Werner Voß / Veronika Khlavna / Nadine M. Schöneck

Einführung in die Datenanalyse und Datenmanagement mit SPSS

Bochum 2012

Inhaltsverzeichnis

Vorwort

Dieses Skriptum dient dazu, die wichtigsten statistischen Methoden zu erläutern, bei deren Einsatz das Statistikprogramm SPSS verwendet wird. Dieses Programm befreit von der Last mühsamer Berechnungen, so dass man sich ganz den sachlogischen Problemen der Datenauswertung und der inhaltlichen Interpretation der Ergebnisse zuwenden kann.

Auf die folgenden Anmerkungen machen wir einleitend besonders aufmerksam:

1. Das Programm SPSS besteht aus einem Basissystem und einer Reihe von Einzelmodulen. Je nachdem, welche Module installiert wurden, stehen unterschiedliche Verfahren und Methoden zur Verfügung. Dem entspricht es, dass in einigen der Abbildungen dieses Skriptums Menüpositionen oder Befehle auftauchen, die gegebenenfalls auf Ihrem Bildschirm nicht erscheinen. Wir stützen uns im Folgenden auf die Version SPSS 20. Sollten Sie eine andere Version verwenden, können Unterschiede in den Abbildungen zum SPSS-Bildschirm auftreten, denen aber keine besondere Aufmerksamkeit geschenkt werden muss.
2. Die Funktionen des Programms können auch durch Eingabe von Befehlen in der sog. SPSS-Sprache (SPSS-Syntax) ausgeführt werden. Über diese SPSS-Syntax können noch viel mehr und zusätzliche Möglichkeiten bereitgestellt werden, als über die SPSS-Menübefehle. Diese zusätzlichen Möglichkeiten werden aber im Folgenden nicht besprochen. Sie sind zu finden im SPSS Syntax Guide (aufzurufen über die Menüposition HILFE).
3. In den einzelnen Kapiteln dieses Skriptums werden möglichst einfache und überschaubare, meist recht kleine Zahlenbeispiele verwendet. Dies ist mit dem wesentlichen Vorzug verbunden, dass die Verfahren, um die es geht, leichter nachvollzogen und verstanden werden können.
4. In einigen Anwendungsbeispielen wird auf einen Datenbestand zurückgegriffen, der unter dem Namen „B00.sav“ im Ordner „Daten“ zu finden ist.

Veronika Khlavna
Nadine Schöneck
Prof. Dr. Werner Voß

1 Statistische Methoden

1.1 Aufgaben der Statistik

Überall da, wo wissenschaftliche Erkenntnisse auf der Basis empirischer Informationen gewonnen werden, sind Statistiker mit ihren Erhebungs-, Auswertungs- und Analysemethoden an der Arbeit. Das gilt insbesondere auch in den Sozialwissenschaften.

Man kann die Ergebnisse der statistischen Arbeit nur dann richtig verstehen und bewerten, wenn man weiß, wie die einzelnen statistischen Methoden funktionieren, und was die Begriffe, mit denen gearbeitet wird (Mittelwert, Streuung, statistische Signifikanz, Prognose, Korrelation usw.), überhaupt bedeuten. Vor allem muss man wissen, welche statistischen Auswertungsmethoden bereitstehen, was man damit machen kann, wie sie eingesetzt werden, und welche Befunde erzielt werden können. Deshalb werden in diesem Skriptum die wichtigsten dieser Methoden vorgestellt.

Der Statistiker hat es fast immer mit umfangreichen Datenbeständen zu tun, die ausgewertet werden müssen. Deshalb empfiehlt sich der Rechnereinsatz und der Einsatz von SPSS.

Allerdings, das soll in diesem Zusammenhang nicht verschwiegen werden, wer die Hintergründe der Verfahren kennt, auch wenn diese Kenntnisse nicht mehr zwingend erforderlich sind, der versteht natürlich mehr von den Methoden, die er einsetzt, und kann deshalb auch die Ergebnisse leichter und angemessener interpretieren.

1.2 Zum Begriff der Statistik

Tabellen mit Angaben (Daten) zur Konjunkturlage, grafische Darstellungen etwa der Wählerstruktur bei einer Bundestagswahl, Angaben zur Entwicklung der Arbeitslosenquote in den letzten Jahren, Auswertungen einer Marktumfrage u.ä. werden im Allgemeinen mit dem Begriff Statistik überschrieben. Genaugenommen sind solche Statistiken aber nur die Ergebnisse statistischer Arbeit. Man erhält diese Ergebnisse, wenn man sich *statistischer Methoden*bedient. Deshalb wollen wir im Folgenden unter „Statistik“ vor allem die „statistischen Methoden“ verstehen.

Folgt man einer Definition von *R. Wagenführ*, kann man deren Zweck folgendermaßen umreißen: *Statistische Methoden* werden benötigt, um Massenerscheinungen zu quantifizieren, zu beschreiben, zu beurteilen, Schlüsse aus ihnen zu ziehen und ihre Erklärung vorzubereiten.

Bevor wir uns dem Einsatz von SPSS zuwenden, ist es angebracht, die große Zahl unterschiedlicher statistischer Methoden und Verfahren zu klassifizieren, um einen ersten Überblick zu gewinnen. Dabei bieten sich verschiedene Unterteilungskriterien an.

Deskriptive und induktive Statistik

Von *deskriptiver Statistik*, bzw. von *deskriptiven statistischen Methoden* spricht man, wenn das Ziel der eingesetzten Verfahren die Beschreibung des Ausgangsdatenbestandes ist. Zum Beispiel zählt die Berechnung eines arithmetischen Mittels, eines Durchschnitts also, zu dieser Gruppe, weil Sie mit der Mittelwertberechnung Ihren Datenbestand zusammenfas-

send beschreiben können. Von *induktiven statistischen Methoden* hingegen spricht man, wenn auf der Grundlage von Stichprobendaten Rückschlüsse auf die Grundgesamtheit angestrebt werden, aus der die jeweilige Stichprobe stammt. Diese Rückschlüsse führen zu wahrscheinlichkeitsbehafteten Aussagen, weshalb die Verfahren dieser Gruppe auch manchmal dem Begriff der *Wahrscheinlichkeitsstatistik* untergeordnet werden. Häufig spricht man auch in diesem Zusammenhang von *schließender Statistik* oder von *beurteilender Statistik*.

Uni-, bi- und multivariate Methoden

Bei der statistischen Auswertungsarbeit interessiert sich der Statistiker häufig nur für eine einzige Untersuchungsvariable, zum Beispiel für das monatliche Nettoeinkommen einer großen Zahl abhängig Beschäftigter in der Bundesrepublik Deutschland. Alle Methoden, die er einsetzt, um die Einkommensangaben zu analysieren, zählen zu den Methoden der *univariaten Statistik* (uni = eins).

Wenn es aber um die Betrachtung von zwei Variablen gleichzeitig geht – zum Beispiel um den tagesdurchschnittlichen Zigarettenkonsum zufällig ausgewählter Erwachsener einerseits und um Angaben zum oberen Blutdruckwert dieser Personen andererseits –, wenn man sich also dafür interessiert, ob es vielleicht Zusammenhänge zwischen diesen beiden Variablen gibt, dann bedient man sich der Methoden der *bivariaten Statistik* (bi = zwei).

Schließlich kommt der Statistiker auf die Idee, dass die Angaben zum Blutdruck nicht nur vom Zigarettenkonsum, sondern vielleicht auch vom Alter, vom Geschlecht und vom Beruf der befragten Personen beeinflusst sein könnten. Will er solchen gemeinsamen Beeinflussungen auf die Spur kommen, benötigt er die Methoden der *multivariaten Statistik* (multi = viele = drei oder mehr).

Skalenabhängige Methoden

Bei diesem Einteilungskriterium geht es um die Frage der sogenannten *Skalenqualität* von statistischen Untersuchungsvariablen. Mit diesem Begriff ist der *Informationsgehalt* von Daten angesprochen. Es soll kurz erläutert werden, was damit gemeint ist:

Die Variable „Familienstand“ mit den Ausprägungen ledig, verheiratet, geschieden oder verwitwet ist eine *nominalskalierte Variable*, weil ihre Werte nur Etiketten sind, gewissermaßen also Namen. Diese erlauben nur Unterscheidungen zwischen einzelnen Personen – sonst nichts. Der Informationsgehalt beschränkt sich also auf die Feststellung von Unterschieden oder von Identitäten.

Anders ist es zum Beispiel mit der Variablen „Zeugnisnote“. Diese Variable beinhaltet nicht nur Unterscheidungs-/Identitätsinformationen, sondern zusätzlich auch eine Rangordnungsinformation. Eine solche Variable nennen wir *ordinalskalierte Variable*.

Können zusätzlich auch die Abstände (oder sogar die Quotienten) zwischen je zwei Werten einer Untersuchungsvariablen inhaltlich interpretiert werden, sprechen wir von einer *metrischen Skala*. Beispielsweise ist die Variable Körpergröße metrisch skaliert.

Natürlich können Sie zum Beispiel auch bei einer Ordinalskala Abstände oder Quotienten zwischen je zwei Werten berechnen – aber die Rechenergebnisse besagen inhaltlich nichts.

Aus diesen Überlegungen ergibt sich, dass bestimmte mathematische Operationen bei bestimmten Skalen inhaltlich sinnvoll sind, bei anderen aber nicht. Und dies wiederum bedeutet, dass es statistische Methoden gibt, die bei bestimmten Skalen eingesetzt werden können, bei anderen aber nicht.

1.3 Wichtige Grundbegriffe

Merkmale und Merkmalsträger

Interessiert sich der Statistiker zum Beispiel für die Körpergröße von erwachsenen Personen, dann wird die Körpergröße als *Merkmal* oder als *Variable* bezeichnet. Die Werte der Variablen, also die einzelnen beobachteten oder gemessenen Körpergrößen, werden *Ausprägungen* (*Variablenausprägungen*) oder *Merkmalswerte* (kurz *Werte*) genannt.

Zweckmäßigerweise unterscheidet man zwei Typen von Variablen:

Variablen, die nur ganz bestimmte Werte annehmen können, die streng voneinander getrennt sind, so dass keine Zwischenwerte möglich sind, werden *diskrete Variablen* genannt. Ein typisches Beispiel ist das Merkmal „Geschlecht“, das als Werte nur „männlich“ oder „weiblich“ annehmen kann. Weitere Beispiele wären etwa die Merkmale Familienstand, Kinderzahl, gewählte politische Partei usw.

Dagegen hat man es mit einer *stetigen Variablen* (Merkmal) zu tun, wenn die Variable im Prinzip jeden Wert und jeden Zwischenwert als Ausprägung annehmen kann. Beispiele dafür sind Einkommen oder generell Geldgrößen (wenn man davon absieht, dass zwischen benachbarten Centangaben im Allgemeinen keine Zwischenwerte angegeben werden), metrische Angaben wie Körpergrößen oder -gewichte, gemessene Zeiten, Temperaturen, Prozentangaben usw.

Die *Merkmalsträger* sind bei diesen Beispielen immer einzelne Personen gewesen. Merkmalsträger können aber auch Nationen sein, vielleicht mit den Merkmalen Bevölkerungszahl, Fläche, Bruttosozialprodukt; oder Gemeinden mit den Merkmalen Steueraufkommen, Grünflächenanteil, Bevölkerungszahl; oder Straßenkreuzungen mit dem Merkmal Zahl der Unfälle pro Monat; Werkstücke mit den Merkmalen Durchmesser, Gewicht, Schadhaftigkeit; Zuchtsauen mit den Merkmalen Gewicht, Zahl der Ferkel usw.; Autos mit den Merkmalen Farbe, Hubraum usw.

Stichprobe und Grundgesamtheit

Wenn sich der Statistiker für bestimmte Sachverhalte, Tatbestände oder Entwicklungen quantitativer Art interessiert, dann muss er versuchen, die entsprechenden Daten zu finden. Es bieten sich generell zwei Wege an: Entweder er betrachtet die *Grundgesamtheit* aller in Frage kommenden Merkmalsträger, oder er beschränkt sich auf eine *Teilerhebung* aus dieser Grundgesamtheit, die man auch *Stichprobe* nennt. Es leuchtet unmittelbar ein, dass eine Stichprobe rascher zu Ergebnissen führt, und dass die Datenerhebung auf Stichprobenbasis viel preiswerter ist als eine *Totalerhebung* (Auszählung der Grundgesamtheit). Weiterhin ist einleuchtend, dass eine Totalerhebung dann vorzuziehen ist, wenn es – aus welchen Gründen auch immer – unbedingt erforderlich ist, alle in Frage kommenden Merkmalsträger zu untersuchen. Genauso einleuchtend ist es, dass in manchen Fällen Totalerhebungen nicht möglich oder nicht sinnvoll sind – beispielsweise dann, wenn produzierte Autos zum Zweck der Qualitätskontrolle zerstörenden Crash-Tests unterzogen werden.

2 Statistische Daten

2.1 Start von SPSS

Wenn Sie das Programm SPSS starten, gelangen Sie zu einem Startfenster, das in Abbildung 2.1 vorgestellt ist.

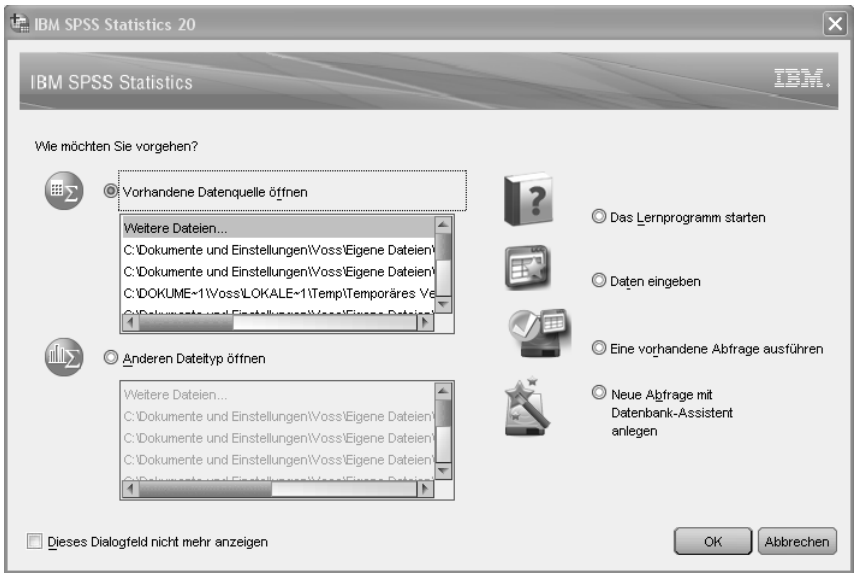


Abb. 2.1: Startfenster

Im Startfenster der Abbildung 2.1 werden Ihnen mehrere Möglichkeiten des Arbeitens mit SPSS geboten. Insbesondere haben Sie die Möglichkeit, direkt Dateien öffnen, mit denen Sie zuletzt gearbeitet haben, und Sie können Dateien aus anderen Anwendungen öffnen.

Wir interessieren uns zunächst für die Option DATEN EINGEBEN, weshalb wir den runden Schalter bei dieser Option anklicken und danach die Schaltfläche OK. Damit gelangen wir zum Startbildschirm von SPSS (siehe Abbildung 2.2).

In diesem Startbildschirm erscheint das erste von zwei Blättern, die sog. *Datenansicht*. Über das Blatt *Variablenansicht* wird weiter unten gesprochen. Sollte bei Ihnen diese Variablenansicht erscheinen, klicken Sie bitte auf Datenansicht.

In der zweiten Bildschirmzeile der Abbildung 2.2 (es handelt sich um den sog. SPSS Daten-Editor) erkennen Sie das SPSS-*Hauptmenü* mit den Positionen DATEI, BEARBEITEN, ANSICHT, DATEN... usw. (Menüpositionen, Schaltflächenbeschriftungen und Begriffe, die in den SPSS-Dialogfenstern auftauchen, schreiben wir im folgenden in KAPITÄLCHEN, damit Sie diese sofort als solche erkennen können). Zugleich erhalten Sie eine leere Tabelle, die der Aufnahme Ihrer Ausgangsdaten dient. Diese Tabelle trägt zunächst den Namen UNBENANNT1.

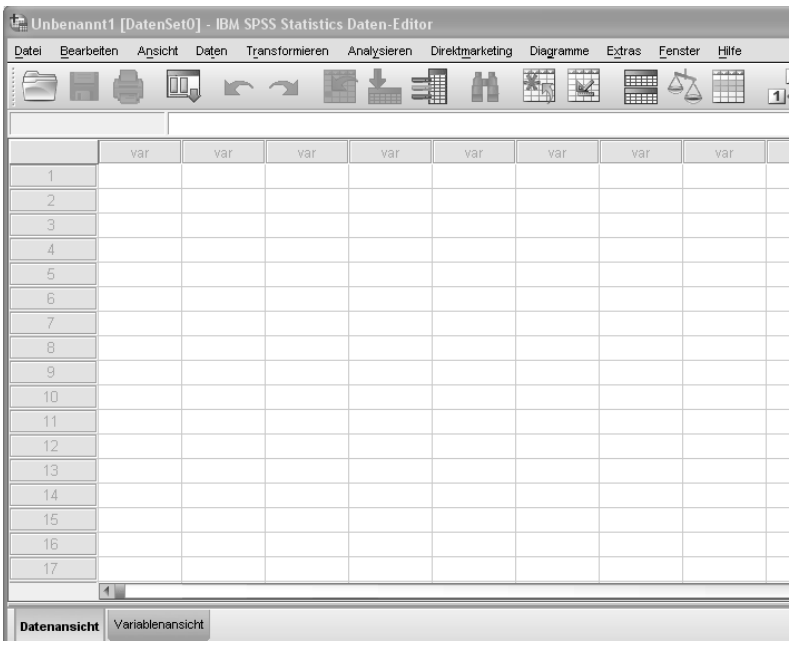


Abb. 2.2: Startbildschirm (Datenansicht)

Hinweis:

Wenn Sie Daten eingegeben haben und erste statistische Auswertungen durchführen, so erscheinen die Auswertungsergebnisse in einem zweiten Fenster. SPSS unterscheidet also prinzipiell zwischen dem Dateneingabefenster und zwischen Fenstern zur Präsentation von Ergebnissen.

Dem entspricht es, dass Ausgangsdaten und Auswertungsergebnisse in unterschiedlichen Dateien gespeichert werden. Erstere erhalten die Typenkennung .SAV, letztere die Typenkennung .SPV.

Bekanntlich öffnen Sie ein Menü durch Anklicken mit der Maus oder durch gemeinsames Drücken der Alt-Taste und dem in der Menüposition unterstrichenen Buchstaben. Jedes Menü können Sie durch Drücken der Esc-Taste wieder verlassen, oder indem Sie eine Stelle außerhalb des Menüs anklicken – wenn Sie zum Beispiel nur kurz einmal in das Menü hineinschauen wollten.

Ganz rechts in der Hauptmenüzeile finden Sie das Menü HILFE, mit dem Sie Hilfsinformationen aufrufen können.

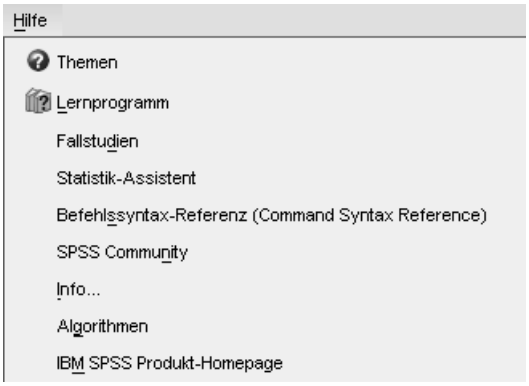


Abb. 2.3: Menü HILFE

2.2 Gewinnung statistischer Daten

Vor jeder statistischen Auswertung steht der Prozess der *Datengewinnung*. Welche Verfahren stehen hier zur Verfügung? Wir wollen uns an einem konkreten Beispiel orientieren: Es soll untersucht werden, ob das Wahlverhalten der Deutschen von demografischen Variablen beeinflusst wird. Um diese Frage zum Beispiel auf dem Weg einer Umfrage zu beantworten, müssen zunächst die folgenden Teilfragen beantwortet werden:

1. Wer soll befragt werden?
2. Wie soll befragt werden?
3. Wonach soll gefragt werden?

Die erste Frage könnte man beispielsweise so beantworten, dass aus der Gesamtheit der wahlberechtigten Bundesbürger eine *repräsentative Stichprobe* gezogen wird. Der einfachste Weg besteht darin, aus der Grundgesamtheit eine *Zufallsstichprobe* zu ziehen, indem zum Beispiel 200 Wahlberechtigte nach dem Zufallsprinzip (Losverfahren) ausgewählt werden.

Die zweite Frage, wie überhaupt befragt werden soll, zielt auf Überlegungen, ob beispielsweise *mündliche Interviews* oder aber eine *schriftliche Befragung* (postalische Fragebogenaktion oder z.B. eine Online-Befragung) angemessen ist. Die Verfahren weisen unterschiedliche Vor- und Nachteile auf, über die hier nicht gesprochen werden soll.

