

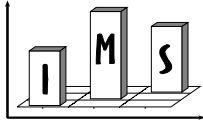
Institut für Medizinische Statistik der Universität Wien
Vorstand: o. Univ. Prof. Dr. P. Bauer

STATISTIK MIT SPSS

Alexandra Auterith

Institut für Medizinische Statistik
Universität Wien
Schwarzspanierstr. 17
1090 Wien

Wien, Dezember 2002



INHALT

Der Dateneditor.....	3
Berechnen von neuen Variablen.....	4
Zählen des Auftretens bestimmter Werte.....	4
Bilden von Subgruppen anhand einer metrischen Variable.....	5
Analyse für eine Auswahl an Fällen durchführen.....	6
Analyse getrennt für Subgruppen durchführen.....	6
Auswertung der Daten	
Analysieren – Deskriptive Statistik – Häufigkeiten.....	7
Analysieren – Deskriptive Statistik – Explorative Datenanalyse.....	8
Analysieren – Deskriptive Statistik – Kreuztabellen.....	8
Analysieren – Mittelwerte Vergleichen.....	10
Analysieren – Korrelation Bivariat.....	10
Analysieren – Regression – Linear.....	11
Analysieren – Überlebensanalyse – Kaplan Meier.....	11
Analysieren – Mittelwerte Vergleichen – t-Test bei unabhängigen Stichproben....	12
Analysieren – Mittelwerte Vergleichen – t-Test bei gepaarten Stichproben.....	13
Der Viewer.....	13
Der Syntax-Editor.....	13

Der Dateneditor

Eine neue Datendatei wird über DATEI-NEU-Daten angelegt. Eine bereits existierende SPSS-Datendatei wird mit dem Befehl DATEI-ÖFFNEN-DATEN geöffnet. Mit dem gleichen Befehl kann auch ein Datensatz aus einem anderen Programm (z.B. Excel) importiert werden. Zu einem Zeitpunkt kann immer nur eine Datendatei geöffnet sein. Der Dateneditor (Datenansicht und Variablenansicht) wird als Datei mit der Endung ".sav" gespeichert.

Der Dateneditor stellt zwei Ansichten der Daten bereit.

Datenansicht: Die Datenansicht enthält die Tabelle mit den Rohdaten. Die Zeilen enthalten die Beobachtungseinheiten, die Variablen werden in Spalten eingegeben. Zwischen den Zellen kann man sich mit der Maus oder mit den Pfeiltasten auf der Tastatur fortbewegen (wie in Excel). Es besteht allerdings keine Möglichkeit in der Datentabelle selbst Berechnungen durchzuführen oder Formeln einzugeben. Der Wert in den Zellen kann durch Anklicken der Zelle bzw. Drücken der Taste F2 verändert werden.

Variablenansicht: Die Variablenansicht ist eine Art "Legende" und enthält Informationen zu den Variablen des Datenblattes. Unter "**Name**" wird der Name einer Variable eingegeben. Dabei ist zu beachten, dass Variablennamen in SPSS maximal 8 Zeichen lang sein dürfen, keine Sonderzeichen enthalten und mit einem Buchstaben beginnen müssen. In der zweiten Spalte wird der Datentyp festgelegt. Folgende Datentypen sind vorhanden.

Datentypen:

Numerisch: Numerische Variablen sind Variablen, deren Werte Zahlen sind. Unter "Dezimalstellen" kann angegeben werden, wie viele Dezimalstellen gespeichert und im Dateneditor angezeigt werden.

Datum: Man kann aus einer Liste das gewünschte Format auswählen. Bei zweistelligen Jahresangaben wird das Jahrhundert durch die Einstellung in den Optionen bestimmt. Dieses kann im Menü Bearbeiten-Optionen im Register "Daten" verändert werden.

String: Variablen, die als String formatiert sind, können sowohl Buchstaben als auch Ziffern enthalten. Sie können nicht für Berechnungen verwendet werden.

Punkt: Variable, deren Werte mit Punkten als Tausender-Trennzeichen und Komma als Dezimaltrennzeichen angezeigt werden.

Komma: Variable, deren Werte mit Kommata als Tausender-Trennzeichen und Punkt als Dezimaltrennzeichen angezeigt werden.

Wissenschaftliche Notation: Die Werte einer numerische Variable, deren Werte mit einem E und einer Zehnerpotenz mit Vorzeichen angezeigt werden.

