben werden, solange dieses Copyright nicht entfernt wird (Script war viel Arbeit!)

1 EDV I: Definitionen.....	1
2 Statistische Funktionen von Excel 5.0 (deutsch).....	4
3 Statistikfunktionen mit SPSS 5.02 (englisch)	5
4 Praktische Aufgaben mit SPSS und EXCEL	6
4.1 T-Test mit EXCEL	6
4.2 Korrelationskoeffizient zwischen 2 metrischen Untersuchungsvariablen EXCEL/SPSS	8
4.3 Bivariater Korrelationskoeffizient.....	8
4.4 χ^2 -Unabhängigkeitstest mit irgendeinem Programm oder Programmiersprache	9
4.5 Ausschluß falscher Datensätze.....	11
4.6 Gemeinsamkeiten/Unterschiede EXCEL-SPSS	12

1 EDV I: Definitionen

Anwendersoftware:

Software, die kunden- oder anwenderspezifische Probleme löst; typisch sind Tabellen- und Textverarbeitungsprogramme

ASCII-Code:

American Standard Code for Information

Der ASCII-Code ist eine genormte Zuordnungsregel, die die Darstellung von Zeichen in Form von binären Zahlen ermöglicht. Dabei ist jedem Zeichen (von 256) eine eindeutige Nummer zugewiesen.

Ursprung: Übertragung von Daten durch Fernschreiber, dort wurde allerdings ein 7-Bit Code verwendet (128 Zeichen).

Batch-Programm:

Stapeldatei, Textdatei, die aus einer Anzahl einzelner Befehle besteht, die i.d.R. nacheinander vom Betriebssystem abgearbeitet werden

Benutzeroberfläche:

Als Benutzeroberfläche werden jene Teile eines Hardware- oder Softwaresystems bezeichnet, mit denen der Mensch beim Gebrauch in Kontakt kommt. Moderne Benutzeroberflächen erleichtern durch graphische Symbole (Icons) die Benutzung des Betriebssystems sowie der Programme.

Betriebssystem:

Das Betriebssystem ist ein Bündel von Dienst- Hilfs- und Organisationsprogrammen, die insbesondere für die Verwaltung von Informationen zuständig sind, sowie den Informationsaustausch innerhalb des Rechners und zwischen Rechner und peripheren Geräten organisieren, also die Verwaltung der Rechenanlage übernehmen.

Das Betriebssystem ist also die Schnittstelle von Anwendungsprogramm und Gerät; dient der Steuerung und Verwaltung der internen Rechnerfunktionen; erst das B. versetzt den Computer in die Lage zu arbeiten.

Betriebssystembefehl:

Anweisung aus dem Befehlsvorrat des Betriebssystems, z.B. „copy“ zum Kopieren von Dateien (MS-Dos), „ls“ (zum Auflisten von Verzeichnissen) (Unix).

Binärcode:

Zahlensystem zur Basis 2. Die kleinste Einheit in der Informationsverarbeitung ist das Bit, welches ein binären (0 oder 1, An oder Aus, Strom oder nicht) repräsentiert. Komplexere Zahlen müssen aus diesen Bits zusammengesetzt werden.

Binärsystem (auch Dualsystem):

Zahlensystem, das mit der Basis 2 arbeitet. Das Binärsystem ist in der EDV von grundlegender Bedeutung, weil mit seiner Hilfe einfachste elektronische Grundsaltungen zum Ausführen von Berechnungen verwendet werden können.

Bit oder Binärzeichen (binary digit):

Kleinste Informationseinheit: Ein Bit ist jedes der Zeichen aus einem Zeichenvorrat von zwei Zeichen: Flipflops, die meistgebrauchten Speicherelemente für Chips können zwei Zustände speichern, also ein Bit.

Booten:

das selbstständige Laden des Betriebssystems und bestimmter durch den Benutzer festgelegter Programme nach dem Einschalten des Computers

Byte:

Ein Byte ist eine Folge von acht Bits, die gemeinsam in einer elektronischen Datenverarbeitungsanlage verarbeitet werden.

Ein Byte ist die kleinste adressierbare Speicherstelle. Die acht Bits ermöglichen die Darstellung von 256 verschiedenen Zeichen (siehe ASCII-Code).

CD-ROM

compact disc - read only memory

optisches Speichermedium, kann im laufenden Betrieb nur gelesen werden, die Kapazität beträgt ca. 650 MB. Ein Laserstrahl trägt digitale Informationen auf die Platte auf, die so entstehende Spur wird beim Lesevorgang abgetastet. CD-ROMs sind ein gutes Medium für die Speicherung von relativ stabilen und umfangreichen Datenbeständen, die an unterschiedlichen Einsatzorten verfügbar sein müssen.

Code:

eine Verschlüsselungsvorschrift, nach der Informationen von einer Darstellungsform in eine andere überführt werden können; gibt z.B. an, wie die dem Menschen gemäße Schriftform in eine vom Computer zu verarbeitende Binärform gebracht werden kann.

Codebuch:

Übersicht mit Codierungsregeln

Informationen der Erhebung (Ausgangsdaten) werden vom Computer in bestimmte, einfach verstehbare Zahlenwerte transformiert; der Inhalt beschreibt diese Zuordnungen, z.B. Name der Variablen, Wertebereich, missing values, Feldgröße, Variablentyp.

Codierung:

Die Ausgangsdaten werden so transformiert, daß sie in einer EDV-Anlage übergeben werden können. Codieren bedeutet, daß Transformationsregeln vorgegeben werden müssen.

Datei (file):

Sammlung inhaltlich zusammengehörender Informationen, die gemeinsam auf einem Speicher abgelegt sind; Informationen können strukturiert sein, wobei dann eine Aufteilung in Datensätze oder Datenfelder erfolgt

*(dies ist aber nur eine logische Teilung einer Datei, physikalisch enthält eine Datei nur Hexdezimalzahlen).