Unter inhaltlichen Gesichtspunkten ist die dritte die wichtigste Frage: Wonach soll gefragt werden. Ausgehend von der Themenstellung unseres kleinen Beispiels könnten die folgenden Sachverhalte bedeutsam sein:

1. Welcher Partei würden die Befragten ihre Stimme geben, wenn am kommenden Sonntag Wahlen zum Deutschen Bundestag stattfinden würden?
2. Geschlecht der Befragten
3. Alter der Befragten
4. Wohnort der Befragten
5. Konfession der Befragten
6. Bildungsstand der Befragten

etc.

Diese Aufzählung verdeutlicht, dass wir von der folgenden *Untersuchungshypothese* ausgehen: Das Geschlecht, das Alter, der Wohnort, die Konfession und der Bildungsstand wahlberechtigter Bundesbürger beeinflussen ihre Parteienpräferenz.

Bevor Sie sich dazu entschließen, nun tatsächlich eine eigene Erhebung durchzuführen, die ja Geld, Zeit und Mühe kostet, sollte man im konkreten Fall zunächst die Frage prüfen, ob man nicht an anderer Stelle schon entsprechende Befunde finden kann.

Sekundärstatistik

Bei vielen Fragestellungen ist es nicht notwendig, den mühsamen Weg einer eigenen Erhebung zu beschreiten. Häufig kann man nämlich auf vorhandene Daten zurückgreifen, die von anderen schon erhoben wurden, wenn man weiß, wo derartige Daten veröffentlicht sind.

Schon vorliegende Daten können nicht so aktuell sein wie die einer eigenen Erhebung; möglicherweise sind sie für den eigenen Untersuchungszweck schon veraltet. Des weiteren muss bedacht werden, dass schon vorhandene Daten häufig im Hinblick auf andere Zielsetzungen erhoben wurden, die nicht notwendigerweise mit denen der eigenen Untersuchung übereinstimmen. Unter Umständen muss auch berücksichtigt werden, dass bei ihrer Zusammenstellung vielleicht Erhebungsmethoden verwendet worden sind, die man selbst nicht eingesetzt hätte. In jedem einzelnen Fall muss deshalb überlegt werden, ob diese Daten für die eigenen Zwecke verwendbar sind.

Bei der Verwendung schon vorhandener Daten spricht man von *Sekundärstatistik*.

Primärstatistik

Wenn die sekundärstatistischen Quellen nicht die Informationen bereitstellen können, die Sie benötigen, muss man die notwendigen Daten selbst erheben. Dazu entwirft man einen *Fragebogen*, den man per Post verschickt, oder den man zu einer persönlichen Befragung durch Interviewer mitnimmt. Dabei ist die folgende Frage besonders wichtig:

Sollten Antworten vorgegeben werden, die dann nur angekreuzt werden müssen, oder lässt man den Befragten seine Antworten selbst formulieren? Beide Möglichkeiten sind mit Vor- und Nachteilen verbunden. Die Vorgabe von Antworten beispielsweise engt die Antwortmöglichkeiten ein und verzerrt so vielleicht die Ergebnisse, wobei auch eine Rolle spielt, dass der Befragte oft bestimmte Antworten vorzieht (man stellt häufig die Tendenz fest, dass lieber „Ja“ geantwortet und angekreuzt wird als „Nein“).Verzichtet man deshalb auf Antwortvorgaben, muss man im nachhinein die verschiedenen Antworten gruppieren oder klassifizieren, was oft nicht einfach ist.

Zieht man eine postalische Befragung vor, so werden einige (unter Umständen sehr viele) Befragte den Fragebogen gar nicht ausfüllen, nichts zurückschicken, und man weiß nicht genau warum. Man erhält dann unter Umständen eine kaum zu kontrollierende Verfälschung der Ergebnisse.

Es ergeben sich also einige Probleme, deren Aufzählung beileibe nicht vollständig ist. Man kann aber schon jetzt erkennen, dass die praktische Durchführung der Datenbereitstellung mit beträchtlichen Schwierigkeiten verbunden sein kann – weniger bei Variablen wie Geschlecht, Alter usw., sondern bei komplizierteren und interessanteren Größen, so dass die statistischen Ergebnisse, die auf diese Weise vorbereitet worden sind, eigentlich immer mit großer Vorsicht betrachtet werden müssen.

Hintergründe der Datenerhebung

Ein Statistiker interessiert sich für die Frage, welche Größen für den Rückgang der Kinderzahl pro Familie in der Bundesrepublik Deutschland (Abnahme der Geburtenrate) bedeut-

sam sind, und er vermutet, dass dies vor allem der gestiegene Verbrauch der Antibabypille sei. Ohne sich über die inhaltlichen Probleme seiner Themenstellung weiter Gedanken zu machen, erfasst er deshalb Geburtenrate und Verbrauch der Pille und stellt in der Tat fest, was auch zu erwarten war, dass die eine mit der anderen Größe statistisch eng zusammenhängt: Zunehmender Verbrauch der Pille = abnehmende Geburtenrate. Er hat aber dabei wichtige soziale und wirtschaftliche Größen übersehen: Es ist ja durchaus vorstellbar, dass wirtschaftliche und gesellschaftliche Gründe dafür verantwortlich sind, dass die Frauen die Pille nehmen (Notwendigkeit bzw. Wunsch der Ehefrau, berufstätig zu sein; zu kleine bzw. zu teure Wohnungen ohne ausreichende Kinderzimmer u.ä.). Derartige Zusammenhänge hat unser Forscher wegen der nicht durchgeführten notwendigen Vorüberlegungen nicht entdecken können. Seine Untersuchung hat kaum dazu beigetragen, neue Erkenntnisse zu erarbeiten. Die Aussage, dass zunehmender Pillenverbrauch geringere Kinderzahl bedeutet, ist wissenschaftlich wertlos, sie ist trivial.

Man sieht an diesem Beispiel sehr deutlich, wie wichtig es ist, vor einer konkreten Erhebung ein gegebenes Problem möglichst genau zu durchleuchten. Es kann sonst geschehen, dass man die eigentlich interessierende Wirklichkeit in statistisch fassbare Größen umsetzt, die diese Wirklichkeit gar nicht oder nur sehr unvollständig beschreiben. Wenn dies der Fall ist, dann können auch noch so viele Daten und die anspruchsvollsten Auswertungsmethoden keine Resultate liefern, die Erkenntnisse über diese Wirklichkeit bedeuten.

2.3 Arbeiten mit SPSS-Syntax

Warum SPSS-Syntax?

SPSS kann grundsätzlich auf zwei verschiedene Arten bedient werden. Die Befehle können zum einen über Menüs und Dialogfenster abgegeben werden. Zum anderen können die Befehle in der internen SPSS-Befehlssprache (SPSS-Syntax) verfasst und anschließend zur Ausführung gebracht werden. Der zweite Weg empfiehlt sich immer dann, wenn Prozeduren ausgeführt werden müssen, die über die Menüs nicht angeboten werden.

Die Verwendung von Syntax eignet sich auch zur Dokumentation ausgeführter Prozeduren. Dies ist vor allem dann von Vorteil, wenn verschiedene Personen gemeinsam an einem Datensatz arbeiten. Des Weiteren bietet sich das Arbeiten mit SPSS-Syntax für Berechnungen an, die wiederholt mit nur wenigen Änderungen vorgenommen werden. Auch können mit Hilfe von Syntax mehrere Befehle „gesammelt“ und dann gleichzeitig ausgeführt werden. Ferner lassen sich manche Befehle schneller manuell in der Befehlssprache erfassen, als menügesteuert.

> **Hinweis:**
>
> Wenn Sie Auswertungsergebnisse erzeugen, werden in der Ausgabedatei von SPSS auch die entsprechenden Syntax-Befehle ausgegeben. Diese können nach Anklicken separat in eine neue Syntaxdatei kopiert werden (siehe unten: Öffnen und Speichern der Syntaxdateien) und stehen dann – als Dokumentation oder zur späteren Wieder- oder Weiterverwendung – zur Verfügung. Sie können bei Bedarf wie z.B. Word-Dateien im Syntax-Editor verändert werden und kommen zur Ausführung durch die Menüposition AUSFÜHREN/ALLE im Syntaxfenster.

Öffnen und Speichern der Syntaxdateien

Das Verfassen der Befehle in Syntaxsprache erfolgt in einer von SPSS zur Verfügung gestellten Syntaxdatei. Ist SPSS bereits gestartet, so kann eine neue Syntaxdatei über das Menü DATEI/NEU/SYNTAX erstellt werden. Eine bereits bestehende Syntaxdatei kann über das Menü DATEI/ÖFFNEN/SYNTAX geöffnet werden. Dateien von Typ Syntax können an der Typenkennung SPS erkannt werden. Zur Speicherung einer Syntaxdatei ist das Menü DATEI/SPEICHERN UNTER... zuständig.

Der Vorteil gespeicherter Syntaxdateien liegt nicht nur darin, dass man so immer die Möglichkeit hat, auf bestimmte z.B. oft verwendete Prozeduren mit wenigen Änderungen, schnell und ohne großen Aufwand zuzugreifen. Auch um Speicherplatz zu sparen, sollten statt sehr voluminöser Ausgabedateien, die Syntaxdateien – diese werden im Prinzip wie Texte behandelt und benötigen somit wenig Speicherplatz – gespeichert werden.

Bearbeiten einer Syntaxdatei

Das Schreiben und Bearbeiten der Syntaxdateien erfolgt auf die gleiche Art wie bei einem beliebigen Text-Editor. Die SPSS-Befehle werden in englischer Sprache formuliert. Es können aber neben den normalen Buchstaben auch Ziffern von 0 bis 9 sowie Sonder- und Leerzeichen verwendet werden. Nicht zugelassen sind hingegen Umlaute sowie ß. Selbstverständlich müssen die Befehle in der Art verfasst sein, dass sie den Regeln der SPSS-Befehlssprache genügen. SPSS bietet aber auch die Möglichkeit, einen Befehl aus einem Menü- bzw. Dialogfenster direkt in die Syntaxdatei zu übertragen. Dies geschieht mittels der Schaltfläche EINFÜGEN, die sich in jedem Menü- und Dialogfenster befindet. Dieser Weg – also vom Menü zur Syntax – scheint nur auf den ersten Blick verwirrend. Zieht man die oben angeführten Vorteile der Syntax in Betracht, lohnt es sich meistens, einen Befehl einmal „doppelt“ abzugeben, um später viel Arbeit zu sparen.

An dieser Stelle sei noch darauf aufmerksam gemacht, dass die ausgeführten Befehle auch in der Ausgabedatei angezeigt werden. So hat man immer ein Überblick darüber, wie die Ergebnisse zustande kamen. Diese Einstellung wird – falls es noch erforderlich ist – im Menü BEARBEITEN/OPTIONEN vorgenommen. Dazu muss in diesem Menü im Register TEXT-VIEWER ein Häkchen bei BEFEHLE IM LOG ANZEIGEN gemacht werden.

Regeln der SPSS-Befehlssprache

Wie bereits erwähnt, müssen beim Erfassen von SPSS-Befehlen bestimmte Regeln beachtet werden. Diese sollen im Folgenden kurz vorgestellt werden.

Jeder SPSS-Befehl beginnt mit dem Namen des Hauptbefehls (command), dem weitere Unterbefehle (subcommands) folgen. Ein Befehl kann auch Schlüsselwörter (keywords) sowie Spezifikationen z.B. in Form von Variablennamen enthalten. Folgendes Beispiel mag dies verdeutlichen:

```
FREQUENCIES VARIABLES v00001 TO v00010
  /STATISTICS=MEAN
  /HISTOGRAM=NORMAL.
```

Dieser Befehl würde SPSS dazu veranlassen, die Häufigkeitstabellen sowie Mittelwerte und Histogramme der Variablen V00001 bis V00010 auszugeben. FREQUENCIES wäre dabei der Hauptbefehl, während VARIABLES, STATISTICS und HISTOGRAM die Unterbefehle wären. Das Wort TO ist ein Schlüsselwort und wurde in diesem Beispiel für die

Angabe der Spezifikationen (hier waren es die Variablennamen v00001 und v00010) benötigt.

Ein Befehl beginnt immer in einer neuen Zeile und muss immer mit einem Punkt abgeschlossen werden. Alle weiteren Zeilen, die die Unterbefehle enthalten, sollten eingerückt werden. Mehrere Unterbefehle können in einer Zeile hintereinander stehen, müssen dann aber durch einen Schrägstrich getrennt werden. Es sollte ferner beachtet werden, dass eine Zeile maximal 80 Zeichen lang sein darf. Unterbefehle können wie im obigen Beispiel Gleichheitszeichen enthalten. Darauf kann auch verzichtet werden, allerdings müssen die einzelne Anweisungen und Angaben dann durch mindestens ein Leerzeichen getrennt werden. Obiger Befehl könnte auch wie folgt geschrieben werden:

```
FREQUENCIES VARIABLES v00001 TO v00010
  /STATISTICS MEAN /HISTOGRAM NORMAL.
```

Es sei auch darauf hingewiesen, dass eine Leerzeile von SPSS als Abschlusspunk interpretiert wird.

Für die Variablennamen gilt, dass jede Variable einzeln in einen Befehl einbezogen werden kann. Die Variablenangaben können aber auch in Form von Aufzählung erfolgen, wobei die einzelnen Variablen dann entweder durch Leerzeichen oder Kommas getrennt werden (v00001 v00002 v00003 usw.). Es können auch ganze Variablenbereiche zusammengefasst werden; dies geschieht wie im Beispiel unter der Verwendung des Schlüsselworts TO. Durch die Angabe VARIABLES=ALL können schließlich alle Variablen eines Datensatzes in eine Auswertung einbezogen werden.

Beim Schreiben eines Befehls ist es nicht von Bedeutung, ob er mit Groß- oder Kleinbuchstaben geschrieben wird. Die Schreibweisen FREQUENCIES, frequencies und Frequencies sind gleichwertig. Das Gleiche gilt für die Variablennamen. SPSS selbst verwendet allerdings in der Regel (mit der Ausnahme der Variablennamen) die Großbuchstaben.

Sowohl Befehle, als auch Unterbefehle und Schlüsselwörter können auf minimal 3 Zeichen gekürzt werden. Entsteht dabei Ähnlichkeit mit anderen Befehlen oder Variablen, so muss die Abkürzung um ein Zeichen verlängert werden. Der Befehl „FREQUENCIES“ kann in der Schreibweise zum Beispiel auf „FREQ“, „FRE“ oder einfach „fre“ reduziert werden. Im den folgenden Ausführungen werden dem Leser auch die möglichen Abkürzungen der Befehle vorgestellt.

Kommentare in SPSS-Syntax

Vor allem im Hinblick auf die dokumentierende Funktion des SPSS-Syntax erscheint es sinnvoll, den geschriebenen oder eingefügten Befehlen Kommentare hinzufügen zu können, in denen z.B. kurz das Ziel der folgenden Berechnung erläutert wird. Auch diese Möglichkeit wird von SPSS angeboten.

Die Kommentare können an beliebiger Stelle in einer Syntaxdatei stehen (aber natürlich nicht mitten in einer Befehlsfolge). Um in eine Syntaxdatei einen Kommentar zu schreiben, der von dem Programm als solcher erkannt (und entsprechend akzeptiert) wird, bietet SPSS zwei Wege. Zum einen steht der Befehl COMMENT zur Verfügung, dem ein beliebiger Text folgen kann. Ein Kommentar kann aber auch mit einem „*“ eingeleitet werden (zwei und mehr * dürfen es auch sein). Hier ein paar Beispiele für verschiedene Kommentierungsmöglichkeiten:

```
COMMENT dies ist ein Beispiel für ein Kommentar.
*hier ein zweites Beispiel.
***ein drittes Beispiel***
```

Wird der Kommentar mit dem Befehl COMMENT eingeleitet, so sollte er mit einem Punkt abschließen. Kommentare, die mit * eingeleitet werden, können entweder mit einem Punkt oder mit * abschließen. Es empfiehl sich die Kommentare in mehrere Sternchen einzuschließen, denn so werden diese bei der Durchsicht einer Syntaxdatei sofort erkannt.

```
******************************************************
****so ein Kommentar fällt doch sofort auf, oder?****
******************************************************
```

Ausführen der SPSS-Befehle in einer Syntaxdatei

Um einen (oder mehrere) Befehl(e) in einer Syntaxdatei auszuführen, gibt es verschiedene Möglichkeiten. Wählt man Menü AUSFÜHREN/ALLE, so werden alle Befehle, die sich in der geöffneten Syntaxdatei befinden, ausgeführt. Möchte man nur einen oder mehrere bestimmte Befehl(e) ausführen, so sollten diese Befehle zuerst markiert werden und danach durch das Menü AUSFÜHREN/AUSWAHL zur Ausführung gebracht werden. Durch das Menü AUSFÜHREN/AKTUELLEN BEFEHL wird nur derjenige Befehl ausgeführt, in dem bzw. hinter dem sich der Cursor befindet. Schließlich bietet das MENÜ AUSFÜHREN/BIS ENDE die Möglichkeit, von dem Punkt in der Syntaxdatei aus, an dem sich der Cursor befindet, alle nachstehenden Befehlen zur Ausführung zu bringen. Die Befehle können schneller durch die Benutzung der Schaltfläche ▶ ausgeführt werden. Klickt man auf diese Schaltfläche (sie befindet sich in der Symbolleiste unter der Menüleiste) so werden die markierten Befehlsfolgen ausgeführt bzw. diejenige Befehlsfolge, in der sich der Cursor befindet.

Befinden sich in einer Syntaxdatei mehrere Befehle, so empfiehlt es sich der besseren Übersichtlichkeit wegen, diese immer durch eine Leerzeile zu trennen (innerhalb einer Befehlsfolge dürfen sich allerdings keine Leerzeilen befinden).

2.4 Vorbereitung der Dateneingabe

Kehren wir zu dem einführenden Beispiel zurück, bei dem es um eine Erhebung zum Wählerverhalten ging. Welche Schritte sind jetzt erforderlich, bevor SPSS zur Datenanalyse eingesetzt werden kann?

Der Fragebogen

Wenn man sich dazu entschließt, die gewünschten Informationen per schriftlicher Befragung zu gewinnen, dann könnte der Fragebogen so aussehen, wie Abbildung 2.4 ausschnittsweise zeigt:

```
1.  Welcher Partei würden Sie Ihre Stimme geben, wenn am kommenden
    Sonntag Wahlen zum Deutschen Bundestag stattfinden würden?
    CDU/CSU  o     SPD  o     FDP  o     Die Grünen  o    Sonstige  o

2. Geschlecht     männlich  o        weiblich  o

3. Geburtsjahr    19.......

4. In welchem Bundesland leben Sie?
          ....................................................

5. Konfession     katholisch  o     evangelisch     o
                  sonstiges   o     konfessionslos  o

6. Welches ist Ihr letzter Bildungsabschluss?
   kein Abschluss  o      Hauptschule  o     Mittlere Reife  o
   Abitur          o      Akad. Examen o     Sonstiges       o
```

Abb. 2.4: Fragebogen

Sie erkennen, es gibt Fragen, die direkt durch Ankreuzen beantwortet werden können (z.B. Frage 1), es gibt solche, die die Eingabe eines Zahlenwertes erfordern (z.B. Frage 3), und es gibt Fragen, bei denen die befragte Person die Antwort selbst formulieren und hinschreiben muss (z.B. Frage 4). Im letzten Fall ist vor der Datenauswertung eine Klassifikation durch den Forscher notwendig. Man spricht hier von offenen Fragen, während die anderen geschlossene Fragen sind.

Die Stichprobe

Aus Zeit- und Kostengründen wird auf eine Totalerhebung (Auszählung der Grundgesamtheit aller Wahlberechtigten) verzichtet. Stattdessen erheben wir eine *reine Zufallsstichprobe*. Eine solche Stichprobe zeichnet sich dadurch aus, dass jedes Element der Grundgesamtheit (jede wahlberechtigte Person) die gleiche Chance hat, in die Stichprobe aufgenommen zu werden. Solche Zufallsstichproben sind tendenziell repräsentativ. Man kann eine solche Zufallsstichprobe in der Weise realisieren, dass man nach dem Zufallsprinzip aus dem bundesweiten Wählerverzeichnis z.B. 200 Adressen auswählt. Steht ein solches Verzeichnis nicht zur Verfügung, muss man sich andere Wege einfallen lassen, zum Beispiel die Nutzung eines Telefonverzeichnisses oder dergl.

Kodierung

Wenn nun die ausgefüllten Fragebögen zurückkommen, schließt sich ein sehr wichtiger Schritt an, die Kodierung der Daten. Hier geht es darum, so weit dies überhaupt möglich ist, den Ausprägungen der einzelnen Variablen numerische Werte zuzuordnen, die dann dem Rechner übergeben werden. Zu diesem Zweck erstellt man ein sog. *Kodebuch*, das schematisch so aussieht, wie ausschnittsweise Abbildung 2.5 zeigt:

Frage-Nr.	Variable	Name	Werte	Kode
	Nummer des Fragebogens	NR	001-200	siehe Werte
1	Partei	V01	CDU/CSU	1
			SPD	2
			FDP	3
			Die Grünen	4
			Sonstige	5
			keine Angabe	*9*
2	Geschlecht	V02	männlich	1
			weiblich	2
			keine Angabe	*9*
3	Geburtsjahr	V03	00 bis 94	siehe Werte
			keine Angabe	*99*
4	Bundesland	V04	Baden-Württem-berg	01
			Bayern	02
			Berlin	03
			Brandenburg	04
			usw.	
			keine Angabe	*99*
usw.				