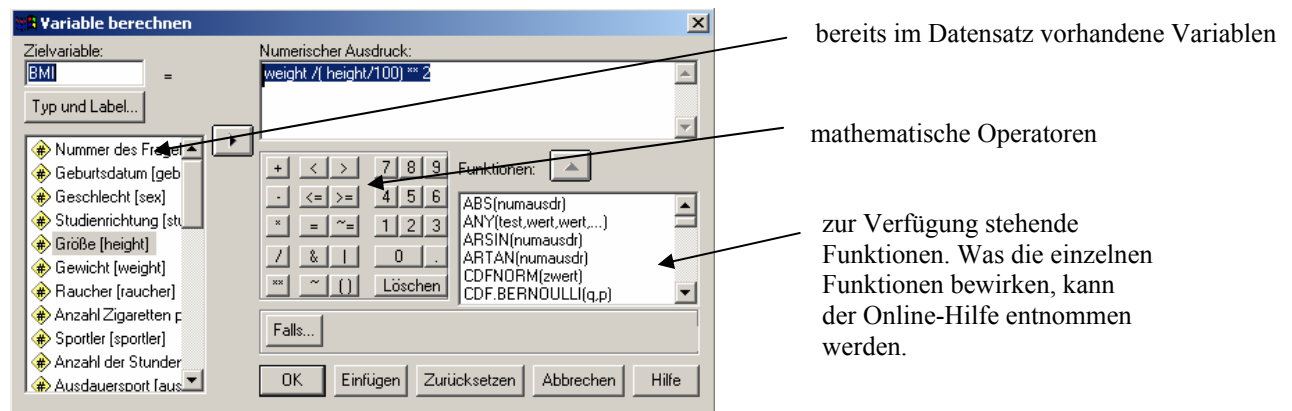
Spezielle Währung: In der Dialogbox BEARBEITEN - OPTIONEN können auf der Registerkarte "Währung" spezielle Währungsformate definiert werden.

Unter "**Spaltenformat**" wird angegeben, wie viele Zeichen eine Variable hat (z.B. 8 Ziffern). Da die kurzen Variablennamen in SPSS nicht sehr sprechend sind, ist es möglich, für jede Variable ein Variablenlabel (ein bis zu 120 Buchstaben langer Text) zu vergeben. Ob im Dateneditor die Variablennamen oder die Variablenlabels angezeigt werden, wird über den Befehl BEARBEITEN – OPTIONEN im Register "Allgemein" festgelegt.

Auch für die (meist numerisch codierten) Ausprägungen einer Variable können "**Wertelabels**" vergeben werden. Hier muss man auf den kleinen Pfeil klicken, damit sich die entsprechende Dialogbox öffnet. Über den Befehl ANSICHT – WERTELABELS wird bestimmt, ob im Dateneditor die Originalwerte oder die Wertelabels angezeigt werden. In der Datenansicht kann weiters die "**Ausrichtung**" der Werte in der Spalte festgelegt werden, sowie das "**Messniveau**". Außerdem können einzelne Merkmalsausprägungen als fehlende Werte definiert werden (z.B. 3 = "keine Angabe").

Berechnen von neuen Variablen

Die Berechnung neuer Variablen erfolgt in SPSS am schnellsten über das Menü TRANSFORMIEREN – BERECHNEN. Es öffnet sich die Dialogbox "Variable berechnen".



In das Feld "Zielvariable" wird der Name für die zu berechnende Variable eingegeben (z.B. BMI). Unter "Numerischer Ausdruck" wird die Berechnungsformel für die neue Variable eingegeben. Dabei können bereits vorhandene Variablen, die Operatoren in der Dialogbox und die Funktionen aus der Liste verwendet werden. Durch Klicken auf den Button "Typ und Label" können für die neue Variable auch gleich Variablen- und Wertelabels vergeben werden.

Wichtige mathematische Funktionen:

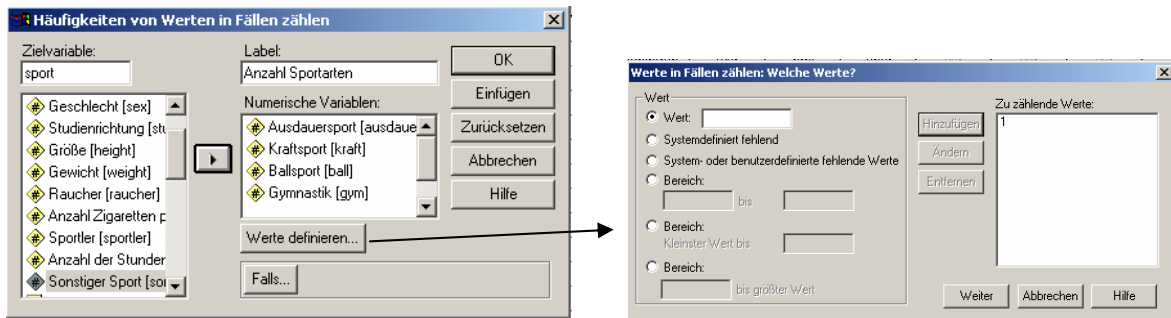
ABS(numerischer Ausdruck)	liefert den Absolutbetrag einer Zahl
SQRT(numerischer Ausdruck)	liefert die Quadratwurzel
LG10(numerischer Ausdruck)	Logarithmus zur Basis 10
LN(numerischer Ausdruck)	natürlicher Logarithmus

Zählen des Auftretens bestimmter Werte

In einer Fragebogenaktion wurde Studierenden folgende Frage gestellt:
Welche Sportarten betreiben Sie regelmäßig? (Mehrfachnennungen möglich)

- Ausdauersportarten (Laufen, Radfahren, Schwimmen)
- Krafttraining
- Ballsportarten
- Gymnastik, Aerobic