Datenflußplan:

s. Programmablaufplan

Datenmatrix:

Die Datenmatrix dient dazu, die Ausgangsdaten gemäß der Regelungen des Codebuches so zu transformieren, daß sie direkt in den Rechner eingegeben werden können.

Beispiel:

1Bochum:	38112.3	959293515.1
2Bottrop	112 2.1	710152815.0
3Dortmund	56816.3	958198917.6
4Duisburg	51515.3	943299516.0

Die Datenmatrix kann (wie jede Matrix) von oben nach unten oder von links nach rechts gelesen werden, bzw. spalten- oder zeilenweise.

In jeder Zeile ist die Bezeichnung des Merkmalsträgers (Stadt), in jeder Spalte die Werte für die Variablen eines Datensatzes (*record, case*) enthalten.

Datensatz:

besteht aus einem Satz zusammengehöriger Daten eines Merkmalsträgers, alle Datensätze einer Datei müssen die gleiche Struktur haben, z.B. Datenfelder Name, Straße, Ort sind Datensatz einer Adressdatei

Datenträger:

Datenträger sind zur materiellen Verkörperung oder dauerhaften Aufnahmen von Daten geeignete physikalische Mittel. Sie unterscheiden sich z.B. in den Punkten „Speicherkapazität“, „Zugriffszeit“, „Datentransferrate“. Beispiele für Datenträger sind: Disketten, Streamerbänder, Magnetplatten, aber auch EC-Karten.

Dos:

(*disc operating system*): Betriebssysteme (PC/Mac), die Festplatten/Disketten etc verwalten können

* besser: Teil eines Betriebssystems zur Unterstützung von E/A-Operationen auf

* Diskettenstationen und Festplatten. Häufig als Synonym für MS-DOS verwendet.

MS-DOS/PC-DOS: altes PC-Betriebssystem aus den 80er Jahren, entwickelt von IBM als 16-Bit-Betriebssystem für den ersten PC (1981). Da Intel ihre PC-Familien mit einem Kompatibilitätsmodus auslieferte, ergaben sich in der ersten Hälfte der 80er Jahre keine Notwendigkeiten, MS-DOS im PC-Marktsegment durch ein leistungsfähigeres Betriebssystem

zu ersetzen. Erst Windows 97 verzichtet endgültig auf eine DOS-Unterstützung.

Dualsystem: s. binäres System

EBCDI-Code

Extenden Binary Coded Decimal Interchange

Codesystem zur Darstellung von Buchstaben, Ziffern und Sonderzeichen. Der EBCDIC benutzt zur Informationsdarstellung eine feste Länge von acht Bits, die in zwei sog. Tetraden zu je vier Bits unterteilt ist. Alphanumerische Daten werden zeichenweise den vereinbarten Bitkombinationen zugeordnet, d.h. für jedes alphanumerische Zeichen wird ein Byte vorgesehen.

Externer (peripherer) Speicher:

Speicher ist nicht Teil der Zentraleinheit; interner Speicher(CPU) ist der Hauptspeicher, alle anderen Speicher sind extern, z.B. Festplatte, Diskette

Hexadezimalsystem:

ein Zahlensystem zur Basis 16; Computerinterne Werte werden fast immer in h. Schreibweise angegeben; ein Byte entspricht einer zweistelligen Hexazahl.

* wird deswegen verwendet, da Speicher zu Worten von 16 Bit zusammengefasst wird

Kilobyte:

Ein Kilobyte ist die Maßeinheit für eine Speicherkapazität von 1024 Byte.

Im EDV-Bereich hat es sich eingebürgert, Kapazitätswerte durch Einheiten in Kilobyte auszudrücken.

Mikroprozessor:

Baustein, der alle notwendigen Steuerfunktionen durch geeignete Schaltungen übernehmen kann. Hauptbestandteile: Steuerwerk mit Befehlsregister, Decoder zur Decodierung und Ausführungsüberwachung von Maschinenbefehlen, Rechenwerk (Arithmetisch-Logische Einheit), Speicherzugriffssteuerung (?).

missing value:

Bei statistischen Analysen kommt es immer wieder vor, daß für bestimmte Merkmalsträger kein Merkmalswert vorliegt (Erfassungsproblem/Antwortverweigerung). Durch die Definition und Benutzung des „missing value“ wird der Computer in die Lage versetzt zu erkennen, daß kein „echter“ Wert vorliegt, der bei folgenden Berechnungen natürlich nicht mitverwendet wird. Beispielsweise könnte man bei einer erwarteten Spannbreite einer Merkmalsausprägung von „100-150“ den „missing value“ mit 0 kodieren.

Problemanalyse:

Die Problemanalyse ist ein Schritt bei der Entwicklung eines selbstgeschriebenen Programmes, bei dem das zu lösende Problem in Einzelprobleme zerlegt wird. Die logische Abfolge der Einzelschritte (Einzelprobleme) muß geklärt werden, nützlich ist hier die Erstellung eines Flußdiagramms (Programmablaufplan).

Problemorientierte Programmiersprache (Problemsprache, Benutzersprachen)

Problemsprachen (Benutzersprachen) dienen dazu, Datenverarbeitungsaufgaben aus bestimmten Problembereichen zu lösen. Beispiel: Fortran, Cobol, Algol. Problemsprachen sind die technische Weiterentwicklung von Maschinensprachen (Assembler), bei denen Einzelanweisungen durch definierte Schlüsselwörter ersetzt wurden.