Abb. 2.5: Kodebuch

Im Kodebuch wird also angegeben, wie die Informationen dem Programm SPSS übergeben werden sollen. Besonders wichtig ist dabei, dass Werte auch für den Fall vorgegeben werden, dass eine bestimmte Frage nicht beantwortet worden ist. Solche Werte werden *missing values* (fehlende Werte oder *Missing-Werte*) genannt. Für sie werden Kodezahlen bereitgestellt, die im realen Datenbestand nicht auftauchen können. Beispielsweise gibt es bei der Variablen „Partei“ (Frage 1) keine Ausprägung, der real die Ziffer 9 zugeordnet werden könnte. Deshalb verwenden wir diese Ziffer als Missing-Wert. Da wir die Bundesländer zweistellig kodieren müssen (Frage 4) – es gibt ja 16 davon –, haben wir als Missing-Wert hier 99 vereinbart. Bei der Frage nach dem Geburtsjahr ist die Wahl von 99 vielleicht problematisch – zumindest dann, wenn man davon ausgehen muss, dass in der Befragung auch Personen des Jahrgangs 1899 auftauchen könnten. Sollte damit zu rechnen sein, würde sich als Missing-Wert vielleicht -1 anbieten.

Bei Bedarf können auch weitere Informationen in das Kodebuch aufgenommen werden, wie zum Beispiel der Typ der jeweiligen Variablen (diskret oder stetig) oder ihre Skalenqualität (nominalskalierte Variable, ordinalskalierte Variable etc.). Zusätzlich haben wir in der dritten Spalte jeder Variablen einen Namen gegeben. Der Einfachheit halber haben wir die Namen V01, V02 usw. gewählt. Aber selbstverständlich hätten wir auch Begriffe wie z.B. „Partei“ als Variablennamen verwenden dürfen.

Auch die Nummer des Fragebogens haben wir als eigene Variable mit dem frei gewählten Namen NR in das Kodebuch aufgenommen – sogar an erster Stelle. Es empfiehlt sich nämlich, die ankommenden ausgefüllten Fragebogen zu nummerieren. Dies hat zum einen den Vorteil, dass bei späteren Datenkontrollen Eingabefehler leichter identifiziert werden können, zum anderen kann es mit ein Auswertungsziel sein, zum Beispiel zu untersuchen, ob sich früh ankommende Fragebögen (niedrige Nummern) typisch von spät ankommenden unterscheiden.

Datenmatrix

Wenn das Kodebuch erstellt ist, kann die sog. Datenmatrix aufgebaut werden. Hier geht es darum, anhand der ausgefüllten Fragebögen einerseits und des Kodebuchs andererseits eine Tabelle aufzubauen – sinnvollerweise benutzt man dazu kariertes Papier –, die ausschnittsweise so aussieht, wie es Abbildung 2.6 zeigt.

NR	Partei	Geschlecht	Jahrgang	Land	Konfession	Abschluss
0001	1	1	20	01	1	1
0002	1	1	22	01	1	2
0003	1	1	27	01	1	3
0004	1	1	33	01	9	9
usw.						

Abb. 2.6: Datenmatrix

Sie werden gleich erkennen, dass dies die geeignete Form ist, die Daten in den Rechner zu geben. Zuvor noch einige Anmerkungen zu dieser Datenmatrix: Jede Zeile repräsentiert eine der befragten Personen; wir sprechen von einem *Datensatz*. Jede Spalte repräsentiert eine *Variable*; wir sprechen in diesem Zusammenhang auch von einem *Feld*. Jede Zelle in der Matrix (Kreuzungsbereich einer Spalte mit einer Zeile) repräsentiert einen *Wert*.

2.5 Eingabetabelle füllen

Wir starten SPSS und gelangen zum Dateneingabebildschirm, den wir schon oben in Abbildung 2.2 vorgestellt haben. Zur Dateneingabe gehen wir folgendermaßen vor (wir haben im Folgenden auf die Eingabe der Fragebogennummer verzichtet):

1. Klicken Sie die erste Zelle der angebotenen Tabelle an (in der Kopfzeile steht „var“, in der Vorspalte steht die Ziffer 1).
2. Schreiben Sie 1 (für CDU/CSU).
3. Drücken Sie die Return-Taste.

Es dauert jetzt einen kurzen Moment, bis SPSS den Wert 1,00 erscheinen lässt. Zugleich wird der Kopfzeilenbegriff von „var“ in „var00001“ geändert. SPSS vergibt also automatisch für die erste Variable (das war die bevorzugte politische Partei) den Namen „var00001“ und wandelt die eingegebene Zahl in eine Dezimalzahl mit zwei Nachkommastellen bei Unterdrückung der führenden Nullen um.

Wenn Ihnen die automatische Namensvergabe oder die Art der Zahlenpräsentation nicht gefällt, wechseln Sie durch Anklicken am unteren Bildrand zur *Variablenansicht* (siehe Abbildung 2.7).

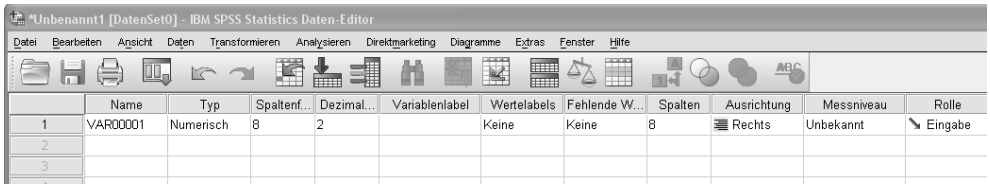


Abb. 2.7: Variablenansicht

Hier sind alle notwendigen Möglichkeiten geboten, die Darstellung Ihrer Daten zu verändern.

Klicken Sie zum Beispiel bei Ihrer ersten und bisher einzigen Variablen var00001 auf das Stichwort NUMERISCH bei TYP und dann auf die am rechten Zellenrand erscheinende Schaltfläche, werden Ihnen andere Typen angeboten, aus denen Sie auswählen können (siehe Abbildung 2.8). Zudem können Sie hier z.B. die Zahl der gewünschten Dezimalstellen einstellen (ersatzweise können Sie dies auch in der Spalte DEZIMALSTELLEN erledigen).

Abb. 2.8: Typfestlegung

Der Abbildung 2.8 können Sie entnehmen, dass die erste Variable als *numerische Variable* definiert ist (das ist in Ordnung so), und dass sie 8 Stellen mit 2 Dezimalstellen belegt. Wenn Sie keine Dezimalstellen wünschen (und in der Tat sind sie an dieser Stelle ja entbehrlich), klicken Sie die Ziffer 2 im Feld DEZIMALSTELLEN: an, löschen Sie diese Ziffer und ersetzen Sie sie durch eine Null. Klicken Sie dann die Schaltfläche OK an.

Klicken Sie in der Variablenansicht bei Ihrer ersten Variablen auf die Spalte BREITE, können Sie die *Spaltenbreite* der Tabelle in der Datenansicht verändern. Klicken Sie auf die Spalte AUSRICHTUNG, können Sie mit der am rechten Zellenrand erscheinenden Schaltfläche die Ausrichtung Ihrer Informationen in den Tabellenzellen der Datenansicht verändern.

Im Feld VARIABLENLABEL können Sie Ihrer ersten Variablen ein Label zuweisen, z.B. „bevorzugte politische Partei“. Dies ist mit dem großen Vorzug verbunden, dass bei späteren Ergebnisausgaben dieses Label mitgeführt wird und nicht der eigentlich nichtssagende Name VAR00001. Dies erleichtert ganz wesentlich das „Lesen“ der Ergebnisse.

Entsprechendes gilt für das Feld WERTELABELS, das derzeit mit „keine“ belegt ist. Klicken Sie den Schalter in diesem Feld an, öffnet sich das Fenster der Abbildung 2.9:

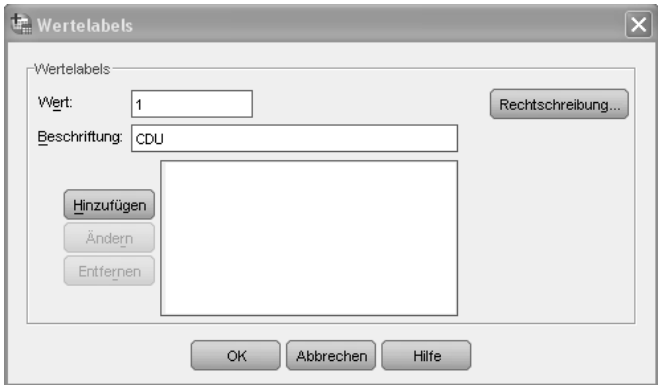


Abb. 2.9: Vereinbarung von Wertelabels

Geben Sie hier bei WERT: die Zahl 1 ein, bei BESCHRIFTUNG: den Text „CDU“ und klicken auf HINZUFÜGEN (und fahren Sie fort mit 2 für SPD usw.).

Ist die Label-Vereinbarung beendet, klicken Sie OK an. In entsprechender Weise können Sie auch für die anderen Variablen Werte-Labels vergeben. Zugegebenermaßen ist das etwas aufwändig, aber dieser Aufwand lohnt sich, weil die späteren Ergebnisausgaben viel leichter und angenehmer lesbar sein werden.

Kehren Sie zur *Datenansicht* zurück, erkennen Sie, dass die Veränderungen, die Sie vorgenommen haben (also etwa das Streichen der Dezimalstellen) von SPSS aufgenommen wurden.

Die weiteren Schritte der Dateneingabe sind die folgenden:

4. Geben Sie unterhalb der 1 erneut eine 1 ein.
5. Drücken Sie Return.
6. Geben Sie erneut die 1 ein.
7. Drücken Sie Return.
8. usw.

Jetzt haben wir die Werte der ersten Spalte der Datenmatrix der Abbildung 2.6 eingegeben, allerdings ohne die führenden Nullen, die ja auch entbehrlich sind.

Wechseln Sie jetzt in die zweite Spalte der Datenansicht, und geben Sie in der ersten Zelle die 1 (für „männlich“) ein. Auch hier produziert SPSS jetzt den Wert 1,00, wenn Sie die Return-Taste drücken. Auch diesen Wert wandeln wir (nach Anklicken des Blattes Variablenansicht) in eine Zahl ohne Nachkommastellen um, und wir bezeichnen die zweite Spalte (die zweite Variable) mit „Sex“, wobei genauso vorzugehen ist, wie es oben beschrieben wurde (als Variablenlabel können wir den Begriff „Geschlecht“ verwenden

Zusätzlich geben wir für diese Variable an, dass als Missing-Wert (*fehlender Wert*) die 9 vereinbart wurde. Dies ist deshalb wichtig, damit SPSS später bei den statistischen Auswertungen diese speziellen Fälle von den anderen, den echten Angaben, unterscheiden kann. Zum Vereinbaren der Missing-Werte klicken Sie in der Variablenansicht auf KEIN bei FEHLENDE WERTE und danach auf die am rechten Zellenrand erscheinende Schaltfläche. Dadurch öffnet sich das Fenster der Abbildung 2.10.

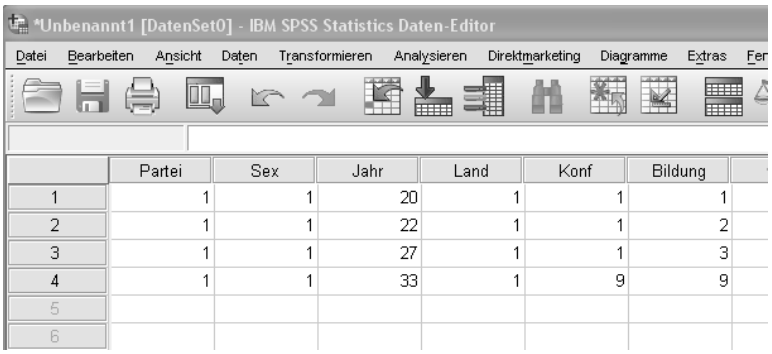
Abb. 2.10: Vereinbarung fehlender Werte

Im Fenster der Abbildung 2.10 sind die folgenden Schritte erforderlich:

1. Klicken Sie bei EINZELNE FEHLENDE WERTE an (Sie sehen übrigens, dass man auch ganze Zahlenbereiche als fehlende Werte definieren kann).
2. Geben Sie in das erste Listenfeld die Zahl 9 ein (hier gibt es zwei weitere Listenfelder, so dass bei Bedarf auch weitere Zahlen als fehlende Werte vereinbart werden könnten).
3. Klicken Sie auf die Schaltfläche OK.

Sie sollten in der Variablenansicht auch das *Messniveau* (*Skalenqualität*) Ihrer Variablen angeben, was aber auf die Auswertung der Daten mit den meisten (Standard-)Prozeduren keinen Einfluss hat, für einige Verfahren allerdings doch.

In der Datenansicht geben Sie nun auch die anderen Werte der Datenmatrix ein, bis Ihre Eingabetabelle so aussieht, wie es Abbildung 2.11 zeigt.



	Partei	Sex	Jahr	Land	Konf	Bildung
1	1	1	20	1	1	1
2	1	1	22	1	1	2
3	1	1	27	1	1	3
4	1	1	33	1	9	9
5						
6						

Abb. 2.11: Beispieldatenbestand

Führen Sie beispielsweise den Mauszeiger genau auf die Nahtstelle zwischen „Partei“ und „Sex“ in der Kopfzeile der Eingabetabelle. Er ändert seine Gestalt, und Sie können jetzt die *Spaltenbreite* verändern, wenn Sie die Maus bei gedrückter linker Maustaste nach links oder rechts bewegen.

2.6 Speichern und Öffnen

Nach der ersten Informationseingabe, also vor allen eventuellen Korrekturen, und bevor Sie mit den statistischen Auswertungen der Daten beginnen, sollten Sie Ihren Datenbestand speichern. Zuständig ist das Menü DATEI/SPEICHERN UNTER... Sie erhalten das aus anderen windowsgestützten Anwendungen bekannte Dialogfenster. In diesem Dialogfenster geben Sie an, wohin gespeichert werden soll (Bereich SPEICHERN IN:). Zudem geben Sie der zu speichernden Datei im Bereich DATEINAME: einen Namen und klicken dann die Schaltfläche SPEICHERN an. SPSS versieht Ihre Datei mit der Typenkennung .SAV. Wenn Sie im Nachhinein Ihre Daten ändern, etwa, indem Sie Eingabefehler beseitigen, genügt zum erneuten Speichern die Menüposition DATEI/ SPEICHERN. Die erste Version Ihrer Datei wird dann automatisch durch die veränderte Version ersetzt.

Wollen Sie nach einer Unterbrechung Ihrer Arbeit mit SPSS Ihren Datenbestand wieder auf dem Bildschirm erscheinen lassen, wählen Sie nach dem Start von SPSS die Menüposition DATEI/ÖFFNEN/DATEN... Sie erhalten dann ein Dialogfenster, in dem Sie Ihre Datei auswählen und mit der Schaltfläche ÖFFNEN öffnen können.

2.7 Umkodierungen

Nicht selten ist es erforderlich, Daten, die schon eingegeben wurden, neu zu kodieren. Das Beispiel der Variablen "Bundesland" mag dies verdeutlichen. Wir hatten zunächst für die einzelnen Bundesländer die Kodezahlen 01 bis 16 vereinbart (Alter Code). Im Nachhinein kommen wir auf die Idee, bestimmte Auswertungen so anzulegen, dass zwischen alten und neuen Bundesländern unterschieden werden kann. Es bietet sich deshalb beispielsweise an, allen alten Bundesländern die Ziffer 1 zuzuweisen, allen neuen Bundesländern die Ziffer 2. Die Kodezahlen 04 (Brandenburg), 08 (Mecklenburg-Vorpommern), 13 (Sachsen), 14 (Sachsen-Anhalt) und 16 (Thüringen) werden also zur 2, die anderen zur 1 umgewandelt. Nur wenn in dieser Weise umkodiert wird, kann auf einfache Art zwischen alten und neuen Bundesländern unterschieden werden.

Es ist dabei zweckmäßig, für die neuen Kodezahlen eine neue Variable zu vereinbaren, damit auch die alte Sechzehner-Unterteilung für detailliertere länderbezogene Auswertungen erhalten bleibt (es wäre auch denkbar, so umzukodieren, dass die neuen Kodezahlen 1 und 2 die alten 1 bis 16 ersetzen; dies ist aber nur selten empfehlenswert, weil dann die ursprünglichen Detailinformationen selbstverständlich verloren gehen und nicht mehr genutzt werden können).

SPSS erledigt dieses gewünschte Umkodieren der Bundesländer automatisch, wenn Sie wie folgt vorgehen:

Wählen Sie im Menü TRANSFORMIEREN die Position UMKODIEREN und dort den Befehl IN ANDERE VARIABLEN.... Sie gelangen zum Dialogfenster der Abbildung 2.12:

Abb. 2.12: Menü TRANSFORMIEREN/UMKODIEREN/IN ANDERE VARIABLEN...

1. Im Fenster der Abbildung 2.12 klicken Sie in der linken Liste auf die Variable „Land“.
2. Klicken Sie auf die Schaltfläche mit dem nach rechts zeigenden Pfeil. Damit wird diese Variable in den mittleren Bereich des Dialogfensters übernommen.
3. Im rechten Bereich (AUSGABEVARIABLE) geben Sie bei NAME: einen Namen für die neue Variable ein, z.B. "Landcode". Wenn Sie es wünschen, können Sie bei LABEL: der neuen Variablen noch ein aussagekräftigeres Label zuweisen.
4. Klicken Sie dann die Schaltfläche ÄNDERN an.
5. Klicken Sie danach auf die Schaltfläche ALTE UND NEUE WERTE ... Sie gelangen damit zu einem weiteren Dialogfenster (siehe Abbildung 2.13).

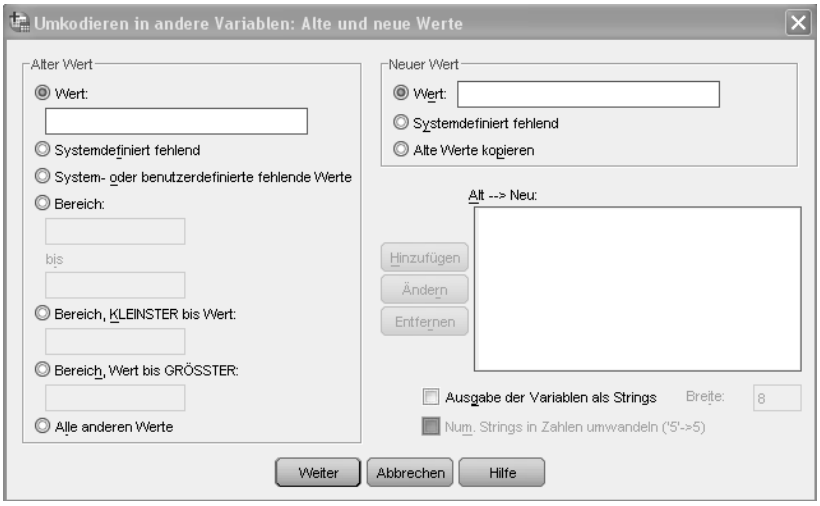


Abbildung 2.13 : Menü TRANSFORMIEREN/UMKODIEREN/IN ANDERE VARIABLEN..., Schaltfläche ALTE UND NEUE WERTE...

Im Fenster der Abbildung 2.13 sind die neuen Kodezahlen einzugeben, die der Variablen „Landcode“ zugewiesen werden sollen:

6. Geben Sie im Bereich ALTER WERT im Listenfeld bei WERT: die Zahl 1 (für Baden-Württemberg) ein.
7. Geben Sie im Bereich NEUER WERT im Listenfeld bei WERT: ebenfalls die Zahl 1 ein (altes Bundesland).
8. Klicken Sie die Schaltfläche HINZUFÜGEN an.
9. Geben Sie im Bereich ALTER WERT im Listenfeld bei WERT: die Zahl 2 (für Bayern) ein.
10. Geben Sie im Bereich NEUER WERT im Listenfeld bei WERT: ebenfalls die Zahl 1 ein (altes Bundesland).
11. Klicken Sie wieder die Schaltfläche HINZUFÜGEN an.

Genauso verfahren Sie auch für das Land Berlin. Beim vierten Land (Brandenburg) muss im Bereich NEUER WERT im Listenfeld bei WERT: die Zahl 2 (neues Bundesland) eingegeben werden. Fahren Sie in dieser Weise fort, bis Sie alle 16 Bundesländer umkodiert haben. Weisen Sie schließlich auch dem alten Wert 99 (fehlender Wert der Variablen "Land") den neuen Wert 9 (fehlender Wert der Variablen "Landcode") zu.