Da diese Frage Mehrfachnennungen erlaubt, muss jede Sportart als eigene Variable mit der Codierung 1 (ja) bzw. 0(nein) gespeichert werden. Möchte man nun für jeden Studierenden zählen, wie viele Sportarten er regelmäßig betreibt, so wählt man aus dem Menü TRANSFORMIEREN den Eintrag ZÄHLEN. Auch hier muss zunächst ein Name für die Zielvariable vergeben werden. Da im Beispiel der Wert "1" in der Variable steht, wenn ein Studierender die jeweilige Sportart betreibt, soll für jeden Studierenden gezählt werden, wie oft der Wert 1 in den Variablen "Ausdauersport", "Kraftsport", "Ballsport" und "Gymnastik" vorkommt. Die Variablen, die die Werte enthalten, die gezählt werden sollen (also die Variablen "Ausdauersport", "Kraftsport", "Ballsport" und "Gymnastik") werden in das Feld "Numerische Variablen" geklickt. Nun muss man über den Button "Werte definieren" festlegen, welche Werte gezählt werden sollen.

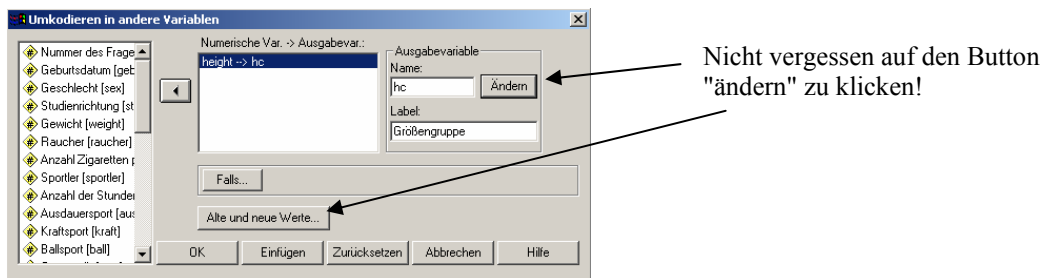


Dialogbox "Häufigkeiten von Werten in Fällen zählen"

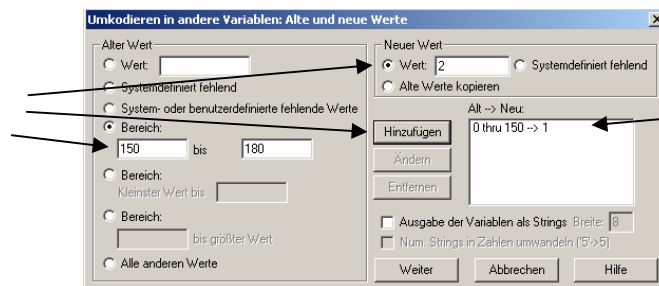
Bilden von Subgruppen anhand einer metrischen Variable

Möchte man eine metrische Variable (z.B. die Körpergröße) zur Bildung von Gruppen (z.B. klein-mittel-groß) verwenden, so erfolgt dies über den Befehl TRANSFORMIEREN – UMKODIEREN – IN ANDERE VARIABLE.

Die metrische Variable wird in das Feld "Eingabevar" geklickt. Dann wird der Name für die (neue) Ausgabevariable in das Feld "Name" eingefügt und auf den Button "Ändern" geklickt. Dann wird über den Button "Alte und neue Werte" eine Dialogbox aufgerufen, über die bestimmt werden kann, welche Werte der metrischen Variable in der neuen Variable in einer Gruppe zusammengefasst werden sollen, und wie diese Gruppe codiert sein soll.



der Bereich 150 bis 180 cm soll in der neuen Variable mit dem Wert 2 codiert werden

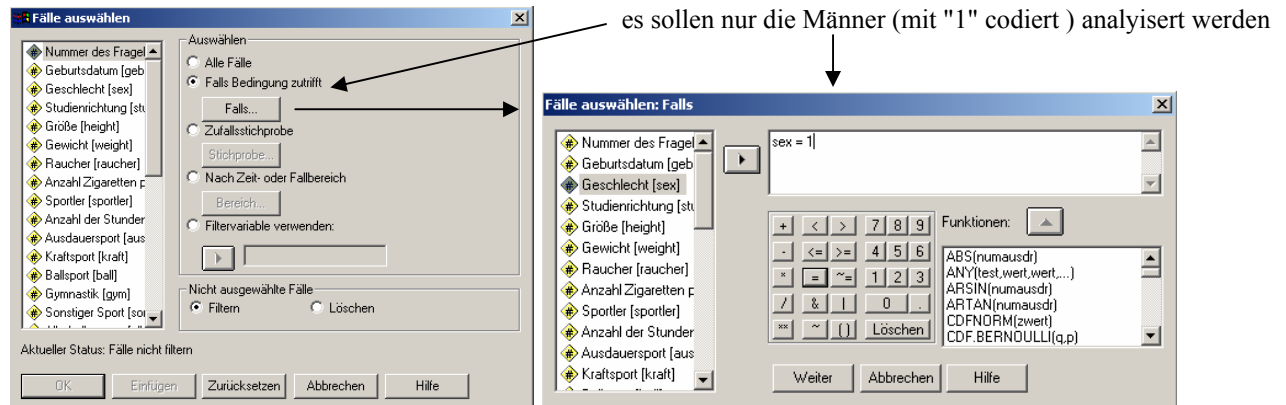


Der Bereich 0 bis 150 cm wird mit dem Wert 1 codiert.