Programm:

Folge von Befehlen an den Computer, die dieser selbständig ausführen kann und die alle zur Lösung einer Datenverarbeitungsaufgabe nötigen Anweisungen und Vereinbarungen enthält.

Programmablaufplan:

graphische Darstellung der Reihenfolge der Befehlsabläufe innerhalb eines Programms

Programmiersprache:

Bezeichnung für Verständigungsmöglichkeit mit dem Computer (Basic, Pascal, C), problemorientierte P. sind nicht direkt auf den Prozessor zugeschnitten, dadurch einfachere Handhabung, müssen von Compilern oder Interpreten in Maschinensprache übersetzt werden

RAM(-Speicher):

random access memory

Bezeichnung für Computerspeicher, der beliebig oft gelesen, beschrieben und gelöscht werden kann. Arbeitsspeichers eines Computers; direkter Zugriff auf jede Speicherzelle möglich

ROM:

read only memory

Bezeichnung für einen Festwert- oder Nurlesespeicher, dessen Daten i.d.R. unveränderbar sind - es sei denn man brennt das ROM z.B. mit einem EPROM-Brenner um - und nur ausgelesen werden können; der Speicherinhalt bleibt auch nach Abschaltung erhalten

Software:

Gesamtheit der fertigen Programme zur Steuerung einer elektronischen Datenverarbeitungsanlage, die zur Erledigung unterschiedlicher Aufgaben eingesetzt werden.

Speicherfeldvariable:

Ein Feld im Speicher des Computers, in dem variable Werte während der Programmbearbeitung unter einem vom User gewähltem Namen stehen.

Speicherkapazität:

Anzahl Bytes, die auf einem Speichermedium (Festplatte, EC-Karte) gesichert werden können

Statistik-Software:

Programme zur Lösung von statistischen Problemen, z.B. Excel, SPSS

Tabellenkalkulationsprogramm: s. S. 12 (Unterschiede Tabellenkalkulation-SPSS)

Typenkenning (Extension)**Unterverzeichnis:**

Ein in der Verzeichnisstruktur dem aktuellen Verzeichnis untergeordnetes Verzeichnis. Trivialer: Verzeichnis, das einem anderen Verzeichnis untergeordnet ist.

Verzeichnis:

Logische Zusammenfassung mehrerer Dateien. Mit Hilfe von Verzeichnissen lassen sich Dateien hierarchisch anordnen.

2 Statistische Funktionen von Excel

ACHSENABSCHNITT(Y_Werte;X_Werte)	Liefert den Schnittpunkt der Regressionsgeraden
ANZAHL	Berechnet, wie viele Zahlen eine Liste von Argumenten enthält
ANZAHL2	Berechnet, wie viele Werte eine Liste von Argumenten enthält
BESTIMMTHEITSMASS(Y_Werte; X_Werte)	Liefert das Quadrat des Pearsonschen Korrelationskoeffizienten
BINOMVERT(Zahl_Erfolge; Versuche; Erfolgswahrsch; Kumuliert)	Liefert Wahrscheinlichkeiten einer binomialverteilten Zufallsvariablen
CHITEST(Beob_Meßwerte; Erwart_Werte)	Liefert die Teststatistik eines Chi-Quadrat-Unabhängigkeitstests
CHIVERT(x; Freiheitsgrade)	Liefert Werte der Verteilungsfunktion (1-Alpha) einer Chi-Quadrat-verteilten Zufallsgröße
FISHER(x)	Liefert die Fisher-Transformation
FTEST(Matrix1; Matrix2)	Liefert die Teststatistik eines F-Tests
FVERT(x; Freiheitsgrade1; Freiheitsgrade2)	Liefert Werte der Verteilungsfunktion (1-Alpha) einer F-verteilten Zufallsvariablen
GAMMAVERT(x; Alpha; Beta; Kumuliert)	Liefert Wahrscheinlichkeiten einer Gammaverteilten Zufallsvariablen
GEOMITTEL(Zahl1; Zahl2; ...)	Liefert das geometrische Mittel
HARMITTEL(Zahl1; Zahl2; ...)	Liefert das harmonische Mittel
HÄUFIGKEIT(Daten; Klassen)	Liefert eine Häufigkeitsverteilung als einspaltige Matrix
KGRÖSSTE(Matrix;k)	Liefert den k-größten Wert einer Datengruppe
KKLEINSTE(Matrix;k)	Liefert den k-kleinsten Wert einer Datengruppe
KONFIDENZ(Alpha; Standabwn; Umfang_S)	Ermöglicht die Berechnung des 1-Alpha Konfidenzintervalls für den Erwartungswert einer Zufallsvariablen
KORREL(Matrix1;Matrix2)	Liefert den Korrelationskoeffizient zweier Reihen von
KOVAR(Matrix1;Matrix2)	Liefert die Kovarianz, den Mittelwert der für alle Datenpunktpaare gebildeten Produkte der Abweichungen
KRITBINOM(Versuche; Erfolgswahrsch; Alpha)	Liefert den kleinsten Wert, für den die kumulierten Wahrscheinlichkeiten der Binomialverteilung größer oder gleich einer Grenzwahrscheinlichkeit sind
MAX(Zahl1; Zahl2; ...)	Liefert den größten Wert innerhalb einer Argumentliste
MEDIAN(Zahl1; Zahl2; ...)	Liefert den Median der angegebenen Zahlen
MIN(Zahl1; Zahl2; ...)	Liefert den kleinsten Wert innerhalb einer Argumentliste
MITTELABW(Zahl1; Zahl2; ...)	Merkmalsausprägungen und ihrem Mittelwert
MITTELWERT(Zahl1; Zahl2; ...)	Liefert den Mittelwert der Argumente
NORMVERT(x; Mittelwert; Standabwn; Kumuliert)	Liefert Wahrscheinlichkeiten einer normalverteilten Zufallsvariablen