Klicken Sie danach WEITER und dann im Dialogfenster der Abbildung 2.12 OK an, werden von SPSS in einer zusätzlichen Spalte der Eingabetabelle, die automatisch mit "Landcode" überschrieben wird, nur Einsen und Zweien (für alte und neue Bundesländer) ausgegeben und ggf. die Ziffer 9 als fehlender Wert.

2.8 Umrechnen von Daten

Manchmal ist es auch erforderlich, dass aus numerischen Ausgangsinformationen neue Werte erzeugt werden sollen.

Zum Beispiel sind Sie daran interessiert, aus den Geburtsjahrangaben des Ausgangsbeispiels das Alter der befragten Personen zu errechnen, um auf dieser Grundlage später eine Altersverteilung erstellen oder das Durchschnittsalter errechnen zu können.

Wenn solche Umrechnungen erforderlich sind, wählen Sie dazu die Menüposition TRANSFORMIEREN/VARIABLE BERECHNEN... (siehe Abbildung 2.14).

1. Geben Sie unter ZIELVARIABLE: einen neuen Namen ein (z.B. „Alter“).
2. Klicken Sie in das Feld im Bereich NUMERISCHER AUSDRUCK:
3. Geben Sie ein: 97- (Anmerkung: Die Ausgangsdaten stammen aus dem Jahr 97)
4. Klicken Sie auf „Jahr“ in der linken Variablenliste.
5. Klicken Sie auf die Schaltfläche mit dem nach rechts zeigenden Pfeil.
   Im Rechenbereich steht jetzt: 97 – Jahr.
6. Klicken Sie dann OK an.

SPSS führt nach dem Anklicken von OK die gewünschte Berechnung durch und gibt die Rechenergebnisse in einer weiteren Spalte der Dateneingabetabelle aus.

Dass man im Dialogfenster der Abbildung 2.14 auch kompliziertere mathematische Funktionen verwenden kann, oder dass man Bedingungen formulieren kann für die Ausführung bestimmter Rechenoperationen, zeigt der Blick auf den rechten unteren Bereich des Dialogfensters.

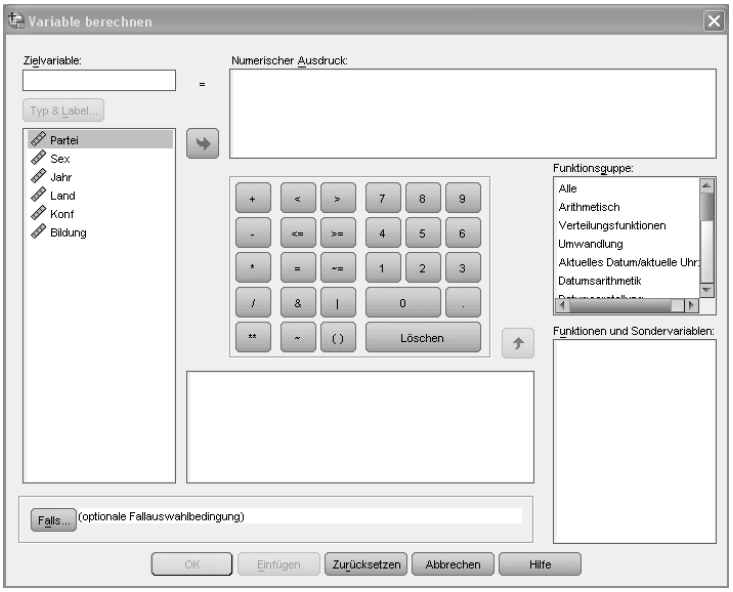


Abb. 2.14: Menü TRANSFORMIEREN/VARIABLE BERECHNEN...

Anzumerken ist in diesem Zusammenhang, dass es manchmal erwünscht ist, in der Ausgangstabelle die Werte-Labels anstelle der Werte selbst sichtbar zu machen. Zuständig dafür ist das Menü ANSICHT/WERTELABELS.

3 Häufigkeitsverteilungen

3.1 Zielsetzungen und statistische Methoden

Nehmen wir an, wir hätten 203 wahlberechtigte Bundesbürger danach gefragt, welche Partei sie wählen würden (zudem wurde nach den Variablen Geschlecht, Geburtsjahrgang, Bundesland, Konfession und zuletzt erreichter Bildungsabschluss gefragt). Beispielsweise beinhalten zwar die Angaben zur bevorzugten politischen Partei die gesamten zur Verfügung stehenden Informationen zu dieser Variablen, aber diese Informationen werden nicht ersichtlich. In der Menge der Daten gehen sie gewissermaßen unter. Es ist deshalb erforderlich, eine Zusammenfassung in der Weise vorzunehmen, dass nur die einzelnen Parteien genannt werden, um dann daneben zu schreiben, wie häufig diese Ausprägungen der Variablen „bevorzugte politische Partei“ aufgetreten sind. Bezieht man diese *absoluten Häufigkeiten* auf die Gesamtzahl der Befragungen, gelangt man zu *relativen Häufigkeiten*. Diese können als Prozentangaben (z.B. 39%) oder als Dezimalzahlen (z.B. 0,39) angegeben werden. Sie werden aus den absoluten Häufigkeiten per Dreisatzrechnung berechnet.

Bei einer stetigen Variablen ist zunächst eine *Klassifizierung* erforderlich. Diese empfiehlt sich allerdings auch dann, wenn die betrachtete Untersuchungsvariable vom diskreten Typ ist, aber sehr viele Ausprägungen aufweist. Beispielsweise dürfte es bei der Variablen „Alter“ zweckmäßig sein, Klassen von jeweils 10 oder von jeweils 5 Jahren zu bilden – je nachdem, wie detailliert die sich ergebende Häufigkeitsverteilung sein soll –, also z.B. „21 bis 25 Jahre“, „26 bis 30 Jahre“ usw. Bei einer „echten“ stetigen Variablen, zum Beispiel „Körpergröße befragter Personen“, ist zu berücksichtigen, dass die Klassen „nahtlos“ aneinander stoßen. Deshalb muss in diesem Fall Sorge dafür getragen werden, dass erfragte Werte, die genau auf eine Klassengrenze fallen, eindeutig zuordenbar sind. Man bildet in einem solchen Fall Klassen in der folgenden Weise: „über 150 bis 160 cm“, „über 160 bis 170 cm“ usw. Damit ist klar, dass eine Person, die genau 160 cm groß ist, in die erste und nicht in die zweite Klasse gehört.

3.2 Diskrete Verteilungen

Wenn man die Häufigkeiten, mit denen die einzelnen Parteien in der Befragung genannt wurden, den Ausprägungen dieser Variable zuordnen will, musste man im Vorcomputerzeitalter per Strichliste die interessierenden Häufigkeiten auszählen – bei einem umfangreichen Datenbestand sicherlich keine sehr angenehme Arbeit. Mit SPSS geht dieses Gruppieren, wie es bei einer *diskreten Variablen* angemessen ist, viel einfacher:

1. Wählen Sie Menü DATEI/ÖFFNEN/DATEN...
2. Geben Sie Ihre Ausgangsdaten ein.
3. Wählen Sie das Menü ANALYSIEREN.

Es erscheint jetzt das Menü der Abbildung 3.1.

Hinweis:

Diese Daten, die auch in weiteren Beispielen genutzt werden, werden in der Datei „B00.sav“ bereitgestellt.

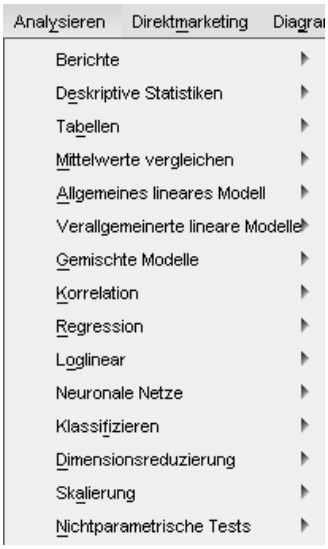


Abb. 3.1: Menü ANALYSIEREN (Ausschnitt)

Wählen Sie DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN..., gelangen Sie zum Dialogfenster der Abbildung 3.2.

Abb. 3.2: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN...

Hier ist nun wie folgt zu verfahren:

1. Klicken Sie in der linken Variablenliste „Partei (Partei)“ an.
2. Klicken Sie auf die Schaltfläche mit dem nach rechts zeigenden Pfeil (Übertragen der Variablen „Partei“).
3. Klicken Sie unten links auf die Markierungsfläche bei HÄUFIGKEITSTABELLE ANZEIGEN, so dass dort ein Häkchen auftaucht (eventuell erübrigt sich das).
4. Klicken Sie OK an.

Jetzt erscheint in einem eigenen Fenster, das mit „Ausgabe1 SPSS Viewer“ überschrieben ist, Ihre erste Häufigkeitsverteilung:

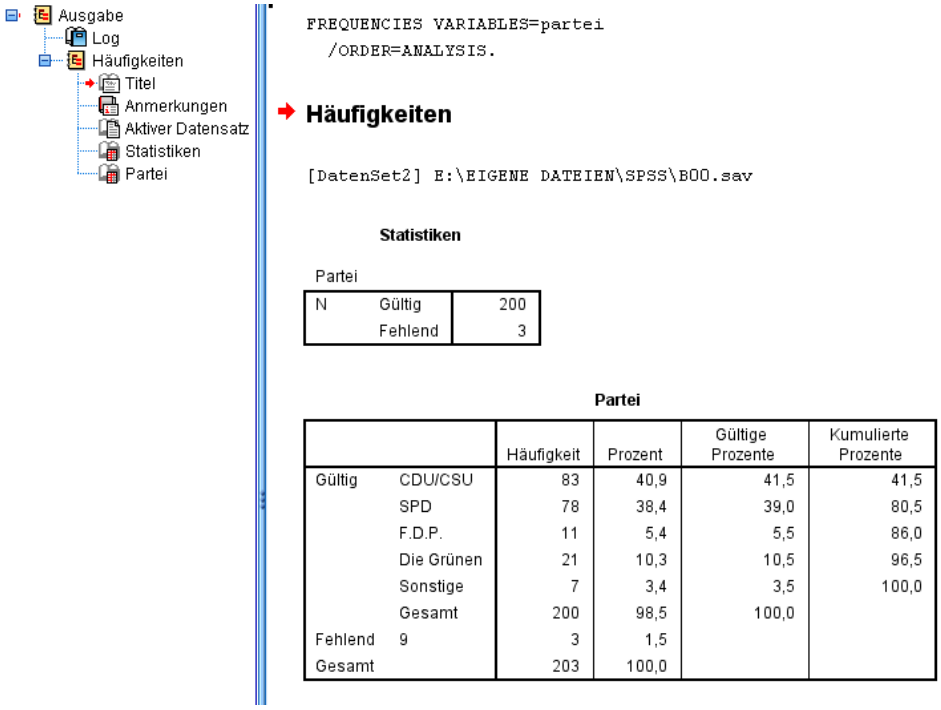


Abb. 3.3: Parteienverteilung

SPSS bietet Ihnen zwei Tabellen an. In der ersten, überschrieben mit „Statistiken“, wird angegeben, wie viele Fälle betrachtet werden (n=203) und wie viele davon gültig sind (200). Dies bedeutet, dass bei der Frage nach der bevorzugten politischen Partei drei der befragten Personen keine Antwort gegeben haben.

Die zweite von SPSS erzeugte Tabelle ist überschrieben mit „Partei“ (das ist das Label, das wir für diese Variable vergeben hatten), und in der Vorspalte stehen die Parteinamen, die als Wertelables vergeben wurden. In der nächsten Spalte tauchen unter der Überschrift HÄUFIGKEIT die absoluten Häufigkeiten auf. Unter der Überschrift PROZENT sind auch die relativen Häufigkeiten (in %) angegeben. Beim Stichwort GESAMT steht die Gesamtzahl der Befragungen (n=203), daneben, unter der Spalte der relativen Häufigkeiten, erwartungsgemäß der Wert 100,0.

In der vorletzten Spalte mit der Überschrift GÜLTIGE PROZENTE ist von SPSS eine zusätzliche Berechnung durchgeführt worden. Hier werden die relativen Häufigkeiten ausgerechnet, die sich ergeben, wenn die drei Fälle, die die Antwort auf diese Frage nach der bevorzugten Partei verweigerten (oder die vergessen haben, anzukreuzen; diese drei Fälle werden unten beim Stichwort FEHLEND gesondert ausgegeben), nicht mitgerechnet werden. Bezogen auf die 200 „echten“ Antworten, machen also die CDU/CSU-Anhänger 41,5% aus.

In der letzten Spalte schließlich werden die Werte der vorletzten Spalte kumuliert. Von einer *Kumulation* spricht man dann, wenn der jeweilige Prozentwert einer Ausprägung zu den schon erreichten Prozentwerten hinzuaddiert wird. Diese Kumulationen sind ganz inte-

ressant bei metrischen Daten, wie in Abschnitt 3.3 gezeigt wird. Da die Variable „Partei“ nicht metrisch ist, sondern nur nominalskaliert, brauchen wir diese letzte Spalte hier nicht weiter zu berücksichtigen.

Hinweis:

Wenn Sie eine derartige Ausgabe vom Layout her verändern wollen (etwa Veränderung der Schriftart; Änderung der Anzahl der Dezimalstellen oder z.B. Löschen der letzten Spalte, da hier die kumulierten Werte uninteressant sind), müssen Sie einen Doppelklick auf die entsprechende Tabelle ausführen, was in den Editiermodus führt. Einzelheiten dazu brauchen hier aber nicht besprochen zu werden – probieren Sie einfach etwas aus, indem Sie im Editiermodus eine Tabellenzelle markieren und dann das Menü Format/Zelleneigenschaften aufrufen. Den Editiermodus verlassen Sie wieder, indem Sie eine freie Stelle außerhalb der Tabelle anklicken.

Wenn Sie diese Ausgabetabelle speichern wollen, wählen Sie die Menüposition DATEI/ SPEICHERN UNTER.... Geben Sie im sich öffnenden Dialogfenster Ihrer Datei einen vernünftigen Namen, und geben Sie den Pfad (Laufwerk, Unterverzeichnis) zum Speichern an. SPSS speichert Ihre Datei mit der Typenkennung .SPV.

Zusätzlich wird von SPSS im oberen Teil der Ausgabe die Syntax ausgegeben. Weiterhin ist zu erkennen, dass im linken Bildschirmbereich eine Art Inhaltsverzeichnis präsentiert wird. Dies ist deshalb ganz günstig, weil beim weiteren Arbeiten mit SPSS die Ausgabedatei ständig verlängert wird (es sei denn, Sie würden sie zwischenzeitlich schließen; weitere Ausgaben würden dann in einer neuen SPV-Datei präsentiert), so dass dieses Inhaltsverzeichnis benutzt werden kann, um zu navigieren. Zudem erleichtert es beispielsweise das Löschen von Ausgaben, die Sie nicht mehr benötigen, weil Sie sie z.B. nur probeweise anschauen wollen. Zum Löschen klicken Sie im Inhaltsverzeichnis auf die entsprechenden Elemente und dann die Entf-Taste. Klicken Sie beispielsweise auf den gelben Schalter neben „Ausgabe“, werden alle Elemente markiert und können dann insgesamt gelöscht werden.

Hinweis:
Wenn Sie im Nachhinein Ihren Datenbestand verändern – etwa um Fehler der Dateneingabe zu korrigieren – müssen Sie Auswertungen, die Sie schon vorgenommen haben, erneut in Gang setzen. Spätestens dann ist es sinnvoll, fehlerbehaftete Ausgaben zu löschen.

Wenn Sie sich die oben vorgestellte Prozedur zur Erstellung der Häufigkeitsverteilung noch einmal anschauen, dann erkennen Sie übrigens, dass Sie mehrere Häufigkeitsverteilungen zugleich hätten erstellen können, wenn Sie die Schritte 4. und 5. für weitere Variablen wiederholt hätten. Für alle diskreten Variablen (Partei, Sex, Land, Konf und Bildung) hätte man so „in einem Aufwasch“ die Häufigkeitsverteilungen erstellen können.

Jetzt haben wir einen informativen, zusammenfassenden Überblick über die Antworten auf die erste Frage des Fragebogens, der viel aussagekräftiger ist, als die Betrachtung der 203 Ausgangswerte. Dies ist die wesentliche Aufgabe der *deskriptiven Statistik*: Zentrale Informationen in zusammenfassender Form sichtbar zu machen. Noch deutlicher würden die Befunde, wenn wir ergänzend eine grafische Präsentation der erzeugten Häufigkeitsverteilung anbieten würden. Dazu verweisen wir aber auf Kapitel 4.

3.3 Stetige Verteilungen

Hat man es mit einer *stetigen Variablen* zu tun, muss man vor der Häufigkeitsauszählung zunächst Klassen bilden. In unserem Beispiel ist dies für die Variable „Jahr“ (Geburtsjahrgang) der Fall, bzw. für die Variable „Alter“, die wir, wie in Abschnitt 2.7 beschrieben wurde, erzeugt haben. Bei der Variablen „Alter“ sind Werte zwischen 18 und maximal 79 zu erwarten, die man sinnvollerweise etwa jahrzehntweise zusammenfasst (es sei daran erinnert, dass die Daten aus dem Jahr 1997 stammen).

Der einfachste Weg, eine solche klassifizierte Häufigkeitsverteilung zu erstellen, besteht darin, die Variable „Alter“ zunächst umzukodieren. Diese Umkodierung erreichen Sie so, wie es weiter oben am Beispiel der Bundesländer beschrieben wurde (siehe Abschnitt 2.6).

Hier noch einmal das Dialogfenster zum Umkodieren selbst:

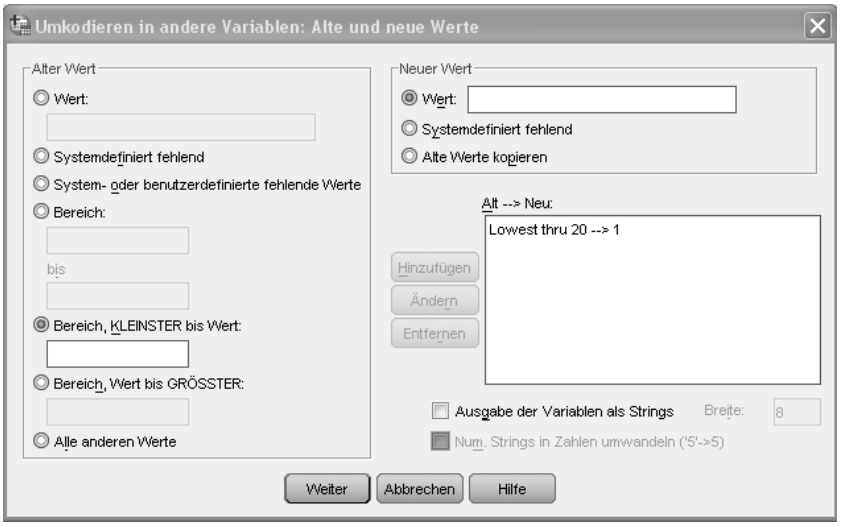


Abb. 3.4: Schaltfläche ALTE UND NEUE WERTE...

In diesem Fenster kodieren Sie die Altersangaben, wie es in Abschnitt 2.6 ausführlich beschrieben wurde. Dabei wählen Sie für die erste Altersklasse die Kategorie BEREICH, KLEINSTER BIS WERT:, geben darunter die Zahl 20 ein, danach rechts oben den Wert 1, gefolgt von einem Klick auf HINZUFÜGEN. Für die zweite Altersklasse wählen Sie rechts BEREICH: und geben darunter die Zahlen 20 und 30 ein, rechts oben bei NEUER WERT, WERT: den Wert 2, wiederum gefolgt von einem Klick auf HINZUFÜGEN usw. Wenn alle Altersklassen definiert sind, klicken Sie auf WEITER und dann auf OK.

Hinweis:

Sie erkennen, dass die Altersklassen sich jeweils in einem Wert überschneiden. SPSS ordnet aber die Altersangaben so zu, dass z.B. der Wert 20 in die erste, nicht in die zweite Klasse einsortiert wird. Die Altersklassen lauten also intern: 1. Klasse: „bis 20“; 2. Klasse: „über 20 bis 30“ usw. Diese Texte könnten Sie der Klarheit halber auch als Wertelabels der neuen Variablen „Altcode“ in der Variablenansicht zuweisen.

Wenn Sie nun eine Häufigkeitstabelle für diese Variable „Altcode“ erstellen (Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN...) erhalten Sie die Tabelle der Abbildung 3.5.

Alter klassifiziert

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 20	5	2,5	2,5	2,5
	über 20 bis 30	43	21,2	21,2	23,6
	über 30 bis 40	59	29,1	29,1	52,7
	über 40 bis 50	29	14,3	14,3	67,0
	über 50 bis 60	35	17,2	17,2	84,2
	über 60 bis 70	20	9,9	9,9	94,1
	über 70 bis 80	12	5,9	5,9	100,0
	Gesamt	203	100,0	100,0	

Abb. 3.5: Altersverteilung

Sie sehen in Abbildung 3.5 eine Häufigkeitstabelle, die im Aufbau der entspricht, die wir im vorangegangenen Abschnitt erzeugt haben. Auch hier werden absolute und relative Häufigkeiten ausgegeben. Die angepassten relativen Häufigkeiten entsprechen in diesem Beispiel den „normalen“ relativen Häufigkeiten, weil hier keine Missing-Werte auftraten.