Nachdem sämtliche Bereiche definiert wurden, wird die Dialogbox durch Klicken auf den Button "Weiter" bzw. "OK" geschlossen und die Berechnung wird durchgeführt.

Analyse für eine Auswahl von Fällen durchführen

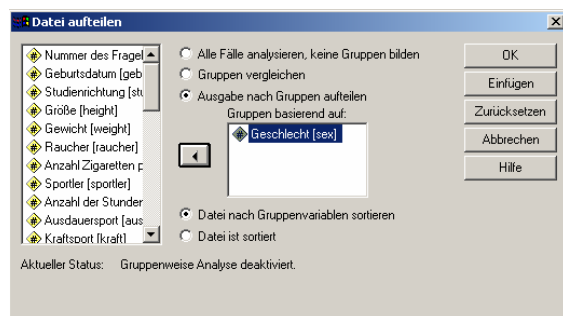
Manchmal möchte man eine Analyse nicht für alle Beobachtungseinheiten durchführen, sondern eine Frage lediglich für eine Subgruppe der Beobachtungseinheiten (z.B. alle Männer) beantworten. Über den Befehl DATEN – FÄLLE AUSWÄHLEN öffnet sich eine Dialogbox, in der man die Option "Falls Bedingung zu trifft" auswählt. Bei Klicken auf den Button "Falls" öffnet sich eine weitere Dialogbox, in der die Bedingungen (mit Hilfe der verschiedenen Schaltflächen in der Dialogbox) genauer definiert werden muss. Im Dateneditor werden die nicht ausgewählten Fälle durch eine durchgestrichene Zeilennummer gekennzeichnet.



Aufgehoben wird die Auswahl der Fälle durch Anklicken der Option "Alle Fälle" in der Dialogbox "Fälle auswählen".

Analyse getrennt für Subgruppen durchführen

Die Datenanalyse kann aber auch getrennt für Subgruppen (z.B. Männer und Frauen) durchgeführt werden. Dazu wählt man aus dem Menü DATEN den Befehl DATEI AUFTEILEN. Die Voreinstellung ist "Alle Fälle analysieren". Wählt man aber die Option "Ausgabe nach Gruppen aufteilen" und gibt in dem darunterliegenden Feld eine entsprechende Gruppierungsvariable an, so werden die Ergebnisse bei der Ausgabe in separaten Tabellen (Grafiken) dargestellt. Der Datensatz muss allerdings nach der Gruppierungsvariable sortiert sein, ansonsten bekommt man eine Fehlermeldung. Die Option "Datei nach Gruppenvariablen sortieren" bewirkt, dass diese Sortierung automatisch durchgeführt wird. Aufgehoben wird die Analyse nach Subgruppen indem man in der Dialogbox "Datei aufteilen" die Option "alle Fälle analysieren,..." wählt. Die Sortierung kann durch Sortierung über das Menü DATEN – FÄLLE SORTIEREN (etwa nach der Fragebogennummer) wieder rückgängig gemacht werden.



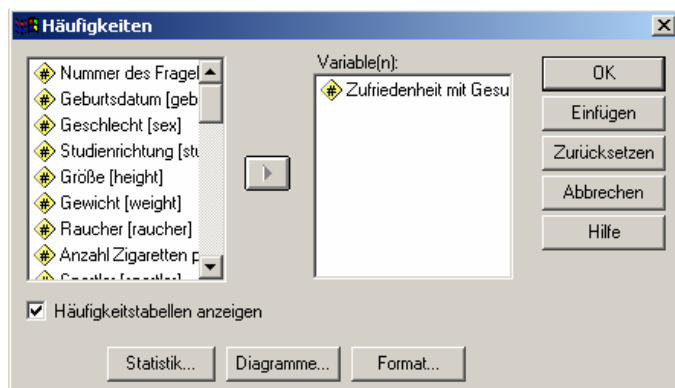
Auswertung der Daten mit SPSS


Es gibt in SPSS sehr viele Möglichkeiten Daten zu analysieren. Im Folgenden werden einige, insbesondere für die deskriptive Analyse wichtige Dialogboxen erklärt. Eine vollständige Auflistung kann hier freilich nicht erfolgen. SPSS verfügt über ein ausgezeichnetes Hilfesystem, das man bei Unklarheiten aufrufen sollte. In jeder Dialogbox ist ein Hilfe-Button vorhanden. Bei Klick mit der rechten Maustaste auf einen unbekanntenen Begriff erhält man meist eine hilfreiche Erklärung. Statistische Begriffe muss man gegebenenfalls in einem Statistikbuch nachschlagen.

Sämtliche für die Auswertung der Daten relevanten Dialogboxen können über die Menüs "Analysieren" und "Grafiken" aufgerufen werden.

ANALYSIEREN – DESKRIPTIVE STATISTIK – HÄUFIGKEITEN

Mit diesem Dialog lassen sich Variablen aller Skalenniveaus deskriptiv analysieren.