PEARSON(Matrix1;Matrix2)	Liefert den Pearsonschen Korrelationskoeffizienten
RANG(Zahl; Bezug; Reihenfolge)	Liefert den Rang, den eine Zahl innerhalb einer Liste von Zahlen einnimmt
RGP(Y_Werte; X_Werte; Konstante; Stats)	Liefert die Parameter eines linearen Trends
STABW(Zahl1;Zahl2;...)	Schätzt die Standardabweichung ausgehend von einer Stichprobe
STABWN(Zahl1;Zahl2;...)	Berechnet die Standardabweichung ausgehend von der Grundgesamtheit
STANDARDISIERUNG(x; Mittelwert; Standabwn)	Liefert den standardisierten Wert
STANDNORMVERT(z)	Liefert Werte der Verteilungsfunktion einer standardnormalverteilten Zufallsvariablen
STEIGUNG(Y_Werte; X_Werte)	Liefert die Steigung der Regressionsgeraden
SUMQUADABW	Liefert die Summe der quadrierten Abweichungen
TREND(Y_Werte; X_Werte; Neue_x_Werte; Konstante)	Liefert Werte, die sich aus einem linearen Trend ergeben
TTEST(Matrix1; Matrix2; Seiten; Typ)	Liefert die Teststatistik eines Student'schen t-Tests
TVERT(x; Freiheitsgrade; Seiten)	Liefert Werte der Verteilungsfunktion (1-Alpha) einer (Student) t-verteilten Zufallsvariablen
VARIANZ(Zahl1; Zahl2; ...)	Schätzt die Varianz, ausgehend von einer Stichprobe
VARIANZEN(Zahl1; Zahl2; ...)	Berechnet die Varianz, ausgehend von der Grundgesamtheit
VARIATIONEN(n;k)	Liefert die Anzahl der Möglichkeiten, um k Elemente aus einer Menge von n Elementen ohne Zurücklegen zu ziehen
WAHRSCBEREICH(Beob_Werte; Beob_Wahrsch; Untergrenze; Obergrenze)	Liefert die Wahrscheinlichkeit für ein von zwei Werten eingeschlossenes Intervall

2.1 Analyse-Funktionen (Add-In) von Excel

- Berechnung von Populationskenngrößen
- Einfaktorielle Varianzanalyse
- Exponentielles Glätten
- Fourieranalyse
- Gleitender Durchschnitt
- Histogramm
- Korrelation
- Kovarianz
- Rang und Quantil
- Regressionsanalyse
- Stichprobenziehung
- Zufallszahlengenerierung
- Zweifaktorielle Varianzanalyse mit Meßwiederholung
- Zweifaktorielle Varianzanalyse ohne Meßwiederholung
- Zweistichproben-F-Test
- Zweistichproben-t-Test bei abhängigen Stichproben
- Zweistichproben-t-Test unter der Annahme gleicher Varianzen
- Zweistichproben-t-Test unter der Annahme unterschiedlicher Varianzen
- Zweistichproben-Test bei bekannten Varianzen

3 Statistikfunktionen mit SPSS 5.02 (englisch)

Summarize

- Frequencies
- Descriptives
- Explore
- Crosstabs

Custom Tables

- Basic Tables

- General Tables
- Tables of Frequencies

Compare Means

- Means
- Independent samples T test
- Paired samples T test
- One-way ANOVA

ANOVA Models

- Simple Factorial
- General Factorial
- Multivariate

Correlate

- Bivariate
- Partial
- Distances

Regression

- Linear
- Logistic
- Probit
- Nonlinear

Loglinear

- General
- Hierarchical
- Logit

Classify

- K-Means Cluster Analysis
- Hierarchical Cluster Analysis
- Discriminant Analysis

Data Reduction

- Factor Analysis

Scale

- Reliability Analysis
- Multidimensional Scaling

Nonparametric tests

- Chi-square
- Binomial
- Runs
- One-sample Kolmogorov-Smirnov
- Two independent samples
- Several independent samples
- Two related samples
- Several related samples

Survival

- Life Tables
- Kaplan-Meier
- Cox Regression
- Cox w/Time Dep Cov

Multiple Response

- Define Sets
- Frequencies
- Crosstabs

4 Praktische Aufgaben mit SPSS und EXCEL

4.1 Einfaktorielle Varianzanalyse mit EXCEL

Lösung:

- Daten eingeben:

Düngung	Ertrag	Messung 1	Messung 2	Messung 3	Messung 4
1		67	67	55	42
2		98	96	91	66
3		60	69	50	35
4		79	64	81	70
5		90	70	79	88

Aus dem Menü <Extras><Analyse-Funktionen> wählt man den Menüpunkt „Einfaktorielle Varianzanalyse“ aus und gibt die entsprechenden Zellen an, in denen die Daten sich befinden (Eingabebereich).

Excel wirft folgende Tabelle mit den Ergebniswerten aus:

Einfaktorielle Varianzanalyse

Eingabe

Eingabebereich:

Geordnet nach: Spalten Zeilen

Beschriftungen in erster Zeile

Alpha:

Ausgabe

Ausgabebereich:

Neues Tabellenblatt:

Neue Arbeitsmappe

OK
Abbrechen
Hilfe

Anova: Einfaktorielle Varianzanalyse**ZUSAMMENFASSUNG**

Gruppen	Anzahl	Summe	Mittelwert	Varianz			
2	2	9	4,5	0,5			
2	2	10	5	2			
ANOVA							
Streuungsursache	Quadratsummen (SS)	Freiheitsgrade (df)	Mittlere Quadratsumme (MS)	Prüfgröße (F)	P-Wert	kritischer F-Wert (Rückweisungspunkt)	
Unterschiede zwischen den Gruppen	0,25	1	0,25	0,2	0,698488655	18,51276465	
Innerhalb der Gruppen	2,5	2	1,25				
Gesamt							
	2,75	3					