Da es sich hier um eine metrische Variable handelt, sind nun die kumulierten Prozentwerte, im Gegensatz zum Beispiel zuvor, interpretierbar. Mit Blick auf die letzte Spalte der ausgegebenen Tabelle können Sie beispielsweise feststellen, dass 67% der Befragten 50 Jahre oder jünger sind.

Auch hier könnte mit einer zusätzlichen grafischen Präsentation noch mehr Anschaulichkeit erreicht werden. Dazu verweisen wir wieder auf das vierte Kapitel.

3.4 Zählen

Manchmal ist es für bestimmte Zwecke ganz sinnvoll, auszuzählen, wie oft bestimmte Werte bei einer einzigen Person auftauchen. Stellen Sie sich beispielsweise vor, es würden die Schulzensuren für fünf Fächer bei mehreren zufällig ausgewählten Schülern erfasst, so dass sich etwa der Datenbestand der Abbildung 3.6 ergibt:

	deutsch	mathe	physik	chemie	sport
1	2	1	4	1	2
2	1	5	1	1	1
3	1	4	1	3	2
4	3	1	2	1	2
5	4	2	3	5	1
6	2	2	4	5	4
7					

Abb. 3.6: Schulzensuren für ausgewählte Schüler

Sie wollen jetzt wissen, wie viele Einsen bei den einzelnen Schülern auftauchen. In dem kleinen Beispielsdatenbestand ist das sofort zu sehen. Bei umfangreicheren Datenbeständen ist es aber ganz angenehm, SPSS diese Aufgabe zu übertragen. Gehen Sie dazu folgendermaßen vor:

1. Wählen Sie TRANSFORMIEREN/WERTE IN FÄLLEN ZÄHLEN...

Dies führt in das Dialogfenster der Abbildung 3.7

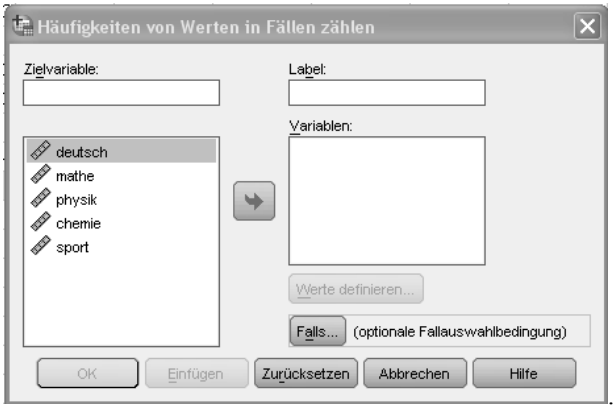


Abb. 3.7 : Menü TRANSFORMIEREN/WERTE IN FÄLLEN ZÄHLEN...

2. Im Fenster der Abbildung 3.7 geben Sie im Bereich ZIELVARIABLE: einen neuen Variablennamen ein (z.B. „Eins“), der Sie im Bereich LABEL: auch ein Label zuweisen können (z.B. „Zahl der Einsen“).
3. Übertragen Sie alle (alten) Variablen (die Zensuren) in den Bereich VARIABLEN:.
4. Klicken Sie auf die Schaltfläche WERTE DEFINIEREN...

Dies führt in das Fenster der Abbildung 3.8.

Abb. 3.8: Menü TRANSFORMIEREN/WERTE IN FÄLLEN ZÄHLEN..., Schaltfläche WERTE DEFINIEREN...

5. Geben Sie im Fenster der Abbildung 3.8 im Bereich WERT bei WERT: die Zahl 1 ein.

6. Klicken Sie auf HINZUFÜGEN.
7. Klicken Sie auf WEITER.
8. Klicken Sie im Fenster der Abbildung 3.7 auf OK.

SPSS berechnet jetzt, wie viele Einsen bei den einzelnen Schülern auftauchen und gibt das Zählergebnis in einer neuen Spalte der Ausgangstabelle aus, wie Abbildung 3.9 zeigt.

	deutsch	mathe	physik	chemie	sport	Eins
1	2	1	4	1	2	2,00
2	1	5	1	1	1	4,00
3	1	4	1	3	2	2,00
4	3	1	2	1	2	2,00
5	4	2	3	5	1	1,00
6	2	2	4	5	4	,00
7						

Abb. 3.9: Zählergebnis

Sie können mit der Schaltfläche FALLS... im Fenster 3.7 auch Bedingungen formulieren, wenn sich beispielsweise das Zählen nur auf bestimmte Fälle erstrecken soll. Auf diese Weise können Sie die Frage beantworten, wie viele Einsen diejenigen Schüler in den übrigen Fächern haben, die in Sport eine 5 haben.

4 Grafische Verteilungen

4.1 Zielsetzungen und statistische Methoden

Das wesentliche Ziel grafischer Darstellungen statistischer Daten besteht darin, Übersichtlichkeit zu gewinnen. Nach dem Motto „ein Bild sagt mehr als 1000 Worte“ können Sie davon ausgehen, dass eine grafische Darstellung gewissermaßen „auf einen Blick“ erkennen lässt, was eigentlich Sache ist. Wir wollen deshalb in diesem Kapitel zeigen, wie man eine univariate Häufigkeitsverteilung grafisch dargestellt werden kann. Die grafische Präsentation anderer Datenbestände wird in späteren Kapiteln behandelt.

Bei der Art der grafischen Darstellung ist es sinnvoll, zwischen nichtmetrischen und metrischen Variablen zu unterscheiden.

4.2 Nichtmetrische Daten – Kreisdiagramme

Hier betrachten wir zunächst das Beispiel der Parteienverteilung, das schon in Kapitel 3 vorgestellt wurde. Es handelt sich bei der Variablen „Partei“ bekanntlich um eine nominalskalierte Variable, also um eine nichtmetrische Variable. Für die grafische Präsentation der Verteilung einer solchen nominalskalierten Variablen eignet sich das Kreisdiagramm. Es ist allerdings erforderlich, in der Variablenansicht bei Messniveau „Nominal“ einzustellen

Partei

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	CDU/CSU	83	40,9	41,5	41,5
	SPD	78	38,4	39,0	80,5
	F.D.P.	11	5,4	5,5	86,0
	Die Grünen	21	10,3	10,5	96,5
	Sonstige	7	3,4	3,5	100,0
	Gesamt	200	98,5	100,0	
Fehlend	9	3	1,5		
Gesamt		203	100,0		

Abb. 4.1: Bevorzugte politische Partei

Um ein Kreisdiagramm mit SPSS zu erzeugen, gehen Sie folgendermaßen vor:

1. Starten Sie SPSS.
2. Öffnen Sie die Ausgangstabelle (B00.SAV) über DATEI/ÖFFNEN/DATEN...
3. Wählen Sie Menü DIAGRAMME/GRAFIKTAFEL-VORLAGENAUSWAHL..
4. Klicken Sie aufPartei und dann auf KREISDIAGRAMM DER HÄUFIGKEITEN

Sie gelangen zur Abbildung 4.2. Wenn Sie in dieser auf Ok klicken, erscheint das Kreisdiagramm im Ausgabefenster (siehe Abbildung 4.3).

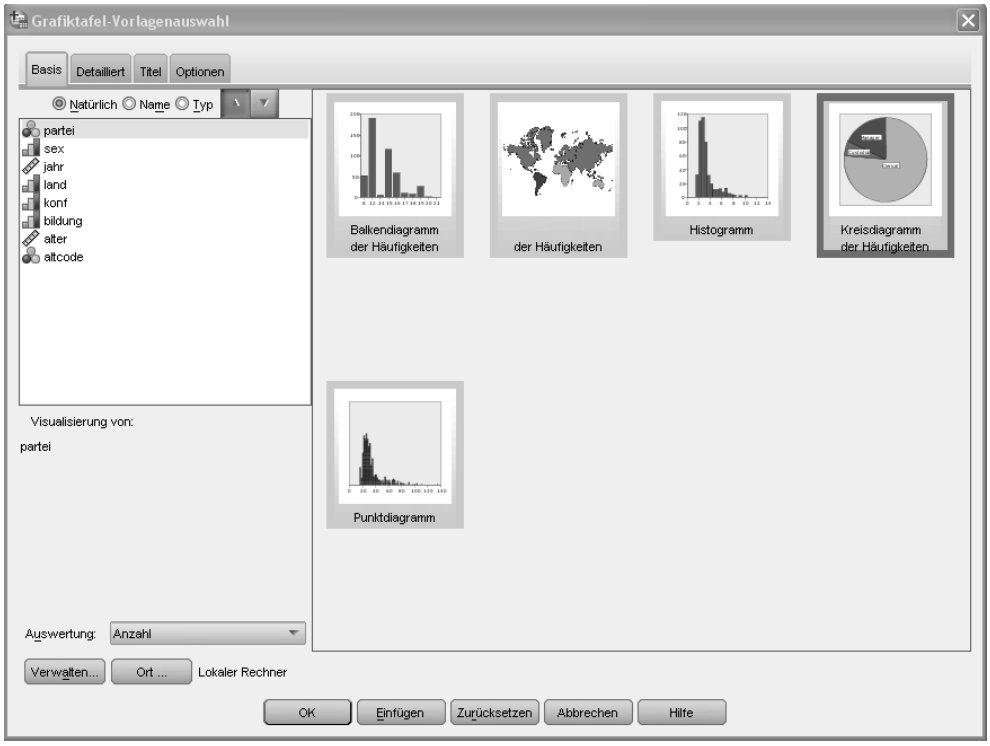


Abb. 4.2: Menü DIAGRAMME/GRAFIKTAFEL-VORLAGENAUSWAHL..

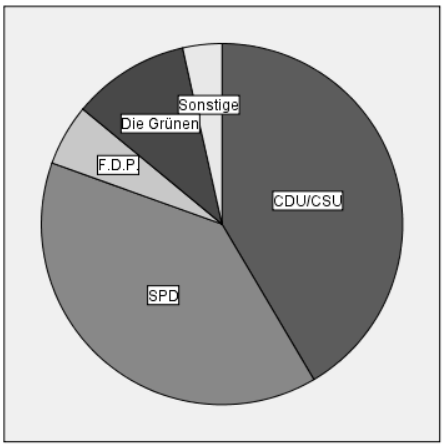


Abb. 4.3: Kreisdiagramm

Es steht zur Bearbeitung einer solchen Grafik eine Reihe von Möglichkeiten zur Verfügung, die hier aber nicht besprochen werden müssen. Sie können aber auch hier schon ein bisschen mit diesen Möglichkeiten experimentieren – nach einem Doppelklick auf das Diagramm –, um zu erkennen, was erreichbar ist.

Bei der grafischen Präsentation der Häufigkeitsverteilung einer ordinalskalierten Variablen (Rangordnungsdaten) verfährt man im Prinzip genauso wie bei der nominalskalierten Variablen, die wir eben betrachtet haben. Auch hier bietet sich ein Kreisdiagramm an.

4.3 Metrische Daten – Balkendiagramme

Betrachten Sie noch einmal die Variable „Alter“ unseres Ausgangsbeispiels, deren Häufigkeitsverteilung in Kapitel 3 schon erzeugt wurde. Bei einer solchen metrischen Variablen wird zur grafischen Präsentation in aller Regel ein Achsenkreuz verwendet. Auf der waagrechten Achse werden die Ausprägungen (oder Klassen) der interessierenden Variablen abgetragen, auf der senkrechten Achse die zugeordneten Häufigkeiten. Dabei ist es für das grafische Erscheinungsbild gleichgültig, ob von den absoluten oder von den relativen Häufigkeiten ausgegangen wird (diese Anmerkung gilt auch für den vorangegangenen Abschnitt).

Um die Altersverteilung grafisch darzustellen, gehen Sie von der Variablen „Altcode“ aus, also von den klassifizierten Altersangaben. Wie beim Kreisdiagramm des vorangegangenen Abschnitts öffnen Sie das Menü DIAGRAMME/GRAFIKTAFEL-VORLAGENAUSWAHL.. (siehe Abbildung 4.2), klicken jetzt aber auf „Altcode“ und dann auf BALKENDIAGRAMM DER HÄUFIGKEITEN. Mit OK gelangen Sie dann zu dem folgenden Diagramm:

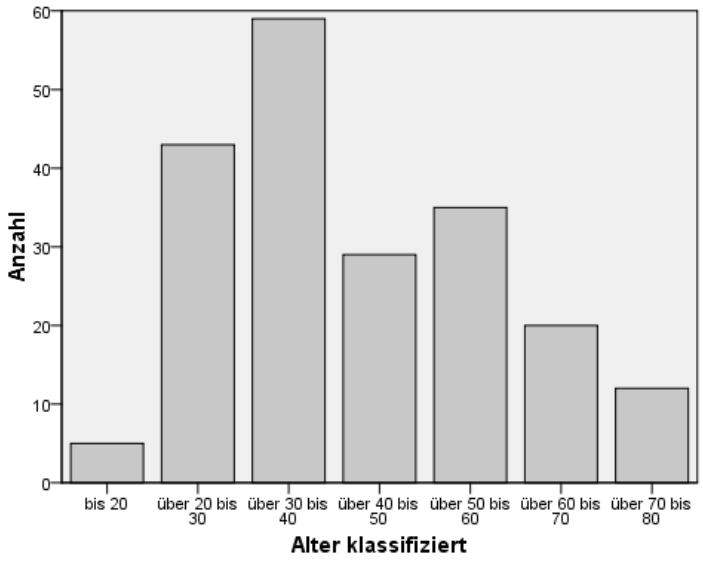


Abb. 4.4: Balkendiagramm

Zum Schluss dieses Kapitels ein wichtiger Hinweis: SPSS bietet eine fast unüberschaubare Zahl von verschiedenen Möglichkeiten der grafischen Präsentation statistischer Daten. Es würde viel zu weit führen, hier alle Einzelheiten zu besprechen. In späteren Kapiteln werden Ihnen einige dieser zusätzlichen Möglichkeiten begegnen.

5 Mittelwerte

5.1 Zielsetzungen und statistische Methoden

Mittelwerte gehören zu den univariaten Maßzahlen. Diese dienen dazu, wichtige Eigenschaften einer gegebenen Häufigkeitsverteilung bzw. des Datenbestandes, aus dem sie erzeugt wurde, zusammenfassend zu charakterisieren. Solche Eigenschaften sind in erster Linie die sog. *zentrale Tendenz*, d.h. die Mitte einer Verteilung, und ihre Streuung (siehe Kapitel 6).

Mittelwerte werden gern auch als *Durchschnittswerte* bezeichnet. Es gibt unterschiedliche Möglichkeiten, die Mitte eines gegebenen Datenbestandes zu bemessen. Die Aufgabe dieser statistischen Maßzahlen ist es, anzugeben, um welchen speziellen Merkmalswert herum sich die einzelnen Merkmalswerte einer Häufigkeitsverteilung konzentrieren.

Die wichtigsten dieser Maßzahlen sind das arithmetische Mittel, der häufigste Wert (Modus) und der Zentralwert (Median).

Arithmetisches Mittel

Wenn von Durchschnittsberechnungen gesprochen wird, meint man meist das *arithmetische Mittel*. Dieser Mittelwert ist so bekannt und gebräuchlich, dass oft mit dem Wort Durchschnitt genau diese Maßzahl gemeint ist. Das arithmetische Mittel ist ein *rechnerischer Mittelwert*, weil bei seiner Bestimmung „per Hand“ gerechnet werden muss. Es kommt in der Weise zustande, dass man alle Merkmalswerte zusammenzählt und diese Summe dann durch die Anzahl der Merkmalswerte dividiert. Geht man so vor, spricht man von einem *ungewogenen arithmetischen Mittel.*

Sie erkennen aus der angegebenen Rechenvorschrift übrigens, dass das arithmetische Mittel nur bei metrischen Daten berechnet werden kann. Bei Nominaldaten (z.B. bevorzugte politische Partei), aber auch bei Ordinaldaten, ist die erforderliche Summenbildung eine sinnlose mathematische Prozedur. Beachten Sie bitte in diesem Zusammenhang, dass SPSS auf Anforderung auch das arithmetische Mittel aus Nominal- oder Ordinaldaten, also eigentlich unzulässigerweise, berechnet. Es liegt an Ihnen, dem Benutzer, SPSS zu veranlassen, nur Sinnvolles zu tun.

Als Beispiel betrachten Sie bitte die folgenden fünf Angaben zur Kinderzahl befragter Ehepaare:

Ehepaar	Kinderzahl
Müller	0
Meier	1
Weber	1
Schmidt	2
Olbrich	4

Die durchschnittliche Kinderzahl erhalten Sie, wenn Sie rechnen:

Summe = 0+1+1+2+4 = 8

Durchschnitt = 8/5 = 1,6

Sie erkennen, dass das arithmetische Mittel sehr wohl eine Zahl sein kann, die in der Wirklichkeit selbst gar nicht beobachtet werden kann – oder kennen Sie jemand, der 1,6 Kinder hat?

Häufigster Wert (Modus)

Der häufigste Wert ist derjenige Merkmalswert in einem gegebenen Datenbestand, der am häufigsten aufgetreten ist. Auch er ist ein charakterisierender Mittelwert, weil er ohne Zweifel eine Aussage über die Mitte einer vorliegenden Häufigkeitsverteilung macht. Allerdings ist dies kein rechnerischer Mittelwert, sondern er wird als *lagetypischer Mittelwert* bezeichnet, weil er ausschließlich durch die Lage (die Position) der einzelnen Merkmalswerte bestimmt wird.

Offensichtlich ist der Modus viel leichter aufzufinden, als ein arithmetisches Mittel. Er kann aber sinnvollerweise nur bei solchen Häufigkeitsverteilungen bestimmt werden, bei denen die Merkmalswerte mit unterschiedlichen Häufigkeiten aufgetreten sind. Treten irgendwelche Einzelwerte gleich häufig auf, ist der Modus nicht bestimmbar. SPSS bestimmt in diesem Fall den ersten von gleich häufig auftretenden Merkmalswerten als Modus.

Im Beispiel ergibt sich der Modus zu 1 (ein Kind), weil dieser Wert häufiger aufgetaucht ist (nämlich zweimal) als einer der anderen Werte.

Nur im Fall symmetrischer Häufigkeitsverteilungen fallen Modus und arithmetisches Mittel zusammen. Je schiefer eine Verteilung ist, desto weiter fallen die beiden Mittelwerte auseinander. Das hat damit zu tun, dass Extremwerte das arithmetische Mittel deutlich beeinflussen, nicht aber den Modus. Damit wird auch klar, welcher Mittelwert zu verwenden ist, wenn beide eingesetzt werden könnten: Will man, dass Extremwerte Berücksichtigung finden, ist das arithmetische Mittel vorzuziehen; will man das nicht, wählt man zweckmäßigerweise den Modus.

Genau genommen haben Sie mit dem Vergleich der beiden Maße der zentralen Tendenz (Modus und arithmetisches Mittel) zugleich ein Maß zur Charakterisierung der *Schiefe* einer Verteilung in Händen.

Der Zentralwert (Median)

Als dritte Maßzahl soll der Zentralwert genannt werden, dem in seiner Eigenschaft als Mittelwert die gleichen Aufgaben zukommen wie den beiden vorher besprochenen Maßzahlen. Auch der Median ist, wie der gerade besprochene Modus, ein lagetypischer Mittelwert. Er ist definiert als derjenige Merkmalswert, der eine der Größe nach geordnete Reihe von Merkmalswerten halbiert. Wenn eine geradzahlige Anzahl von Werten vorliegt, dann liegt der Median zwischen zwei Werten der geordneten Reihe. Man nimmt dann üblicherweise den mittleren Wert aus diesen beiden.

Die Beschreibung dieser Maßzahl lässt erkennen, dass sie nur eingesetzt werden kann, wenn mindestens ordinalskalierte Daten vorliegen, denn nur diese können ja in eine Rangordnung gebracht werden, was Voraussetzung für die Bestimmung des Medians ist.

Im Beispiel erhalten Sie den Wert 1 (ein Kind), weil dieser Wert in der Mitte der fünf geordneten Werte steht.

5.2 Mittelwertberechnungen

Die Bestimmung der Mittelwerte, die bei den obigen Demonstrationsbeispielen schon vorgeführt wurde, ist bei umfangreichen Datenbeständen mit Hilfe von SPSS sehr leicht zu

bewerkstelligen. Prinzipiell bieten sich zwei Wege an, entweder die Berechnung im Zusammenhang mit der Erstellung der Häufigkeitsverteilungen, oder die Berechnung unabhängig von Häufigkeitsverteilungen

Schauen wir uns zunächst den ersten Weg an, und betrachten wir eine neue Variable, nämlich die Kinderzahl befragter Ehepaare. Zehn Ehepaare wurden befragt. Für den Statistiker ist das natürlich ein ausgesprochen magerer Datenbestand, aber für Demonstrationszwecke ist er sicherlich gut geeignet. Übrigens, was für zehn Werte gilt, gilt natürlich auch für 200 oder für 10000 Werte. Die Ausgangsdaten finden Sie in der Dateneingabetabelle der Abbildung 5.1:

	kinder
1	0
2	1
3	3
4	2
5	4
6	1
7	1
8	0
9	1
10	1
11	

Abb. 5.1: Kinderzahl befragter Ehepaare

Zur Erstellung der Häufigkeitsverteilung dieser diskreten Variablen und zur Bestimmung der Mittelwerte dieser kleinen Verteilung gehen Sie wie folgt vor:

1. Wählen Sie ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN...
2. Übertragen Sie im Dialogfenster der Abbildung 5.2 die Variable „Kinder“ nach rechts ins Feld VARIABLE(N):.