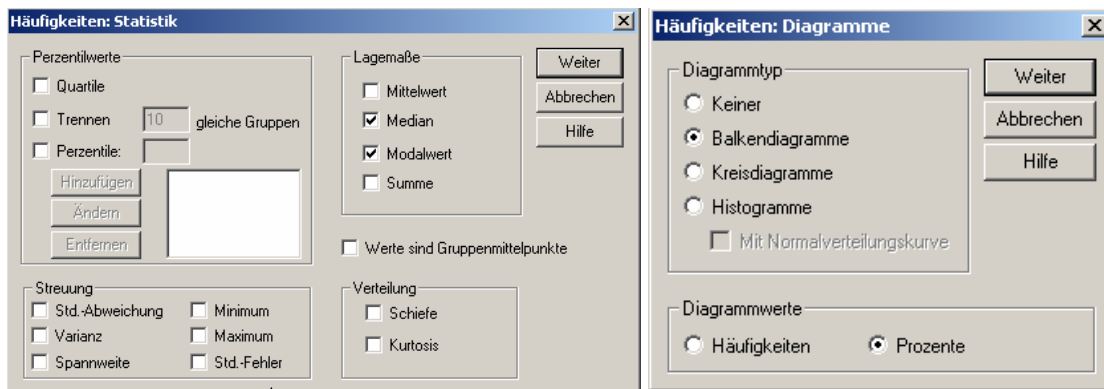


Auf der linken Seite ist zunächst die **Quellvariablenliste** zu sehen. Diese enthält alle Variablen des Datensatzes. Durch das Markieren einer Variable in der Quellvariablenliste und Klicken auf den Transportbutton , kann man die Variable in die **Wahlvariablenliste** klicken. Die Analyse wird für alle Wahlvariablen durchgeführt. Es wird davon abgeraten, Variablen unterschiedlichen Skalenniveaus gleichzeitig zu analysieren.

Auf der rechten Seite ist eine Gruppe von Befehlsschaltflächen zu sehen, die standardmäßig in allen Dialogboxen von SPSS vorkommt:

- OK: Startet die Prozedur, die Dialogbox wird geschlossen
- Einfügen: überträgt die Befehlssyntax in das Syntaxfenster
- Zurücksetzen: macht die Wahlvariablenliste wieder rückgängig
- Abbrechen: schließt die Dialogbox ohne eine Berechnung durchzuführen
- Hilfe: ruft ein Hilfefenster zu jeweiligen Dialogbox auf

Über die Buttons Statistik und Diagramme kann festgelegt werden, welche statistischen Kennzahlen berechnet bzw. welche Grafiken erstellt werden sollen. Vor der Auswahl sollte natürlich überlegt werden, welche Maßzahlen für welches Skalenniveau sinnvoll sind. SPSS gibt keine Warnung aus, wenn man etwa versucht den Median für ein nominalskaliertes Merkmale zu berechnen.

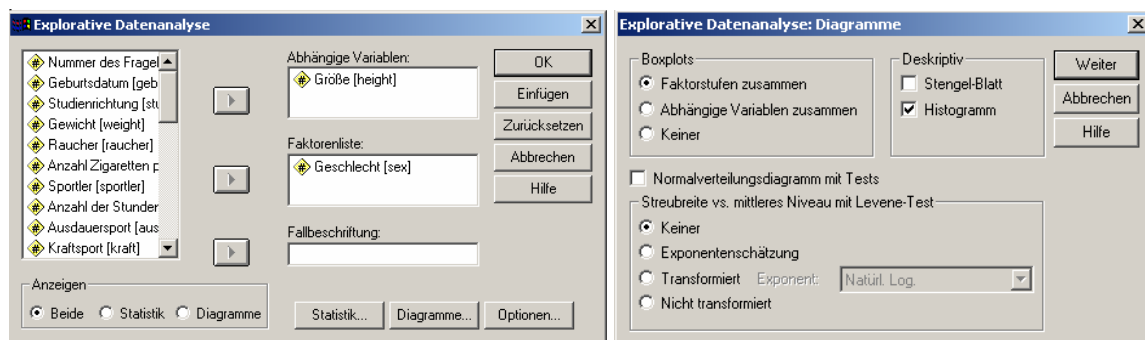


In der Statistik-Unterdialboxen kann man zwischen verschiedenen Lage- und Streuungsmaßen wählen