4.2 T-Test mit Excel

Gegeben sind die Werte einer metrischen Variablen, die auf der Grundlage einer großen Zufallsstichprobe gewonnen wurden. Diese Werte sind nach weiblichen und männlichen Befragten getrennt angegeben. Wie ist vorzugehen, wenn Sie die Hypothese prüfen wollen, ob das arithmetische Mittel der Werte bei den weiblichen Befragten sich signifikant von dem der männlichen Befragten unterscheiden? Zur Prüfung dieser Hypothese soll das Tabellenkalkulationsprogramm EXCEL eingesetzt werden. [WS96, 6 P]

Lösung:

Datensatz eingeben, z.B.

Stunden Fernsehen am Tag									
Männer	3	5	6	3	4	4	3	2	0
Frauen	5	5	7	1	3	8	9	1	6

<Extras><Analyse-Funktionen><Zweistichproben-T-Test für unterschiedliche Varianzen>

Zweistichproben t-Test: Unterschiedlicher Varianzen

Eingabe

Bereich Variable A:

Bereich Variable B:

Hypothetische Differenz der Mittelwerte:

Beschriftungen

Alpha:

Ausgabe

Ausgabebereich:

Neues Tabellenblatt:

Neue Arbeitsmappe

Excel wirft aus:

Zweistichproben t-Test unter der Annahme unterschiedlicher Varianzen

	Männer	Frauen
Mittelwert	3,33333333	5
Varianz	3	8,25
Beobachtungen	9	9
Hypothetische Differenz der Mittelwerte	0	
Freiheitsgrade (df)	13	
t-Statistik	-1,49071198	
P(T<=t) einseitig	0,07994883	
Kritischer t-Wert bei einseitigem t-Test	1,7709317	
P(T<=t) zweiseitig	0,15989766	
Kritischer t-Wert bei zweiseitigem t-Test	2,16036824	

(Rückweisungspunkt)

4.3 Korrelationskoeffizient zwischen 2 metrischen Untersuchungsvariablen EXCEL/SPSS

Wie ist vorzugehen, wenn Sie mit dem einen oder mit dem anderen Programm die Stärke des Zusammenhangs zwischen zwei metrischen Untersuchungsvariablen berechnen wollen [WS96, 3 P]

Lösung mit Excel:

1. Daten eingeben
2. freie Zelle suchen
3. <Einfügen><Funktion><Statistik> „Pearson“
4. Matrix 1: Reihe von unabhängigen Werten
5. Matrix 2: Reihe von abhängigen Werten
6. Der Bravais-Pearsonsche Korrelationskoeffizient kann abgelesen werden

Lösung mit SPSS (hier: englische Version 5.02):

1. Daten eingeben
2. mit <Data><Define variable> überprüfen, ob die Variablen richtig als numerisch definiert sind, evtl. den Variablennamen verändern
3. <Statistics><Correlate><Bivariate> auswählen
4. Die beiden Variablen in das Fenster der ausgewählten Variablen überführen (Klick auf den Pfeil)
5. Pearson auswählen
6. mit <OK> beschäftigen

SPSS wirft folgendes Output-File aus:

PEARSON CORR problem requires 80 bytes of workspace

- - Correlation Coefficient

```

VAR00002    VARIABL1
VAR00002    1,0000    ,4996
              ( 20)    ( 20)
P= ,         P= ,025

VARIABL1    ,4996    1,0000
              ( 20)    ( 20)
P= ,025     P= ,
    
```

(Coefficient / (Cases) / 2-tailed Significance) " . " is printed if a coefficient cannot be determined. Preceding task required 1,04 seconds elapsed

Übersetzungswahrscheinlichkeit
 SPSS führt gleichzeitig einen Test durch, in dem es prüft, ob die beiden Variablen aus einer Stichprobe stammen (Korrelationskoeffizienten-Test), r also nicht zufällig word. . SPSS setzt das SN selbstständig fest, der angegebene Wert P drückt die Überschreitungswahrscheinlichkeit aus. Daß z.B. $r=0,4996$ für Var 1 und Var 2 aus einer Stichprobe stammen, ist mit einer Überschreitungswahrscheinlichkeit kleiner 0,025% anzunehmen

4.4 Bivariater Korrelationskoeffizient

Gegeben ist eine Datenmatrix mit 5 metrischen Variablen (Stichprobenumfang $n=20$). Wie ist vorzugehen, wenn Sie die Matrix der bivariaten Korrelationskoeffizienten bestimmen wollen, indem Sie

- a) das Tabellenkalkulationsprogramm EXCEL
- b) das Statistikprogramm SPSS

einsetzen? (4 Punkte)

Wo sehen Sie die Vorzüge der einzelnen Programme im Vergleich miteinander (2 Punkte)

a) Matrix der bivariaten Korrelationskoeffizienten mit Excel:

	Var 1	Var 2	Var 3	Var 4	Var 5
Var 1	=PEARSON(\$A\$1:\$A\$20;A1:A20)	=PEARSON(\$A\$1:\$A\$20;B1:B20)	=PEARSON(\$A\$1:\$A\$20;C1:C20)	=PEARSON(\$A\$1:\$A\$20;D1:D20)	=PEARSON(\$A\$1:\$A\$20;E1:E20)
Var 2		=PEARSON(\$B\$1:\$B\$20;B1:B20)	=PEARSON(\$B\$1:\$B\$20;C1:C20)	=PEARSON(\$B\$1:\$B\$20;D1:D20)	=PEARSON(\$B\$1:\$B\$20;E1:E20)
Var 3			=PEARSON(\$C\$1:\$C\$20;C1:C20)	=PEARSON(\$C\$1:\$C\$20;D1:D20)	=PEARSON(\$C\$1:\$C\$20;E1:E20)
Var 4				=PEARSON(\$D\$1:\$D\$20;D1:D20)	=PEARSON(\$D\$1:\$D\$20;E1:E20)
Var 5					=PEARSON(\$E\$1:\$E\$20;E1:E20)