Abb. 5.2: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN...

Klicken Sie jetzt aber noch nicht OK an, sondern klicken Sie auf die Schaltfläche STATISTIKEN... Es öffnet sich das Dialogfenster der Abbildung 5.3

In diesem Dialogfenster der Abbildung 5.3 erkennen Sie im rechten Bereich die drei genannten Mittelwerte. Sie werden dort unter der Überschrift LAGEMAßE unter den Namen MITTELWERT, MEDIAN und MODALWERT (das ist der Modus) aufgeführt.

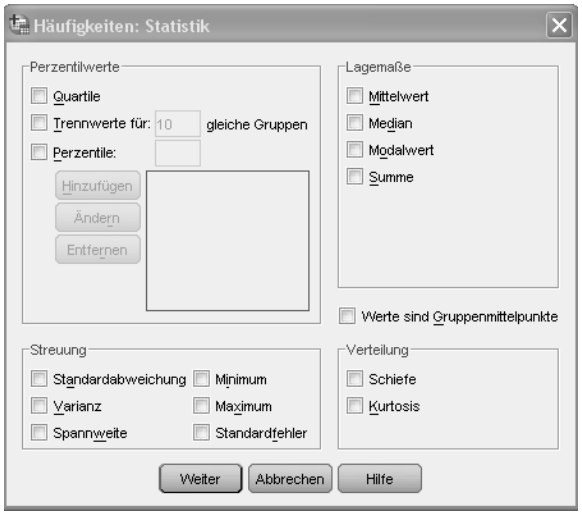


Abb. 5.3: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN..., Schaltfläche STATISTIKEN...

3. Bei allen drei Werten klicken Sie jetzt die quadratischen Kästchen an, so dass dort Häkchen entstehen.
4. Klicken Sie auf WEITER, um zum Fenster der Abbildung 5.2 zurück zu gelangen.
5. Klicken Sie dort OK an.

SPSS erzeugt jetzt die Häufigkeitsverteilung und gibt auch die drei Mittelwerte aus, wie Abbildung 5.4 zeigt. Sie sehen, als arithmetisches Mittel erhalten Sie 1,4 (Kinder), Median und Modus liegen bei jeweils 1 (ein Kind).

Statistiken

kinder

N	Gültig	10
	Fehlend	0
Mittelwert		1,40
Median		1,00
Modus		1

kinder

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	2	20,0	20,0	20,0
	1	5	50,0	50,0	70,0
	2	1	10,0	10,0	80,0
	3	1	10,0	10,0	90,0
	4	1	10,0	10,0	100,0
	Gesamt	10	100,0	100,0	

Abb. 5.4: Diskrete Häufigkeitsverteilung mit Mittelwerten

Der zweite angesprochene Weg zur Berechnung der Mittelwerte geht über das Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/DESKRIPTIVE STATISTIK... Diese Auswahl führt ins Dialogfenster der Abbildung 5.5.

Abb. 5.5: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/DESKRIPTIVE STATISTIK...

In diesem Dialogfenster übernehmen Sie wieder die Variable „Kinder“ nach rechts und klicken dann die Schaltfläche OPTIONEN... an. Damit gelangen Sie zum Dialogfenster der Abbildung 5.6.

Wie Sie dem Dialogfenster der Abbildung 5.6 entnehmen können, wird Ihnen hier, außer Maßzahlen, die wir später noch besprechen werden, nur das arithmetische Mittel unter dem Namen MITTELWERT angeboten.

Versehen Sie, falls das noch erforderlich ist, das dazugehörige Kontrollkästchen mit einem Haken, klicken Sie alle anderen eventuell vorhandenen Häkchen weg, und klicken Sie die Schaltfläche WEITER an.

Im Dialogfenster der Abbildung 5.5 klicken Sie dann auf OK. Ausgegeben wird jetzt, dass die Variable „Kinder“ bei N=10 gültigen Fällen den Mittelwert 1,4 aufweist.

Abb. 5.6: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/DESKRIPTIVE STATISTIK...,
Schaltfläche OPTIONEN...

Wir greifen jetzt einmal auf den umfangreicheren Datenbestand aus Kapitel 2 (203 Befragungen) zurück, um nach dem gleichen Muster wie eben das arithmetische Mittel der Altersangaben zu berechnen. Es ist folgendermaßen vorzugehen:

1. Öffnen der Ausgangsdatei B00.SAV (DATEI/ÖFFNEN/DATEN...).
2. Anwählen von ANALYSIEREN/DESKRIPTIVE STATISTIKEN/DESKRIPTIVE STATISTIK...
3. Auswahl der Variablen „Alter“.
4. Anklicken der Schaltfläche OPTIONEN...
5. Anklicken von MITTELWERT und Beseitigung aller anderen Häkchen.
6. Anklicken der Schaltfläche WEITER.
7. Anklicken von OK.

Es erscheint jetzt die Ausgabe der Abbildung 5.7: Es ergibt sich also als durchschnittliches Alter der Wert 43,07 (Jahre).

Deskriptive Statistik

	N	Mittelwert
alter	203	43,07
Gültige Werte (Listenweise)	203	

Abb. 5.7: Mittelwert der Altersverteilung

6 Streuungsmaße und weitere Maße

6.1 Zielsetzungen und statistische Methoden

Im vorangegangenen Kapitel wurde über Mittelwerte gesprochen. Diese Maßzahlen der zentralen Tendenz sind besonders wichtige Kennziffern eines univariaten Datenbestandes. Sie lassen jedoch ein weiteres wichtiges Charakteristikum dieser Daten unberücksichtigt, nämlich die Streuungsverhältnisse. In einem Zeitungsartikel war kürzlich zu lesen, dass das Netto-Durchschnittseinkommen der Deutschen ungefähr bei 1600 Euro monatlich liegt. Pro Kopf hat also jeder diesen Betrag zur Verfügung, soll das heißen. Allerdings kann eine solche durchaus informative Aussage auch irreführend sein. Wenn nämlich zwei Personen jeweils über 1600 Euro verfügen, dann haben sie im Schnitt tatsächlich 1600 Euro pro Kopf. Der gleiche Durchschnitt ergibt sich aber selbstverständlich auch dann, wenn einer der beiden nichts hat, der andere aber 3200 Euro monatlich. Was sagt also ein Durchschnittswert überhaupt aus? Offensichtlich ist es zweckmäßig, dieser informativen Maßzahl der zentralen Tendenz noch eine zusätzliche Maßzahl beizufügen, die etwas darüber aussagt, wie weit die einzelnen Einkommensgrößen von dem gemeinsamen Durchschnitt abweichen. Nur auf diese Weise kann man Durchschnittsangaben sachgerecht interpretieren.

Spannweite

Die Spannweite (englisch: range) ist nichts anderes als die Differenz zwischen dem größten und dem kleinsten Merkmalswert in einem gegebenen Datenbestand. Wenn man die Ausgangsdaten der Größe nach sortiert, kann man die beiden Extremwerte sofort ablesen und die Spannweite berechnen. SPSS leistet dies, wie Sie sehen werden, ohne vorher sortieren zu müssen.

Die Spannweite ist allerdings als Streuungsmaß nicht allzu beliebt, weil sie zu empfindlich auf Extremwerte reagiert. Sinnvoller, und in der statistischen Praxis weitaus häufiger, ist die Standardabweichung.

Standardabweichung

Diese Maßzahl ist in der praktischen statistischen Arbeit so bedeutsam geworden, dass man meistens sie meint, wenn man nur von „Streuung“ spricht. Es handelt sich hier um eine rechnerische Maßzahl für metrische Daten, die auf folgenden Überlegungen beruht: Man schaut sich an, wie weit die einzelnen erfragten Merkmalswerte von ihrem eigenen arithmetischen Mittel abweichen und notiert diese Abweichungen (falls man per Hand rechnen möchte). Die Summe dieser Abweichungen ist immer null, weil sich negative und positive Abweichungen vom arithmetischen Mittel gegenseitig aufheben. Um dies zu verhindern, werden die Abweichungen quadriert, und alle quadrierten Abweichungen werden aufaddiert. Die entstehende Summe wird durch die Anzahl der Befragungen, also durch n, dividiert (siehe aber den Hinweis im Anschluss an das folgende Rechenbeispiel), und aus der so entstehenden durchschnittlichen quadrierten Abweichung wird die Wurzel gezogen .

Betrachten wir noch einmal die fünf Werte des Beispiels aus Kapitel 5 (Kinderzahlen von fünf Familien). Dort hatte sich als arithmetisches Mittel (durchschnittliche Kinderzahl) der Wert 1,6 ergeben, so dass Sie in der folgenden Arbeitstabelle die Rechenschritte für die Standardabweichung erkennen können:

Arbeitstabelle zur Berechnung der Standardabweichung:

Ehepaar	Kinderzahl	Abweichung	Quadrierte Abweichung
Müller	0	-1,6	2,56
Meier	1	-0,6	0,36
Weber	1	-0,6	0,36
Schmidt	2	+0,4	0,16
Olbrich	4	+2,4	5,78
Summe	*8*	*0*	*9,24*

Standardabweichung = $\sqrt{9{,}24/5}$ = 1,36

Man kann also sagen, die durchschnittliche Kinderzahl liegt bei 1,6 Kindern bei einer Streuung, gemessen mit der Standardabweichung, von 1,36 (Kindern), oder anders formuliert: Im Schnitt weichen die einzelnen Befragungsergebnisse um den Wert 1,36 vom Durchschnittswert (vom arithmetischen Mittel) ab.

Hinweis

In vielen Lehrbüchern finden Sie zur Berechnung der Standardabweichung eine Formel, bei der die Summe der quadrierten Abweichungen vom Mittelwert nicht, wie oben behauptet, durch n, sondern durch (n-1) dividiert wird. So geht auch SPSS vor, d.h. es wird bei dem obigen kleinen Beispiel nicht durch 5, sondern durch 4 dividiert, so dass sich 1,517 ergibt. Zur Begründung dient folgende Überlegung:

Interpretiert man die fünf Zahlenwerte, von denen das obige Beispiel ausgeht, als Zufallsstichprobe aus einer umfangreicheren Grundgesamtheit – und genau genommen interpretiert SPSS alle Daten, die Sie eingeben, als Daten einer Zufallsstichprobe –, so ist es u.a. Anliegen des Statistikers, von den Ergebnissen der Zufallsstichprobe auf die unbekannte Grundgesamtheit zu schließen. Es interessiert ihn beispielsweise die Frage, wie denn die Streuung in der Grundgesamtheit aussieht, ausgehend von der in der Stichprobe berechneten Standardabweichung. Man spricht in diesem Zusammenhang von der „*Schätzung* der Grundgesamtheitsstreuung".

Wenn man sich nun einmal gedanklich vorstellt, dass man aus der Grundgesamtheit **alle** verschiedenen Stichproben vom Umfang n = 5 ziehen würde (das müssen sehr, sehr viele sein), und wenn man sich weiter vorstellt, dass in allen diesen Stichproben die jeweilige Standardabweichung berechnet würde (man würde dann sehr viele unterschiedliche Stichprobenstandardabweichungen erhalten), und wenn man schließlich alle diese Standardabweichungen mitteln würde, dann müsste sich logischerweise die (bis dahin unbekannte) Standardabweichung der Grundgesamtheit ergeben.

Diese Überlegung trifft allerdings nur dann zu – und dies kann mathematisch bewiesen werden –, wenn alle diese Stichprobenstandardabweichungen so berechnet werden, dass nicht durch 5, sondern dass durch 4 (allgemein durch n-1) dividiert wird. Nur dann ist der Mittelwert aller Stichprobenstandardabweichungen identisch mit der Standardabweichung der Grundgesamtheit.

In diesem Fall, also wenn bei der Berechnung der Standardabweichung einer einzigen Stichprobe durch n-1 dividiert wird, nennt man die Stichprobenstandardabweichung eine *erwartungstreue Schätzung* für die unbekannte Standardabweichung der Grundgesamtheit. *Erwartungstreue* wiederum ist eine außerordentlich wichtige und günstige Eigenschaft von Maßzahlen aus Zufallsstichproben, wenn sie zur Schätzung der entsprechenden Maßzahlen einer unbekannten Grundgesamtheit

verwendet werden sollen. Und genau deshalb dividiert SPSS nicht durch n = 5, sondern durch n-1 = 4.

Auch das Quadrat der Standardabweichung wird nicht selten als Streuungsmaß benutzt. Es heißt *Varianz* und ergibt sich hier zu $1{,}36^2 = 1{,}848$.

Häufig wünscht man auch ein *dimensionsloses Streuungsmaß*, insbesondere dann, wenn verschiedene Verteilungen hinsichtlich ihrer Streuungsverhältnisse miteinander verglichen werden sollen. Dann verwendet man den sog. *Variationskoeffizienten*, der in der Weise zustande kommt, dass die Standardabweichung auf das arithmetische Mittel bezogen wird:

Variationskoeffizient = 1,36 / 1,6 * 100 = 85 (%)

Man kann nun sagen, dass die Streuung 85% des Mittelwerts ausmacht.

Quartile

Die bisher besprochenen Streuungsmaße eignen sich nur für metrische Daten. Hat man ordinalskalierte Variablen, so können Streuungsverhältnisse durch Quartile zum Ausdruck gebracht werden (bei metrischen Daten kann damit ebenfalls gearbeitet werden). Dazu wird die geordnete Reihe von Merkmalswerten in vier gleiche Teile geteilt (jeweils 25% also), und man schaut sich an, unter welchem Merkmalswert das Viertel der kleinsten Werte und über welchem Merkmalswert das Viertel der größten Werte liegt. Subtrahiert man diese beiden speziellen Merkmalswerte voneinander und dividiert diese Differenz durch 2, erhält man den sog. *Semiquartilsabstand*, der ebenfalls als Streuungsmaß benutzt wird.

6.2 Berechnung von Streuungsmaßen

Beim Einsatz von SPSS gibt es, wie schon bei den Mittelwerten, zwei Wege:

1. Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/DESKRIPTIVE STATISTIK...
2. Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN...

Um beispielsweise die Streuungsverhältnisse der Altersverteilung unseres Ausgangsbeispiels auf dem zweiten Weg zu bestimmen, ist folgendermaßen vorzugehen:

1. Auswahl von ANALYSIEREN/ DESKRIPTIVE STATISTIKEN/ HÄUFIGKEITEN...
2. Übertragen der Variablen „Alter“ im Dialogfenster der Abbildung 6.1 nach rechts.

Abb. 6.1: ANALYSIEREN/ DESKRIPTIVE STATISTIKEN/ HÄUFIGKEITEN...

Weiterhin sind jetzt die folgenden Arbeitsschritte erforderlich, die dazu dienen, diejenigen statistischen Maßzahlen festzulegen, die Sie von SPSS berechnet haben wollen. Wir wollen uns in diesem Beispiel auf die wichtigsten und bekanntesten dieser Maßzahlen beschränken:

3. Anklicken der Schaltfläche STATISTIKEN...

Sie gelangen zum Dialogfenster der Abbildung 6.2.

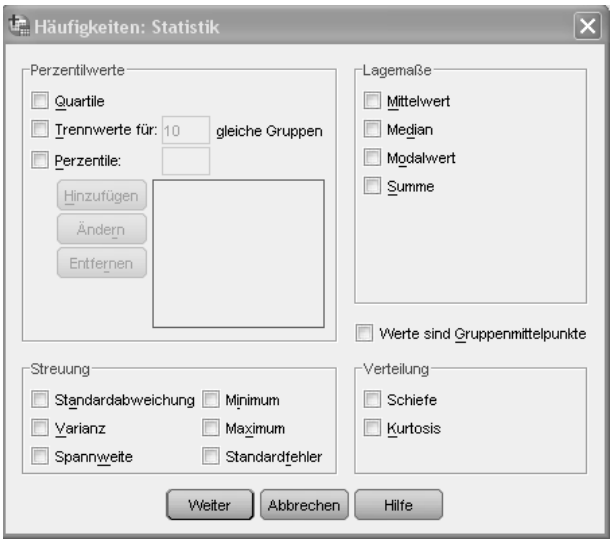


Abb. 6.2: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/HÄUFIGKEITEN, Schaltfläche STATISTIKEN...

4. Im Fenster der Abbildung 6.2 sind Häkchen an den Begriffen MITTELWERT, STANDARDABWEICHUNG, VARIANZ, MINIMUM, MAXIMUM und SPANNWEITE vorzusehen.
5. Anklicken der Schaltfläche WEITER.

Klicken Sie jetzt im Fenster der Abbildung 6.1 das Häkchen bei HÄUFIGKEITSTABELLEN ANZEIGEN weg und dann die Schaltfläche OK an, gelangen Sie zur Ergebnisausgabe der Abbildung 6.3. In diesem Ausgabefenster werden also die interessierenden Maßzahlen zusammenfassend ausgegeben.

Statistiken

alter

N	Gültig	203
	Fehlend	0
Mittelwert		43,07
Standardabweichung		15,523
Varianz		240,975
Spannweite		58
Minimum		19
Maximum		77

Abb. 6.3: Mittelwert und Streuungsmaße der Altersverteilung

Zur Bestimmung des oben genannten Semiquartilsabstandes ist im Fenster der Abbildung 6.2 auch ein Häkchen bei Quartile anzuklicken. Es ergibt sich als erster Quartilswert der Wert 31, d.h. das Alter von 31 Jahren zeichnet sich dadurch aus, dass 25% der Befragten jünger, 75% älter als 31 Jahre sind. Der 50%-Wert, also der zweite Quartilswert, ist nichts anderes als der Median (38 Jahre), der dritte Quartilswert (75%-Wert) liegt bei 55 Jahren. 75% sind jünger, 25% sind älter als 55 Jahre.

Die halbe Differenz zwischen dem dritten und dem ersten Quartilspunkt, darauf wurde weiter oben schon aufmerksam gemacht, ist unter dem Namen *Semiquartilsabstand* als weiteres Streuungsmaß in der Statistik gebräuchlich. Hier ergibt sich (55-31)/2 = 12 (Jahre). Dividiert man diesen Semiquartilsabstand durch den Median, erhält man wieder ein dimensionsloses Streuungsmaß, das *Quartilskoeffizient* genannt wird. Hier ergibt sich 12/38 = 0,3158 oder 31,58%.

6.3 Andere Maßzahlen

In den verschiedenen Dialogfenstern, die geöffnet wurden, um Mittelwerte und Streuungsmaße zu berechnen, sind noch einige Begriffe aufgetaucht, die wir bisher noch nicht besprochen haben. Sie sehen beispielsweise in der Abbildung 6.2 im Bereich VERTEILUNG die Begriffe SCHIEFE und KURTOSIS und im Bereich STREUUNG den Begriff STD.FEHLER.

Schiefe

So wie es Maßzahlen der zentralen Tendenz (Mittelwerte) und Streuungsmaße gibt, so gibt es auch Schiefemaße. Eines davon wird von SPSS berechnet, wenn Sie SCHIEFE anklicken. Da diese Maße aber in der statistischen Praxis kaum benutzt werden, wollen wir darauf nicht näher eingehen.

Kurtosis

Entsprechend das gleiche gilt für die Wölbung (Kurtosis) von Verteilungen.

Std.Fehler

Dieser Begriff, gemeint ist „Standardfehler“, häufig auch mit „s.e.“ für „standard error“ abgekürzt, beleuchtet einen für die Praxis sehr viel wichtigeren Sachverhalt, der Ihnen in anderen Zusammenhängen wieder begegnen wird. Wenn Sie sich für die Altersverteilung diesen Wert ausgeben lassen, ergibt sich der Wert 1,09 (Jahre). Wie kommt dieser Wert zustande und wofür benötigt man ihn?

Um diese Frage zu beantworten, ist eine ähnliche Überlegung erforderlich, wie wir sie oben im Zusammenhang mit der Erwartungstreue einer Stichprobenstandardabweichung angestellt hatten:

Wenn man alle verschiedenen Zufallsstichproben des Umfangs n = 203 aus der gegebenen Grundgesamtheit aller Bundesbürger ziehen würde und in allen diesen (sehr vielen) Stichproben das arithmetische Mittel berechnen würde, dann weichen alle diese Mittelwerte voneinander ab. Zufälligerweise kann natürlich auch einmal ein Mittelwert in einer Stichprobe mit dem einer anderen übereinstimmen. Alle diese Stichprobenmittelwerte, so kann man sagen, streuen. Wenn man nun als Maß für die Streuung dieser speziellen Größen wieder eine Standardabweichung berechnen würde, also die Standardabweichung aller denkbaren Zufallsstichprobenmittelwerte, so ergibt sich eine Größe, die, wie die Mathematiker bewiesen haben, der Grundgesamtheitsstreuung entspricht, geteilt durch die Wurzel aus

dem Stichprobenumfang. Dies wiederum ist ungefähr gleich der Streuung in der (einzigen) Stichprobe, geteilt durch die Wurzel aus dem Stichprobenumfang.