Es kann zwischen drei Diagrammtypen gewählt werden

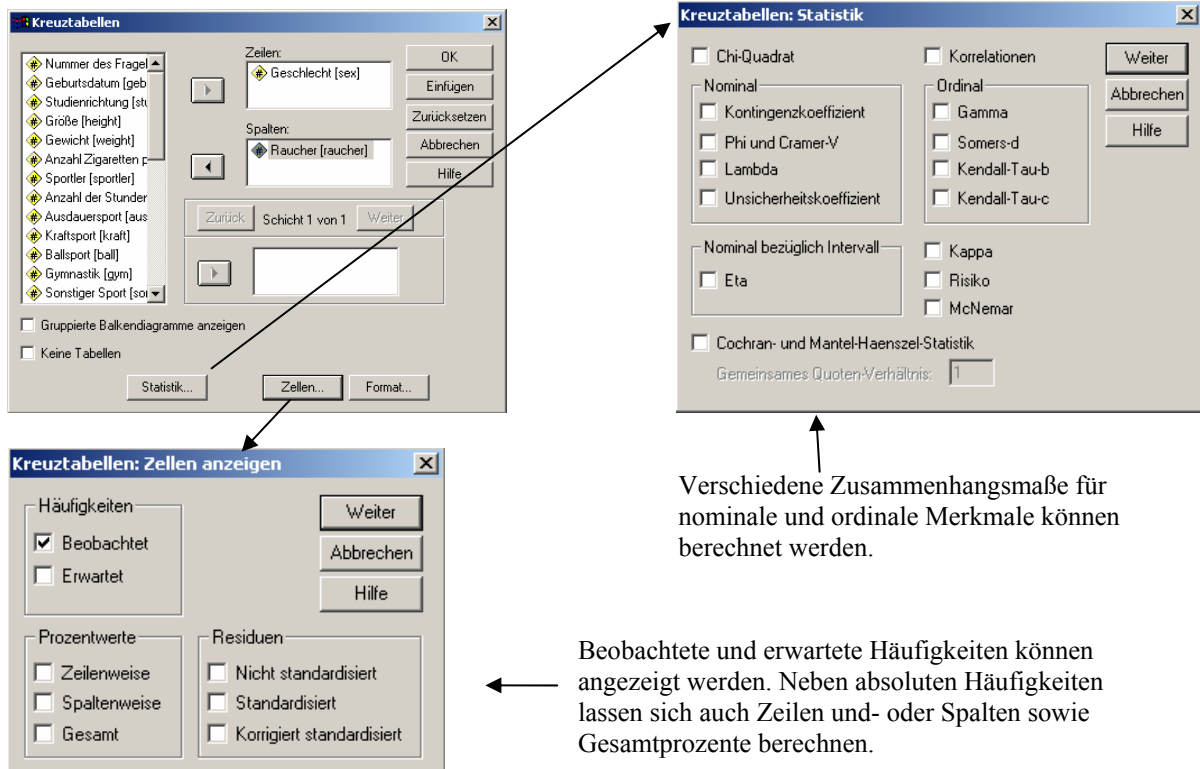
ANALYSIEREN – DESKRIPTIVE STATISTIK – EXPLORATIVE DATENANALYSE

Diese Dialogbox bietet zusätzlich die Möglichkeit, die Datenanalyse (für metrische Variablen!) getrennt für Subgruppen durchführen zu lassen. Die gewünschte Gruppierungsvariable (z.B. Geschlecht) wird dabei in das Feld "Faktorenliste" geklickt. Es werden automatisch sehr viele verschiedene statistische Maßzahlen berechnet. Als Diagramme kann man sich Boxplots und Histogramme ausgeben lassen. Als Fallbeschriftung könnte man etwa die Nummer des Fragebogens angeben. Ausreißer und extreme Werte im Boxplot werden dann etwa mit der Fragebogennummer beschriftet.



ANALYSIEREN – DESKRIPTIVE STATISTIK – KREUZTABELLEN

Das Erstellen von Kreuztabellen kann beispielsweise über diese Dialogbox erfolgen. Die Variablen müssen in die Zeilen bzw. Spalten geklickt werden. Falls man eine Variable in den Bereich "Schicht" klickt, wird für jede Merkmalsausprägung dieser Schichtungsvariable eine eigene Kreuztabelle (entsprechend den Variablen in den Zeilen bzw. Spalten) erstellt. Über den Button "Statistik" kommt man zu einer Dialogbox, in der man zwischen verschiedenen statistischen Zusammenhangsmaßen für Kreuztabellen wählen kann. Unter anderem findet man hier eine Option, die die Durchführung eines Chi-Quadrat Tests bewirkt. Bei Klick auf den Button Zellen öffnet sich ein Dialog, in dem man angeben kann, ob in den Zellen absolute Häufigkeiten bzw. Zeilen- oder Spaltenprozente angezeigt werden sollen. Auch die erwarteten Häufigkeiten können in einer Tabelle ausgegeben werden.



Geschlecht * Raucher Kreuztabelle

Anzahl		Raucher		Gesamt
		nein	ja	
Geschlecht	männlich	24	11	35
	weiblich	13	4	17
Gesamt		37	15	52

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	.348 ^b	1	.555		
Kontinuitätskorrektur ^a	.069	1	.792		
Likelihood-Quotient	.356	1	.551		
Exakter Test nach Fisher				.747	.403
Zusammenhang linear-mit-linear	.341	1	.559		
Anzahl der gültigen Fälle	52				

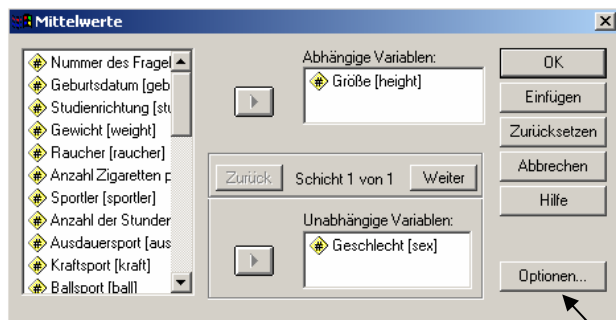
a. Wird nur für eine 2x2-Tabelle berechnet

b. 1 Zellen (25.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 4.90.