Korrelation von Var 0 = 0,4996
 Korrelation von Var 2 = 0,4996
 Korrelation von Var 2 = 0,4996; $r=1$

b) Matrix der bivariaten Korrelationskoeffizienten mit SPSS:

Anzahl der geprüften Elemente: n

Die Eingabe funktioniert genauso wie ein einfacher bivariater Korrelationskoeffizient für zwei Variablen (s.o.)

SPSS wirft folgendes Output-File aus:

PEARSON CORR problem requires 560 bytes of workspace.

```

- - Correlation Coefficients - -
      VAR00001  VAR00002  VAR00003  VAR00004  VAR00005
VAR00001  1,0000      ,4996     -,0482      ,2038     -,1073
          ( 20)      ( 20)      ( 20)      ( 20)      ( 20)
          P= ,      P= ,025     P= ,840     P= ,389     P= ,652
VAR00002  ,4996      1,0000     -,0163      ,5473     -,1544
          ( 20)      ( 20)      ( 20)      ( 20)      ( 20)
          P= ,025     P= ,      P= ,946     P= ,012     P= ,516
VAR00003  -,0482     -,0163      1,0000      ,5498     -,1495
          ( 20)      ( 20)      ( 20)      ( 20)      ( 20)
          P= ,840     P= ,946     P= ,      P= ,012     P= ,529
VAR00004  ,2038      ,5473      ,5498      1,0000      ,1691
          ( 20)      ( 20)      ( 20)      ( 20)      ( 20)
          P= ,389     P= ,012     P= ,012     P= ,      P= ,476
VAR00005  -,1073     -,1544     -,1495      ,1691      1,0000
          ( 20)      ( 20)      ( 20)      ( 20)      ( 20)
          P= ,652     P= ,516     P= ,529     P= ,476     P= ,

```

(Coefficient / (Cases) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
Preceding task required ,88 seconds elapsed.

b) Vorzüge und Nachteile der Programme

Die von mir verwendete Version Excel 5.0 ist gegenüber der Version SPSS 5.01 ungleich komfortabler in der Dateieingabe und Kommentierung. Dafür muß man die Formeln zur Berechnung der Korrelationskoeffizienten (mit ein bißchen Unterstützung durch das Auto-Ausfüllen von Excel) quasi per Hand erledigen. Änderungen im Datenmaterial (z.B. durch Korrekturen wegen falscher Werte) werden sofort in die Matrix übernommen, d.h. die Matrix wird automatisch korrigiert. Die Tabelle, die Excel auswirft, kann formatiert werden und steht optisch aufbereitet für andere Anwendungen zur Verfügung.

Im Gegensatz dazu ist die Berechnung bei SPSS wesentlich einfacher, ein Befehl muß nur ausgeführt werden, die Matrix inkl. Zusatzinformationen wird automatisch ausgeworfen. Ich vermute, daß SPSS für große Datensätze wesentliche Geschwindigkeitsvorteile bietet.

4.5 χ^2 -Unabhängigkeitstest mit irgendeinem Programm oder Programmiersprache

Gegeben ist die bivariate Häufigkeitsverteilung für zwei nominalskalierte Variablen, die auf der Grundlage einer großen Zufallsstichprobe gewonnen wurde. Wie ist vorzugehen, wenn Sie PC-gestützt die Hypothese prüfen wollen, daß zwischen beiden Variablen Unabhängigkeit besteht. Welches Softwareprogramm Sie dabei verwenden, bzw. ob Sie selbst programmieren (und ggf. in welcher Sprache) bleibt Ihnen überlassen [WS96, 6P]

Lösung mit Excel:

Chi-Quadrat-Unabhängigkeitstest. Einfügen-Funktion Chitest(beobachtete Daten, erwartete Daten) liefert die Prüfgröße (Teststatistik) mit den entsprechenden Freiheitsgrade

Lieblingsgetränk	Bier	Cola
Männer	1530	1535
Frauen	1800	1801
	0,8967434	CHITEST(B3:B4;C3:D4)

Lösung mit SPSS:

- variablen definieren (Partei, Geschlecht)
- Daten kodieren (<data value> für „Partei“ → 1=„CDU“, 2=„SPD“...>
- eingeben (jeder Fall muß einzeln eingegeben werden)
1,00 1 (Fall 1: CDU, männlich)
2,00 0 (Fall 4: SPD, weiblich)

4. <statistics><crosstabs> auswählen, die Variable für die Spalte und für die Reihe festlegen

5. unter <statistics> <chi-square> auswählen

SPSS wirft folgende Datei aus:

SEX Geschlecht by PARTEI Partei

		PARTEI					Page 1 of 1
Count		CDU	SPD	fdp	Grüne	sonstige	Row Total
SEX	0	1,00	2,00	3,00	4,00	5,00	14
weiblich		3	5	2	3	1	46,7
	1	4	5	3	2	2	16
männlich							53,3
Column Total		7	10	5	5	3	30
	Total	23,3	33,3	16,7	16,7	10,0	100,0

Chi-Square	Value	DF	Significance
Überschreitungswahrs. ↓			
Pearson (klass chi-quadrat)	,74617	4	,94552
Likelihood Ratio	,75242	4	,94471
Mantel-Haenszel test for linear association	,00035	1	,98508

Minimum Expected Frequency - 1,400
 Cells with Expected Frequency < 5 - 9 OF 10 (90,0%)
 Number of Missing Observations: 1

4.6 Normalverteilung mit EXCEL prüfen

Gegeben sind Angaben zur Körpergröße zufällig ausgewählter erwachsener männlicher Personen auf der Basis einer Zufallsstichprobe (n=1000). Wie ist vorzugehen, wenn sie mit dem Tabellenkalkulation EXCEL die Hypothese testen wollen, daß die Untersuchungsvariable in der GG normalverteilt ist.