Die Stichprobenstandardabweichung lag bei 15,523 (Jahre). Dividiert man also 15,524 durch die Wurzel aus 203, ergibt sich der Wert 1,09 (Jahre). Diese Zahl ist, allgemein gesprochen, ein Maß dafür, wie genau der Stichprobenmittelwert (43,069 Jahre) den unbekannten Mittelwert der Grundgesamtheit „trifft". Diese Größe, die offenbar mit zunehmendem Stichprobenumfang n immer kleiner wird, spielt bei den wahrscheinlichkeitsstatistischen Methoden, die für die Stichprobenstatistik eingesetzt werden (wir werden noch darauf zu sprechen kommen), eine sehr wichtige Rolle. Sie wird auch manchmal „*Stichprobenfehler*" oder „*Standardfehler*" genannt.

6.4 Standardisierung

Nachdem Sie in Kapitel 5 das arithmetische Mittel und in Abschnitt 6.2 die Standardabweichung kennengelernt haben, soll hier auf ein Verfahren aufmerksam gemacht werden, dem die Statistiker den Namen *Standardisierung* (manchmal auch „*Z-Standardisierung*") gegeben haben. Es geht dabei darum, die Werte einer Variablen um das arithmetische Mittel zu bereinigen, um dann die so entstandenen Differenzen auf die Standardabweichung zu beziehen. Betrachten Sie zur Illustration die folgenden vier Werte der Untersuchungsvariablen Körpergröße:

172	168	180	180

Aus diesen vier Werten können Sie das arithmetische Mittel und die Standardabweichung errechnen. Es ergibt sich als Mittelwert der Wert 175 (cm) und als Standardabweichung der Wert 5,196 (cm). Führt man nun die oben erwähnte *Transformierung* (Standardisierung) durch, gelangt man z.B. für die erste Person zu (172-175)/5,196 = -0,577. Dies bedeutet, dass die erste Person mit ihrer Körpergröße um 0,577 Standardabweichungen unter dem Schnitt liegt. Die vier standardisierten Werte stellen sich folgendermaßen dar:

-0,577	-1,347	0,962	0,962

Für diese standardisierten Werte gilt, wie Sie leicht überprüfen können, dass ihr arithmetisches Mittel 0 ist; ihre Standardabweichung ist 1.

Der Sinn solcher Standardisierungen wird durch folgende Überlegung erkennbar: Sie werden in den folgenden Kapiteln Verfahren kennen lernen, die sich mit zwei oder mehr Untersuchungsvariablen gemeinsam beschäftigen. Bei solchen Verfahren ist es häufig erforderlich, die Werte der verschiedenen Variablen gemeinsam zu verrechnen. Stellen Sie sich beispielsweise vor, Sie hätten die Variable „monatliches Haushalts-Nettoeinkommen" und die Variable „Kinderzahl befragter Haushalte" gemeinsam zu betrachten, etwa um der Frage nachzugehen, ob kinderreiche Haushalte tendenziell ärmer sind. Die erste genannte Variable bewegt sich in Größenordnungen von vielleicht 2000 bis 10000, die zweite im Bereich 0 bis vielleicht 7. Wenn nun beide Variablen in einen gemeinsamen Rechenalgorithmus eingehen, würde die erste von beiden ein völlig unangemessenes Gewicht im Vergleich zur zweiten erhalten, einfach weil die einzelnen Werte numerisch viel größer sind. Um dies zu verhindern, werden die Werte beider Variablen standardisiert. Sie bewegen sich dann in gleichen Größenordnungen, gehen mithin mit gleichem Gewicht in die notwendigen Berechnungen ein.

Im Dialogfenster, welches das Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/ DESKRIPTIVE STATISTIK… öffnet (siehe Abbildung 5.5), taucht links unten der Text STANDARDISIERTE WERTE ALS VARIABLE SPEICHERN auf. Standardisierte Werte werden wiederum als Ausgangswerte anderer statistischer Verfahren benötigt – etwa im Rahmen der Faktorenanalyse, auf die wir in Kapitel 17 zu sprechen kommen. Benötigt man also für ein später einzusetzendes Verfahren standardisierte Werte, so können diese hier erzeugt werden. Sie werden von SPSS dann als weitere Spalte der Ausgangstabelle hinzugefügt, die mit ZALTER überschrieben wird. Man nennt diese Standardisierung manchmal auch Z-Standardisierung (*Z-score*), daher das Z am Anfang des neuen Variablennamens.

7 Bivariate Verteilungen

7.1 Zielsetzungen und statistische Methoden

Nun soll der Blickwinkel erweitert werden, indem wir uns mit statistischen Zusammenhängen befassen. Mit der Frage nach solchen Zusammenhängen wird eine Brücke geschlagen zwischen der reinen Deskription erhobener Befunde und ihrer Erklärung. Es steht nicht mehr die Frage im Mittelpunkt des Interesses, **wie** ein bestimmter Datenbestand aussieht, sondern **warum** er so aussieht, wie er sich Ihnen präsentiert.

Ein erster Schritt, solche Fragen zu beantworten, besteht darin, zwei Variablen gleichzeitig zu betrachten. Damit gelangt man zu bivariaten Verteilungen. Auch diese können grafisch oder tabellarisch präsentiert werden, was in diesem Kapitel besprochen werden soll. Zudem können Maßzahlen zur Charakterisierung bivariater Verteilungen berechnet werden, was im Kapitel 9 erörtert wird.

Betrachten Sie das folgende Demonstrationsbeispiel: Zufällig ausgewählte Schüler einer Altersklasse müssen an einem Deutsch-Diktat und an einer Klassenarbeit in Mathematik teilnehmen. Es werden in beiden Arbeiten die Fehler ausgezählt, um der Frage nachgehen zu können, ob zwischen den Leistungen in Deutsch und denen in Mathematik ein Zusammenhang besteht.

Schüler	Fehler in Deutsch	Fehler in Mathematik
Meier	5	4
Müller	2	7
Weber	7	2
Schmidt	0	7
Olbrich	3	5
Adam	6	3
Carius	8	2
Bevermann	2	5
Walter	6	1
Feser	3	6
usw.		

Zur grafischen Darstellung einer bivariaten Verteilung verwendet man ein Achsenkreuz. Auf der X-Achse (waagrechte Achse) trägt man die Werte derjenigen Variablen ab, von der man glaubt, dass sie die andere beeinflusst, auf der Y-Achse (senkrechte Achse) trägt man die Variable ab, die von der anderen beeinflusst wird, bzw. von ihr abhängt. In unserem Schülerbeispiel kann die Frage, welches die unabhängige und welches die abhängige Variable ist, nicht so ohne weiteres beantwortet werden. Letztlich entscheidend ist Ihre Untersuchungshypothese: Wenn Sie vermuten, dass die Deutschleistungen die mathematischen Fähigkeiten in irgendeiner Weise beeinflussen, werden Sie erstere auf der X-Achse, die Mathematikleistungen hingegen auf der Y-Achse abtragen. Gehen Sie von der gegenteiligen Hypothese aus, werden Sie die Achsen vertauschen. Können Sie sich nicht für eine Variante entscheiden, bleibt immer noch die Möglichkeit, mit zwei Achsenkreuzen alternativ zu arbeiten. Sie werden später sehen, dass sich einige wesentliche Befunde der statistischen Methoden, allerdings nicht alle, nicht davon beeinflussen lassen, welche Variable auf welcher Achse steht (siehe Kapitel 9).

Jeder Merkmalsträger (jeder Schüler) lässt sich dann als ein Punkt in einem solchen Achsenkreuz präsentieren. Man spricht von einem Streudiagramm. Um dieses zu erzeugen, benötigen Sie wieder das Fenster, das sich über DIAGRAMME/GRAFIKTAFEL-VORLAGEN-AUSWAHL… ergibt. Hier übertragen Sie „Deutsch“ und „Mathe“ nach rechts und klicken auf STREUDIAGRAMM.

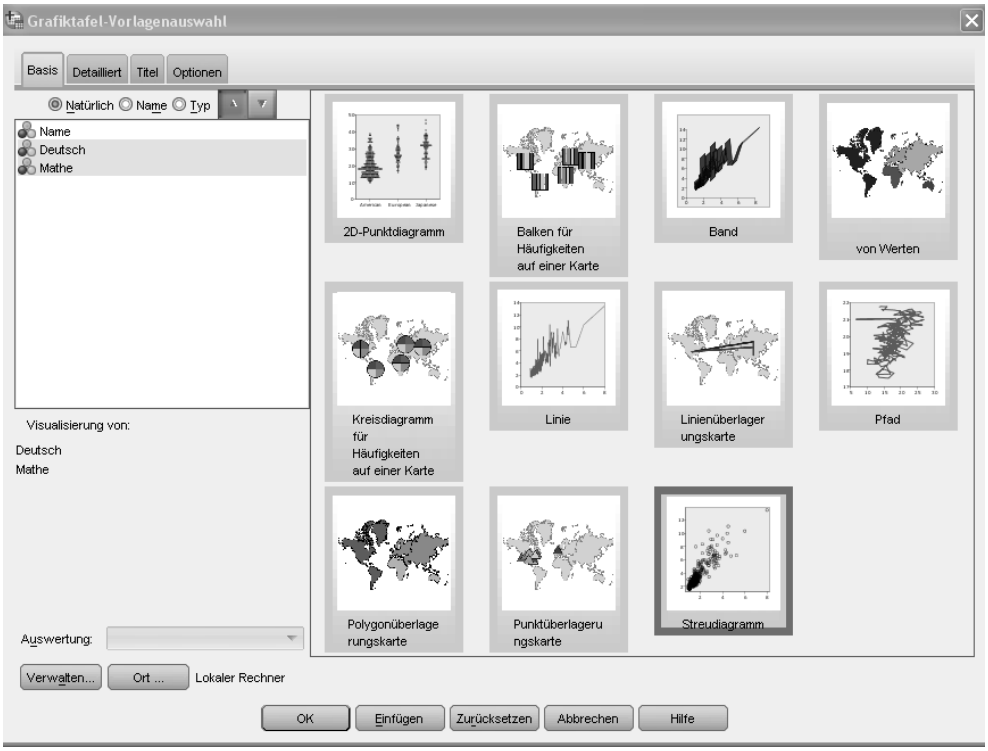


Abb. 7.1: Menü DIAGRAMME/GRAFIKTAFEL-VORLAGENAUSWAHL…

Nach Anklicken von OK ergibt sich das Streudiagramm der Abbildung 7.2.

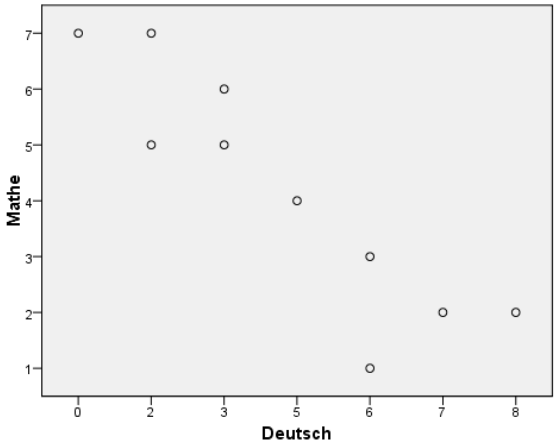


Abb. 7.2: Streudiagramm

Sie erkennen in der Abbildung 7.2, dass sich eine „*Punktwolke*“ ergibt. Sie ist bei 10 Fällen natürlich noch recht dürftig, zeigt aber eine eindeutige Tendenz: Mit zunehmenden Werten der Variablen X (mehr Deutschfehler) nehmen tendenziell die Werte der Variablen Y ab (weniger Mathematikfehler); die Punktwolke fällt von links oben nach rechts unten.

Man spricht in diesem Fall von einem gegenläufigen statistischen Zusammenhang. Dieser gegenläufige Zusammenhang gilt nicht für jeden Vergleich zwischen zwei Schülern, aber er gilt generell für die gesamte Punktwolke, d.h. für die Gesamtheit der untersuchten Schüler.

Würden Sie eine Punktwolke erhalten, die von links unten nach rechts oben steigt, würden Sie von einem gleichläufigen (gleichgerichteten) Zusammenhang sprechen. Weiterhin kann schon an dieser Stelle folgendes behauptet werden: Je enger (schlanker) die Punktwolke ist, desto stärker ist der Zusammenhang zwischen beiden Untersuchungsvariablen. Schließlich ist zu beachten, dass Punktwolken ganz unterschiedliche Gestalt annehmen können, wobei zunächst der Einfachheit halber zwischen linearen und nichtlinearen Punktwolken unterschieden werden soll. Das Beispiel der Abbildung 7.2 repräsentiert eine tendenziell lineare Punktwolke

Betrachtet man aber beispielsweise als X-Variable den Düngemitteleinsatz auf Probefeldern und als Y-Variable den Ernteertrag, kann eine steigende Punktwolke erwartet werden, deren Anstieg sich aber abschwächt, d.h. es dürfte eine gekrümmte Punktwolke entstehen.

Ob im konkreten Fall von einer linearen oder von einer nichtlinearen Punktwolke auszugehen ist, entscheidet zunächst das optische Erscheinungsbild. Wesentlich sind aber inhaltlich-theoretische Überlegungen: Beispielsweise ist jedem Landwirt bekannt, dass mit zunehmendem Düngemitteleinsatz der Ernteertrag zunächst mit zunehmenden, dann mit abnehmenden Zuwächsen steigt, um bei Überdüngung der Probefelder vielleicht sogar wieder rückläufig zu sein. Somit muss logischerweise eine nichtlineare Beziehung unterstellt werden.

Wir werden uns im Folgenden aber bevorzugt mit linearen Beziehungen zwischen zwei Variablen beschäftigen, weil dies auch der statistischen Praxis entspricht, und weil die linearen Beziehungen, wie wir noch sehen werden, besondere Interpretationsmöglichkeiten bieten.

7.2 Bivariate Tabellen

Um eine bivariate Tabellen zu erzeugen – man spricht von einer Kreuztabelle – verwenden wir das Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/KREUZTABELLEN….

Wir greifen auf den Ausgangsdatenbestand der Datei B00.SAV zu, der in einem früheren Kapitel benutzt wurde. Es soll gezeigt werden, wie sich die beiden Variablen „Geschlecht“ und „bevorzugte politische Partei“ in einer gemeinsamen Tabelle darstellen lassen.

1. Wählen Sie Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/ KREUZTABELLEN…

Man gelangt zum Fenster der Abbildung 7.3.

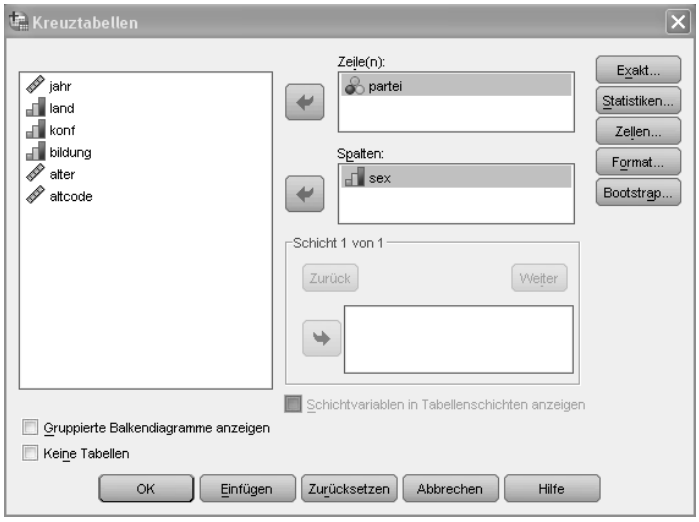


Abb. 7.3: Menü ANALYSIEREN/DESKRIPTIVE STATISTIKEN/ KREUZTABELLEN…

2. Übertragen der Variablen „Partei (Partei)“ mit dem nach rechts zeigenden Pfeil in das Feld ZEILEN: (schon geschehen)
3. Übertragen der Variablen „Geschlecht (Sex)“ mit dem nach rechts zeigenden Pfeil in das Feld SPALTEN: (schon geschehen)
4. Anklicken von OK (alle anderen Bereiche und Schaltflächen im Fenster der Abbildung 7.3 sollen zunächst nicht interessieren).

SPSS erzeugt jetzt die Ausgabe der Abbildung 7.4.

Partei * Geschlecht Kreuztabelle

Anzahl

		Geschlecht		
		männlich	weiblich	Gesamt
Partei	CDU/CSU	43	40	83
	SPD	44	34	78
	F.D.P.	5	6	11
	Die Grünen	6	15	21
	Sonstige	1	6	7
Gesamt		99	101	200

Abb. 7.4: Geschlecht und bevorzugte politische Partei (Kreuztabelle)

Sie sehen in dieser Abbildung, dass es mehr oder weniger deutliche geschlechtsspezifische Unterschiede gibt. Beispielsweise neigen die weiblichen Befragten deutlicher der Partei „Die Grünen“ zu als die Männer.

Wie man solche Unterschiedlichkeiten zu bewerten hat, und wie man sie mit geeigneten statistischen Maßzahlen quantifizieren kann, wird ausführlicher erst in Kapitel 9 besprochen.

Die Interpretation solcher Kreuztabellen wird erleichtert, wenn man auch Prozentwerte ausgeben lässt. Dazu muss im Fenster der Abbildung 7.3 die Schaltfläche ZELLEN angeklickt werden, was zu dem folgenden Fenster führt:

Abb. 7.5: Kreuztabellen/Schaltfläche ZELLEN

In diesem Fenster klicken wir an bei ERWARTET, ZEILENWEISE, SPALTENWEISE UND GESAMT. Dann ergibt sich mit WEITER und OK das folgende Bild:

Partei * Geschlecht Kreuztabelle

			Geschlecht		
			männlich	weiblich	Gesamt
Partei	CDU/CSU	Anzahl	43	40	83
		Erwartete Anzahl	41,1	41,9	83,0
		% innerhalb von Partei	51,8%	48,2%	100,0%
		% innerhalb von Geschlecht	43,4%	39,6%	41,5%
		% der Gesamtzahl	21,5%	20,0%	41,5%
	SPD	Anzahl	44	34	78
		Erwartete Anzahl	38,6	39,4	78,0
		% innerhalb von Partei	56,4%	43,6%	100,0%
		% innerhalb von Geschlecht	44,4%	33,7%	39,0%
		% der Gesamtzahl	22,0%	17,0%	39,0%
	F.D.P.	Anzahl	5	6	11

Abb. 7.6: Kreuztabelle mit Prozentangaben (Ausschnitt)

Am Beispiel der männlichen CDU/CSU-Anhänger sollen die Zahlenwerte kurz erläutert werden:

Zahl	Bezeichnung	Bedeutung
43	Anzahl	Anzahl der Fälle
41,1	erwartete Anzahl	Dazu siehe unten
51,8	% von Partei	Anteil der Männer an allen CDU/CSU-Wählern
43,4	% von Geschlecht	Anteil der CDU/CSU-Wähler bei den Männern
21,5	% der Gesamtzahl	Anteil der männlichen CDU/CSU-Wählern an allen befragten Personen

Mit der „erwarteten Anzahl“ hat es folgendes auf sich:

Sie erkennen in der letzten Zeile der Kreuztabelle (siehe Abbildung 7.4), dass insgesamt 99 Männer und 101 Frauen berücksichtigt wurden. Von den Männern waren 43 CDU/CSU-Anhänger; die entsprechende Zahl der Frauen war 40. Insgesamt wurden also 83 CDU/CSU-Anhänger gezählt.

Wenn man nun unterstellt, dass es zwischen dem Geschlecht und der bevorzugten politischen Partei keinen Zusammenhang gibt, dann wäre zu erwarten, dass sich die 83 CDU/CSU-Anhänger im Verhältnis 99:101 auf die beiden Geschlechter aufteilen müssten. Von den 83 CDU/CSU-Anhängern müssten demnach (bei Unabhängigkeit der beiden Variablen voneinander) 83*99/200 Männer und 83*101/200 Frauen zu erwarten sein.

Rechnet man diese beiden Ausdrücke aus, gelangt man zu 41,1 und 41,9. Dies sind die beiden Zahlen, die Sie unter dem Begriff ERWARTETE ANZAHL in den Tabellenzellen finden. Für die anderen Parteien lassen sich entsprechende Berechnungen durchführen, was SPSS ja auch schon erledigt hat. Diese Erwartungswerte werden Ihnen im Kapitel 10 wieder begegnen. Schon jetzt kann aber gesagt werden, dass die Abhängigkeit zwischen „Geschlecht“ und „bevorzugter politischer Partei“ um so größer sein wird, je weiter die beobachteten Häufigkeiten von den bei Unabhängigkeit zu erwartenden Werten abweichen.