Aus den umfangreichen Ergebnistabellen müssen nun die richtigen Werte abgelesen werden. Der p-Wert des Chi-Quadrat Tests beträgt 0.555. Bei einem Signifikanzniveau von 0.05 würde man die Nullhypothese, dass es keinen Zusammenhang zwischen Geschlecht und Rauchen gibt daher nicht verwerfen. Es ist allerdings fraglich, ob der Chi-Quadrat Test in diesem Fall geeignet ist, da eine Zelle eine erwartete Häufigkeit kleiner 5 hat (wie in der Fußnote b angemerkt wird). In diesem Fall würde man den exakten Test nach Fisher mit einem zweiseitigen p-Wert von 0.747 bevorzugen.

ANALYSIEREN – MITTELWERTE VERGLEICHEN

Diese Dialogbox eignet sich insbesondere, um Mittelwerte verschiedener Subgruppen übersichtlich in einer Tabelle darzustellen.



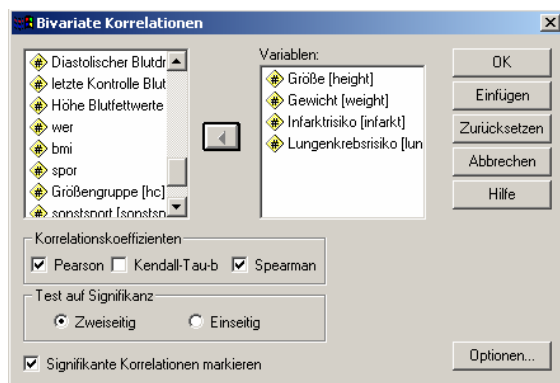
Bericht

Größe			
Geschlecht	Mittelwert	N	Standardabweichung
männlich	181.00	34	5.18
weiblich	171.88	17	4.83
Insgesamt	177.96	51	6.63

Außer dem arithmetischen Mittel können noch andere Maßzahlen für die Subgruppen berechnet werden.

ANALYSIEREN – KORRELATION – BIVARIAT

Zum Berechnen der Korrelationskoeffizienten nach Pearson bzw. Spearman wird die Dialogbox "Bivariate Korrelationen" aufgerufen. Im Variablenfeld können mehrere metrische Variablen angegeben werden. Es werden die paarweisen Korrelationskoeffizienten berechnet und in einer sogenannten Korrelationsmatrix angezeigt. Grafiken müssen über das Menü GRAFIKEN – STREUDIAGRAMM erstellt werden. Standardmäßig wird auch der p-Wert des Tests angezeigt, ob der Korrelationskoeffizient in der Grundgesamtheit von Null verschieden ist.



"Korrelationsmatrix"

Korrelationen

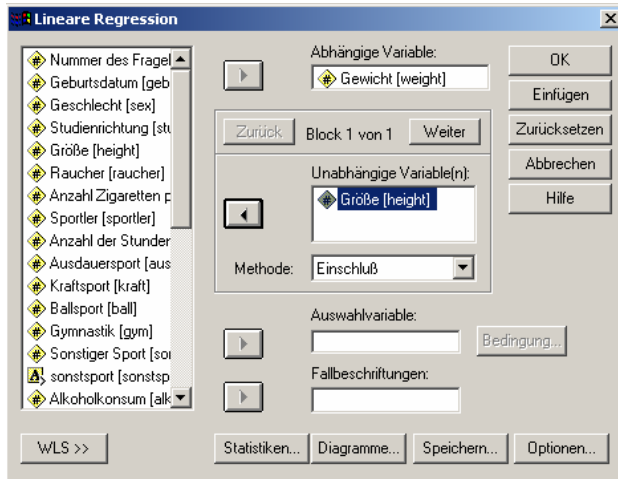
		Größe	Gewicht	Infarktisiko	Lungenkrebsrisiko
Größe	Korrelation nach Pearson	1.000	.620**	-.152	-.152
	Signifikanz (2-seitig)	.	.000	.286	.286
	N	51	51	51	51
Gewicht	Korrelation nach Pearson	.620**	1.000	.023	-.115
	Signifikanz (2-seitig)	.000	.	.874	.423
	N	51	52	51	51
Infarktisiko	Korrelation nach Pearson	-.152	.023	1.000	.558**
	Signifikanz (2-seitig)	.286	.874	.	.000
	N	51	51	51	51
Lungenkrebsrisiko	Korrelation nach Pearson	-.152	-.115	.558**	1.000
	Signifikanz (2-seitig)	.286	.423	.000	.
	N	51	51	51	51

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Im Beispiel zeigt sich eine auf dem Niveau 0.01 signifikante (positive) Korrelation zwischen Größe und Gewicht, sowie zwischen Infarktisiko und Lungenkrebsrisiko. Erwartungsgemäß liegt kein nennenswerter Zusammenhang zwischen Größe (bzw. Gewicht) und Infarkt- (bzw. Lungenkrebs-) Risiko vor.