Lösung:

Liefert die zweiseitige Prüfstatistik für einen Gausstest (Normalverteilung). Bei einem Gausstest wird, bezogen auf eine Datenmenge, (Matrix) für x ein standardisierter Wert erzeugt und als Ergebnis die zweiseitige Wahrscheinlichkeit der Normalverteilung geliefert. Mit dieser Funktion können Sie die Wahrscheinlichkeit schätzen, daß eine bestimmte Beobachtung aus einer bestimmten Grundgesamtheit stammt.

GTEST(Matrix; x; Sigma)

Matrix ist die Matrix oder der Datenbereich, gegen die/den Sie x testen möchten.

x ist der zu testende Wert.

Sigma ist die bekannte Standardabweichung der Grundgesamtheit. Fehlt dieses Argument, wird mit der Standardabweichung der jeweiligen Stichprobe gearbeitet

4.7 Demografische Angaben verarbeiten

Im Zuge einer empirischen sozialwissenschaftlichen Untersuchung auf der Grundlage einer schriftlichen (postalischen) Befragung sind demografische Variablen erfaßt worden (Geburtsjahr, Geschlecht, Familienstand, letzter Bildungsabschluß u.ä.). Skizzieren Sie nicht zu knapp die Arbeitsschritte, die erforderlich sind, um solche demografischen Angaben einer PC-gestützten statistischen Auswertung zugänglich zu machen.

Vorgehensweise:

Nach der Phase der Datenbereitstellung erhält man einen Datenbestand, der, da die Befragung postalisch vorgenommen wurde, schon vorgeordnet ist. Der erste Arbeitsschritt, der vor der elektronischen Weiterverarbeitung nötig ist, ist die Codierung.

Codieren bedeutet in diesem Zusammenhang, daß die Ausgangsdaten so transformiert werden, daß sie einer EDV-Anlage übergeben werden können. Der Code gibt dabei an, in welcher Weise die gegebenen Daten transformiert werden sollen. „Codieren“ bedeutet deshalb auch, daß Transformationsregeln vorgegeben werden müssen. Derartige Codierungsregeln werden in einer Übersicht festgelegt, die man Codebuch nennt.

Für dem demografische Daten könnte das Codebuch beispielsweise folgendermaßen aussehen:

Position	Inhalt	Werte	Name	Länge	Dezimalstellen	Typ
1.	laufende Nummer	01-1000	NR	4	0	numerisch

2.	Geburtsjahr	1890-1997	GEBURT	4	0	numerisch
3.	Geschlecht	Texte: w,m	sex	1	0	String
4.	Familienstand	Texte, z.B. ledig, verheiratet, geschieden, verwitwet	Famstand	10	0	String
5.	letzter Schulabschluß	Texte, z.B. Sonderschule Volksschule, Hauptschule Fachoberschule, Fachhochschule Fachabitur...	Bildung	30	0	String

Wichtig ist es, vor der Datenauswertung „missing values“ zu definieren. Durch die Definition und Benutzung des „missing value“ wird der Computer in die Lage versetzt zu erkennen, daß kein „echter“ Wert vorliegt, der bei folgenden Berechnungen natürlich nicht mitverwendet wird. Beispielsweise könnte man bei einer erwarteten Spannweite der Merkmalsausprägung „Geburtsjahr“ mit 1890-1997 „missing value“ mit „0“ kodieren. Bei den nominalskalierten Variablen, die nicht weiter kodiert werden, müssen die missing values Texte sein.

Die Wertebereiche der nominalskalierten Variablen sollte man im zweiten Schritt ebenfalls kodieren, z.B. ledig=0, verheiratet=1, verwitwet=2, keine Angabe=42.

Als nächster Arbeitsschritt folgt nun die Anlage des Datenübertragungsblattes. Diese Datenmatrix dient dazu, die Ausgangsdaten gemäß der Regelungen des Codebuches so zu transformieren, daß sie direkt dem Rechner eingegeben werden können. Im obigen Beispiel ergäbe sich beispielsweise 1 1972003 für Person 1, geboren 1972, Geschlecht 0=männlich, Familienstand 0=ledig, Schulabschluß=3 (Fachhochschulreife).

Mit der Datenmatrix ist die Ausgangsbasis für eine folgende computergestützte Auswertung erreicht.

4.8 Ausschluß falscher Datensätze

Im Zuge einer empirischen sozialwissenschaftlichen Untersuchung auf der Grundlage einer schriftlichen (postalischen) Befragung treffen einige hundert ausgefüllte Fragebogen ein. Unter anderem ist dabei die Variable „Geschlecht“ erfaßt. Skizzieren Sie nicht zu knapp die Arbeitsschritte, die erforderlich sind, um PC-gestützt zu prüfen, ob die Werte dieser Variablen im zulässigen Bereich sind, bzw. um diesbezüglich fehlerhafte Datensätze zu identifizieren [WS96, 6P]

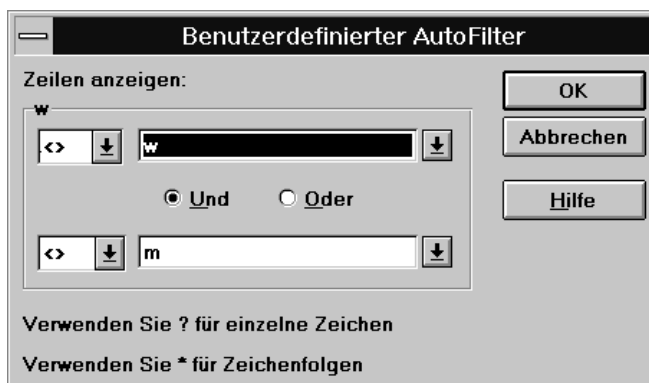
Lösung mit Excel:

Zunächst tippe ich die Daten ab oder (besser) scanne sie ein. Dieses eingescannte Rohdatenmaterial könnte mit einem Basic-Programm (besser: Perl-Script) auf falsche bzw. fehlende Werte überprüft werden.