ANALYSIEREN – REGRESSION- LINEAR

Die Berechnung einer Regressionsgerade erfolgt über den Befehl REGRESSION – LINEAR. Abhängige und unabhängige Variable müssen in das entsprechende Feld geklickt werden. Auf die verschiedenen Optionsmöglichkeiten kann an dieser Stelle nicht eingegangen werden.



Koeffizienten^a

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	-96.623	29.991		-3.222	.002
	Größe	.930	.168	.620	5.525	.000

a. Abhängige Variable: Gewicht

Aus der Tabelle "Koeffizienten" können Achsenabschnitt (-96.623) und Steigung (0.930) der Regressionsgerade abgelesen werden.

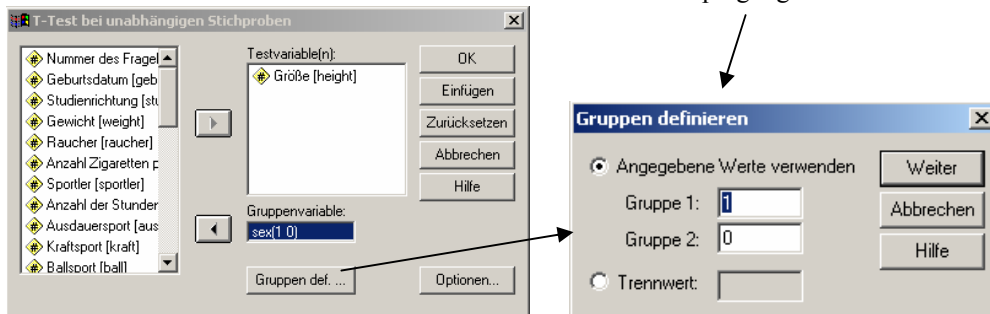
ANALYSIEREN – ÜBERLEBENSANALYSE – KAPLAN MEIER

Als "Zeitvariable" wird eine Variable gewählt, die angibt, wie lange die Beobachtungseinheiten jeweils überlebt haben (nichtzensierte Fälle) oder verfolgt wurden (zensierte Fälle). Die "Statusvariable" gibt an, ob für die jeweilige Beobachtungseinheit ein Event eingetreten ist oder nicht. Welche Ausprägungen dabei einem Event entsprechen muss über die Dialogbox "Ereignis definieren" festgelegt werden.

ANALYSIEREN – MITTELWERTE VERGLEICHEN – T-TEST BEI UNABHÄNGIGEN STICHPROBEN

Möchte man die Mittelwerte eines metrischen Merkmals für zwei von einander unabhängige Gruppen vergleichen, kann dies unter gewissen Verteilungsannahmen mit dem Zweistichproben t-Test erfolgen. Man könnte beispielsweise überprüfen, ob es zwischen der Körpergröße von Männern und Frauen einen "signifikanten" Unterschied gibt. Die metrische Variable (Körpergröße) wird in das Feld "Testvariablen" gezogen, die Gruppierungsvariable wird in das Feld Gruppenvariable geklickt. Die Ausprägungen der Gruppierungsvariable müssen in einer Dialogbox spezifiziert werden.

Die Gruppierungsvariable hat die Ausprägungen 1 und 0



Gruppenstatistiken

	Geschlecht	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Größe	weiblich	17	171.88	4.83	1.17
	männlich	34	181.00	5.18	.89

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
Größe	Varianzen sind gleich	.263	.611	-6.058	49	.000	-9.12	1.50	-12.14	-6.09
	Varianzen sind nicht gleich			-6.201	34.162	.000	-9.12	1.47	-12.11	-6.13

p-Wert für den Test auf Varianzgleichheit

p-Wert für den t-Test

Der Output ist etwas ungewohnt zu lesen. Zunächst muss man überprüfen, ob die Varianzen in den beiden Gruppen annähernd gleich sind. Sind sie das (meist kommt man zu diesem Ergebnis wenn der p-Wert für den Levene-Test auf Varianzgleichheit größer 0.05 ist), dann liest man die Signifikanz des t-Tests in der Zeile "Varianzen sind gleich" ab, ansonsten in der Zeile "Varianzen sind nicht gleich". Im Beispiel zeigt sich erwartungsgemäß ein hochsignifikanter Unterschied in der Körpergröße von Männern und Frauen. In der Tabelle ist als p-Wert der Wert 0.000 ausgewiesen, in einer Arbeit sollte er allerdings als $p < 0.001$ angegeben werden.

ANALYSIEREN – MITTELWERTE VERGLEICHEN – T-TEST BEI GEPAARTEN STICHPROBEN

Wird ein Merkmal an einer Beobachtungseinheit mehrfach gemessen (z.B. Gewicht vor und nach einer Diät) kann man mittels t-Test für gepaarte Stichproben überprüfen, ob zwischen den Vor- und Nachwerten ein signifikanter Unterschied besteht. Die Variablen, die die Vor- und Nachwerte enthalten werden dazu in das Feld "Gepaarte Variablen" geklickt. Die Eingabe wird mit OK abgeschlossen. Aus der Ergebnistabelle können sowohl der p-Wert als auch das Konfidenzintervall für die mittlere Differenz abgelesen werden.