Falls mir diese Möglichkeit nicht zur Verfügung steht, muß ich direkt aus meiner Software (hier: Excel) heraus die fehlerhaften Eingaben finden.

Dazu benutze ich (bei Excel 5.0) die Option des AutoFilters:

1. ersten Datenwert der Variable Geschlecht markieren
2. <Daten><Filter><Autofilter> auswählen
3. In der betreffenden Zelle erhalte ich ein Popdown-Menü, in der ich nach meinen Vorgaben bestimmte Werte herausfiltern kann, z.B. alle männlichen, alle weiblichen, alle=„k.A.“, alle=„Weiß nicht“ etc.
4. Alle Werte, die nicht männlich und weiblich sind, filtere ich heraus, indem ich einen <benutzerdefinierten Filter> entsprechend der obigen Abbildung einrichte.



5. Diese kann ich dann durch <Bearbeiten><Ersetzen> ersetzen, bzw. den „missing value“ einheitlich für alle Werte definieren.

4.9 Gemeinsamkeiten/Unterschiede EXCEL-SPSS

Skizzieren Sie die Gemeinsamkeiten und die Unterschiede zwischen einem Tabellenkalkulationsprogramm (zum Beispiel EXCEL) und einem Statistik-Programm (zum Beispiel SPSS). [WS96, 3 P]

Tabellenkalkulationsprogramme

dienen dazu, Berechnungen innerhalb von Zahlenmaterial durchzuführen und das Zahlenmaterial und die Ergebnisse dann optisch ansprechend darzustellen. Die Grundidee der Tabellenkalkulation ist (war) es, die Arbeitsweise mit Tabelle auf einem normalen Blatt Papier und mit einem Rechenstift nachzuvollziehen. Auf dem Bildschirm wird ein in Zellen gegliedertes elektronisches Arbeitsblatt dargestellt. Jede Zelle ist durch die zugehörige Zeilen- und spaltennummer eindeutig bestimmt. In den verschiedenen Zellen können in beliebiger Abfolge Zahlen, Texte, arithmetische und logische Ausdrücke mit oder ohne Bezugnahme auf andere Zellen eingetragen werden. Damit ist es dem Benutzer möglich, auf sehr flexible Art und Weise *individuelle Rechenschemata (Rechentabellen)* samt erklärendem Text zu gestalten. Tabellenkalkulationsprogramme bieten für den Statistiker vorformulierte Makros in Form von Funktionen an, z.B. zur Mittelwertberechnung, zum Korrelationskoeffizienten r etc. Der Benutzer kann diese Formeln zur Berechnung einsetzen, aber auch eigene Programme mit der zugehörigen Makrosprache programmieren, die ihm die Berechnung erleichtern. Insgesamt läßt sich sagen, daß der „statistische Sachverstand“ von Tabellenkalkulationen sehr gering ist, und die Interpretation und richtige Anwendung der Funktionen dem Benutzer überlassen wird. Ein großer Vorteil von Excel ist es, daß sich Formeln dynamisch dem Datenmaterial anpassen, d.h. wird eine Variable verändert, verändern sich automatisch alle darauf beziehenden Variablen.

Softwareprodukte wie SPSS

sind speziell auf die Lösung statistischer Aufgaben zugeschnittene Programme. SPSS bietet im Gegensatz zu Excel die Lösung von anspruchsvolleren Verfahren wie der Cluster- oder Faktorenanalyse, kann Zeitreihenanalysen ebenso durchführen wie nichtparametrische und parametrische Testverfahren. Die zugrundeliegenden Algorithmen sind optimiert, damit ist die schnelle Abwicklung auch umfangreicher Datensätze gegeben. Die Berechnung vollzieht sich in zwei Fenstern, im Datenfenster dürfen nur die Daten stehen, Ergebnisse werden ins Output-Fenster als unformatierten Text ausgegeben.

Ein Nachteil ist, daß Daten nur als Rohdaten eingegeben werden können, liegen bereits Häufigkeitsverteilungen vor, muß man die Daten wieder nach Einzelfällen aufschlüsseln. Will man Veränderungen im Datensatz vornehmen, müssen alle statistischen Verfahren nochmals neu mit dem Datenmaterial durchgeführt werden. Im Gegensatz zu Excel berücksichtigt SPSS auch Ausnahmen bzw. spezielle Voraussetzungen für Testverfahren. Es bietet wesentlich mehr Optionen als Excel. Der „statistische Sachverstand“ ist also beträchtlich höher als bei einer Tabellenkalkulation. Die Interpretation der Output-Daten bleibt dem Anwender überlassen, im Gegensatz zu Excel betitelt SPSS zumindest die Ergebnisse.

© Dorthe Lübbert, Dorthe.Luebbert@ruhr-uni-bochum.de

Dieser Text kann frei weitergegeben werden, solange dieses Copyright nicht entfernt wird (Script war viel Arbeit!)