7.3 Streudiagramm

Bivariate Verteilungen metrischer Daten lassen sich anschaulich als Punktwolken darstellen, wie das ja schon in Abschnitt 7.1 angedeutet wurde. Man spricht in diesem Zusammenhang von einem Streudiagramm.

Nehmen Sie an, eine Reihe zufällig ausgewählter Personen wäre nach Körpergröße (in cm) und nach Gewicht (in kg) befragt worden. Die Ausgangsdaten sind in eine SPSS-Tabelle unter den Variablennamen „cm“ und „kg“ eingegeben worden (siehe Abbildung 7.7).

Erzeugen Sie aus diesen Daten ein Streudiagramm, wie es in Abschnitt 7.1 besprochen wurde, so ergibt sich Abbildung 7.8.

	cm	kg
1	172,00	72,00
2	174,00	71,00
3	168,00	66,00
4	170,00	67,00
5	180,00	81,00
6	192,00	85,00
7	185,00	88,00
8	171,00	71,00
9	166,00	59,00
10	169,00	65,00
11	177,00	71,00
12	181,00	75,00
13	182,00	79,00
14	173,00	71,00
15	169,00	65,00
16	195,00	91,00
17	188,00	95,00
18	159,00	58,00

Abb. 7.7: Ausgangsdaten

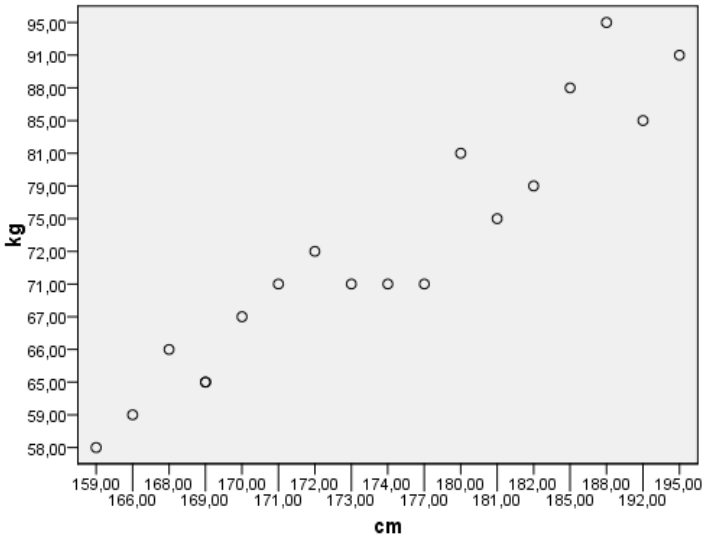


Abb. 7.8: Körpergröße und Körpergewicht

Liegen nichtmetrische Daten vor, sind Streudiagramme sinnlos. Allerdings kann man auch dann illustrative Grafiken erzeugen. In diesem Fall wird kein Streudiagramm erstellt, aber man kann so vorgehen, wie es im Folgenden am Beispiel der beiden Variablen „Geschlecht“ und „bevorzugte politische Partei“ aus der Datei B00.SAV beschrieben wird.

1. Wählen Sie Menü DIAGRAMME/VERALTETE DIALOGFELDER/BALKEN…

Es öffnet sich das Dialogfenster der Abbildung 7.9.

Abb. 7.9: Menü DIAGRAMME/VERALTETE DIALOGFELDER/BALKEN…

2. Wählen Sie im Fenster der Abbildung 7.9 die mittlere Variante GRUPPIERT.

3. Klicken Sie im Bereich DATEN IM DIAGRAMM auf den Optionsschalter bei AUSWERTUNG ÜBER KATEGORIEN EINER VARIABLEN, falls das erforderlich ist.
4. Klicken Sie die Schaltfläche DEFINIEREN an.
5. Übertragen Sie die Variable im Fenster der Abbildung 7.10 „partei“ in den Bereich KATEGORIENACHSE:.
6. Übertragen Sie die Variable „sex“ in den Bereich GRUPPEN DEFINIEREN DURCH: und klicken Sie auf OK.

Diese Prozedur führt zu dem Diagramm der Abbildung 7.11.

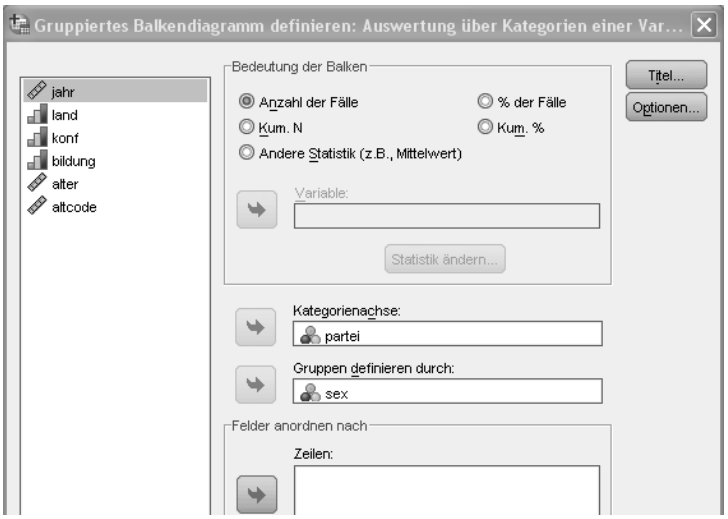


Abb. 7.10: Schaltfläche DEFINIEREN (Ausschnitt)

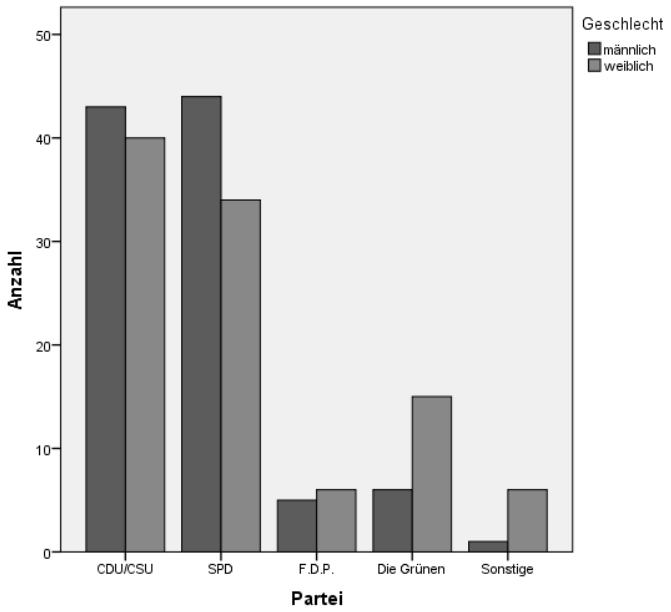


Abb. 7.11: Geschlecht und bevorzugte politische Partei

8 Wahrscheinlichkeitsstatistik

8.1 Ausgangslage

Einige der Ergebnisse, die SPSS bei statistischen Auswertungen produziert, können nur sachgerecht interpretiert werden, wenn man etwas von Wahrscheinlichkeitsstatistik versteht. Deshalb werden wir in diesem Kapitel, bevor weitere SPSS-Anwendungen besprochen werden, einige der diesbezüglichen Grundkonzepte vorstellen, damit die Erläuterungen der dann folgenden Kapitel besser verständlich werden.

Ausgangspunkt ist die Feststellung, dass die statistische Arbeit letztlich dazu dient, *Hypothesen* zu überprüfen. Wenn Sie beispielsweise eine Stichprobe ziehen, in der wir zufällig ausgewählte wahlberechtigte Personen nach ihrer Parteipräferenz befragen und zusätzlich demografische Variablen, wie z.B. „Alter", „Geschlecht" oder „Konfession" erheben, dann deshalb, weil Sie die Hypothese untersuchen wollen, dass diese demografischen Größen das Wahlverhalten beeinflussen könnten. Wenn Sie in der univariaten Statistik ein durchschnittliches Einkommen berechnen, dann vielleicht deshalb, weil Sie die Hypothese überprüfen wollen, das Durchschnittseinkommen betrage 3200 DM.

Die Überprüfung einer Hypothese erfolgt im Lichte empirischer Befunde, d.h. man stellt der Hypothese statistische Informationen, z.B. Befragungsergebnisse, gegenüber, um dann aufgrund dieser Gegenüberstellung über die vorher formulierte Untersuchungshypothese zu entscheiden. Bei dieser Hypothesenentscheidung können vier Fälle voneinander unterschieden werden:

1. Die Hypothese trifft in Wirklichkeit zu (was Sie aber nicht wissen; wüssten Sie es, bräuchten Sie keine empirischen Stichprobendaten mehr zu sammeln) und der empirische Befund führt dazu, dass Sie die Hypothese bestätigen. Dies wäre offensichtlich eine korrekte Entscheidung.
2. Die Hypothese trifft in Wirklichkeit zu, Sie verwerfen sie aber aufgrund des empirischen Befundes. Sie hätten dann einen Entscheidungsfehler begangen.
3. Die Hypothese trifft in Wirklichkeit nicht zu, der empirische Befund führt auch zu ihrer Verwerfung – korrekte Entscheidung.
4. Die Hypothese trifft in Wirklichkeit nicht zu, der empirische Befund empfiehlt aber ihre Annahme – Fehlentscheidung.

Den Fehler, der im zweiten Fall begangen wird, nennen die Statistiker *Fehler vom Typ I* oder *α-Fehler*. Im vierten Fall spricht man von einem *Fehler vom Typ II* oder *ß-Fehler.*

Die Statistik dient nun u.a. dazu, die Wahrscheinlichkeiten der Fehlentscheidungen zu minimieren, zugleich also die für korrekte Entscheidungen zu maximieren. Diese Überlegungen machen es erforderlich, den Begriff der Wahrscheinlichkeit näher zu beleuchten.

8.2 Wahrscheinlichkeit

Wenn Sie eine Münze werfen, dann ist die Wahrscheinlichkeit dafür, „Kopf" zu werfen, 1/2, d.h. es besteht dafür eine Wahrscheinlichkeit von 1/2 oder 0,5 oder 50%. Die Wahrscheinlichkeit, eine Vier mit einem Würfel zu werfen, beträgt entsprechend 1/6. Die Wahr-

scheinlichkeit dafür, aus den 49 Lottozahlen 1 bis 49 sechs bestimmte Zahlen auszuwählen, beträgt 1/13983816.

Wie gelangt man zu solchen Angaben? Man überlegt sich, wie viele Möglichkeiten überhaupt existieren (bei der Münze gibt es zwei Möglichkeiten, beim Würfel sechs Möglichkeiten, beim Lotto gibt es 13983816 Möglichkeiten, sechs aus 49 Kugeln auszuwählen), und bezieht die Anzahl der im Sinn der Fragestellung günstigen Fälle (bei den Beispielen ist das jeweils ein Fall) darauf.

Betrachten Sie noch ein Beispiel: Wie groß ist die Wahrscheinlichkeit, mit drei Münzen genau zweimal „Kopf" zu werfen? Kürzt man der Übersichtlichkeit halber „Kopf" mit K und „Wappen" mit W ab, so können Sie zunächst feststellen, dass beim dreifachen Münzwurf die folgenden Möglichkeiten geboten sind:

WWW, WWK, WKW, KWW, WKK, KWK, KKW und KKK

Dies sind acht Möglichkeiten. Drei davon sind im Sinn der Fragestellung günstig (die fünfte, sechste und siebte der oben aufgeführten Möglichkeiten). Die gesuchte Wahrscheinlichkeit ist demnach 3/8.

Man spricht in diesem Zusammenhang von einer *A-priori-Wahrscheinlichkeit*, weil a-priori (von Vornherein) festgestellt werden kann, dass zum Beispiel die Wahrscheinlichkeit für „Kopf" im einfachen Münzwurf 1/2 ist, also bevor Sie auch nur eine einzige Münze werfen. Die so bestimmte Wahrscheinlichkeit nennt man auch „*klassische Wahrscheinlichkeit*", gemäß des klassischen Wahrscheinlichkeitsbegriffs von *Laplace.*

Der Vollständigkeit halber möchten wir auch auf den Begriff der *A-posteriori-Wahrscheinlichkeit* aufmerksam machen (a posteriori = im Nachhinein; *statistische Wahrscheinlichkeit*). Stellt man beispielsweise in einer langen Reihe von Beobachtungen fest, dass der Anteil von Knabengeburten an der Gesamtzahl aller Geburten bei 51,7% (=0,517) liegt, kann man sagen, dass die Wahrscheinlichkeit dafür, dass bei einer Geburt ein Knabe geboren wird, bei 0,517 liegt.

Ein weiterer Begriff ist in diesem Zusammenhang wichtig, nämlich der der Zufallsvariablen. Was eine Variable ist, haben Sie ja schon in den vorangegangenen Kapiteln erfahren. Eine *Zufallsvariable* nun ist eine Variable, deren Ausprägungen zufällig auftreten. Zweifelsohne ist die Variable „Augenzahl beim einfachen Würfelwurf" eine Zufallsvariable (sie kann „zufällig" die Werte 1, 2, 3, 4, 5 oder 6 annehmen). Und natürlich ist auch die „Zahl der Kopfwürfe im dreifachen Münzwurf" eine Zufallsvariable (sie kann die Werte 0, 1, 2 oder 3 annehmen). Aber auch die Variable „Körpergröße einer zufällig ausgewählten Person" ist eine Zufallsvariable, wie leicht einsichtig ist – eben wegen der Zufälligkeit der Auswahl der Person. Diese zuletzt genannte Zufallsvariable ist eine *stetige Variable;* die anderen waren *diskrete Variablen*. Sie sehen, hier tauchen wieder Begriffe auf, die Sie im Zusammenhang mit statistischen Untersuchungsvariablen auch schon kennen gelernt haben.

8.3 Wahrscheinlichkeitsverteilungen

Bei komplizierteren Fragestellungen (wie groß ist die Wahrscheinlichkeit, beim 128-fachen Münzwurf zwischen 60 und 70 Mal „Kopf" zu erhalten? Wie groß ist die Wahrscheinlichkeit für sechs Richtige im Lotto?) ist das Auszählen von möglichen und günstigen Fällen mühsam. Glücklicherweise hat man für solche Fälle Instrumente geschaffen, die das Be-

rechnen von Wahrscheinlichkeiten vereinfachen. Diese Instrumente sind die Wahrscheinlichkeitsverteilungen.

Betrachten Sie noch einmal das Beispiel des dreifachen Münzwurfs. Wenn man sich für die Zufallsvariable „Anzahl der Kopfwürfe im dreifachen Münzwurf“ interessiert, dann kann man rasch erkennen, dass diese spezielle Zufallsvariable vier denkbare Ausprägungen annehmen kann, nämlich die Werte 0, 1, 2 und 3. Diese Variable ist also vom diskreten Typ. Stellt man diesen vier denkbaren Ausprägungen die entsprechenden Wahrscheinlichkeiten gegenüber, gelangt man zu der folgenden Übersicht:

Ausprägungen	Wahrscheinlichkeit
0	1/8
1	3/8
2	3/8
3	1/8
Summe	1

Dies ist die Wahrscheinlichkeitsverteilung für die Zufallsvariable „Anzahl der Kopfwürfe im dreifachen Münzwurf“. Sie sehen übrigens, die Summe der Wahrscheinlichkeiten ist 1 (dies gilt für alle Wahrscheinlichkeitsverteilungen, also nicht nur in diesem speziellen Beispiel).

Eine solche Wahrscheinlichkeitsverteilung kann man auch grafisch darstellen, wobei sich ein Bild ergibt, das einer Verteilung relativer Häufigkeiten in der deskriptiven Statistik entspricht. Die obige Verteilung führt zur Abbildung 8.1.

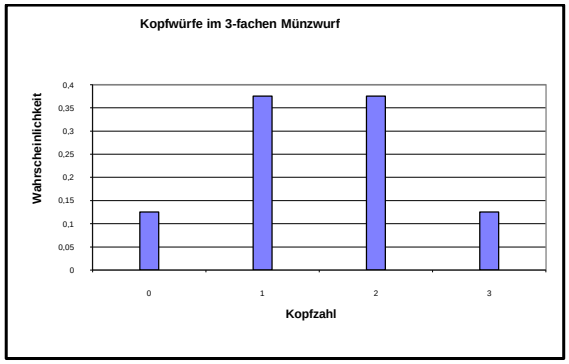


Abb. 8.1 : Diskrete Wahrscheinlichkeitsverteilung

Sie sehen, auf der waagrechten Achse werden die Ausprägungen der interessierenden Zufallsvariablen, auf der senkrechten Achse werden die zugeordneten Wahrscheinlichkeiten abgetragen.

Auch für eine solche Wahrscheinlichkeitsverteilung kann man – wie Sie es in der univariaten Statistik schon gesehen haben (siehe Kapitel 5 und 6) –, charakteristische Maßzahlen berechnen.

Dabei sind das arithmetische Mittel und die Standardabweichung, bzw. ihr Quadrat, die Varianz, besonders bedeutsam. Gewichtet man alle Merkmalswerte (0, 1, 2 und 3) der hier betrachteten Zufallsvariablen mit ihren Wahrscheinlichkeiten und summiert die Produkte auf, erhält man das arithmetische Mittel, das in diesem Zusammenhang auch *Erwartungswert* genannt wird:

Erwartungswert $E(X) = 0*1/8 + 1*3/8 + 2*3/8 + 3*1/8 = 12/8 = 1,5$

Dieser Wert besagt, dass im Schnitt beim dreifachen Münzwurf anderthalb Mal „Kopf“ zu erwarten ist.

Entsprechend erhält man die *Varianz*, wenn man die quadrierten Abweichungen der Merkmalswerte vom Erwartungswert mit den Wahrscheinlichkeiten gewichtet und wieder aufaddiert:

$$\text{Varianz} = VAR(X) =$$
$$= (0-1,5)^2*1/8 + (1-1,5)^2*3/8 + (2-1,5)^2*3/8 + (3-1,5)^2*1/8 =$$
$$= 2,25/8 + 0,75/8 + 0,75/8 + 2,25/8 = 6/8 = 0,75$$

Zu einer stetigen Wahrscheinlichkeitsverteilung gelangt man, wenn man folgende Überlegung anstellt: Stellen Sie sich vor, Sie hätten eine große Zahl von Körpergrößenangaben erfasst. Stellt man diese in einer klassifizierten Häufigkeitsverteilung dar, könnte sich das folgende Bild ergeben:

Größenklasse	Prozent
150 bis unter 160	5
160 bis unter 170	15
170 bis unter 180	33
usw.	usw.

15% der befragten Personen waren also zwischen 160 bis unter 170 cm groß. Man kann deshalb sagen, dass die Wahrscheinlichkeit dafür, dass **eine** zufällig ausgewählte Person zwischen 160 bis unter 170 cm groß ist, 0,15 oder 15% beträgt.

Grafisch könnte diese Häufigkeitsverteilung als Histogramm dargestellt werden (eine spezielle Variante des Balkendiagramms), in dem eine bestimmte Zahl (entsprechend der Klassenanzahl) von nahtlos aneinanderstoßenden Balkenflächen die relativen Häufigkeiten repräsentiert. Die Summe all dieser Balkenflächen ist 1 oder 100%. Wenn man nun, um etwa eine detailreichere Darstellung zu erhalten, alle Klassen noch einmal halbiert und die Urliste der Daten neu auszählt, könnte sich beispielsweise folgendes Bild ergeben:

Größenklasse	Prozent
150 bis unter 155	2
155 bis unter 160	3
160 bis unter 165	8
165 bis unter 170	7
usw.	usw.

Jetzt kann man z.B. sagen, dass eine Wahrscheinlichkeit von 7% dafür besteht, dass eine zufällig ausgewählte Person zwischen 165 bis unter 170 cm groß ist.

Das Histogramm hätte jetzt doppelt so viele, halb so breite Klassen, wie das, das aufgrund der ersten Tabelle gezeichnet werden kann. Die Oberkante aller Rechtecke wird eine Treppenlinie bilden, wobei die Zahl der „Stufen“ im Vergleich zum ersten Fall doppelt so groß ist, die „Stufenhöhen“ hingegen sind im Schnitt nur halb so groß.

Wenn man nun die Klassenbreiten gedanklich immer weiter verkleinert, mithin die Zahl der Klassen ständig erhöht, so geht ihre Anzahl gegen unendlich und die Klassenbreiten gehen gegen null. Die Zahl der „Stufen“ geht damit auch gegen unendlich, die „Stufenhöhen“ gehen gegen null. Damit nähert sich die „Stufenlinie“ der Gestalt einer gekrümmten Kurve

(siehe Abbildung 8.2). In diesem gedanklichen Fall gehen die Balkenflächen ebenfalls gegen null, gleichwohl bleibt die Summe aller Balkenflächen, also die Gesamtfläche unter der „Stufenlinie“ und im Grenzfall unter der gekrümmten Linie gleich 1 (100%).

Weil die Balkenflächen gegen null gehen, gehen auch die Wahrscheinlichkeiten gegen null. Die Wahrscheinlichkeit dafür, dass eine bestimmte Person zwischen 161,2 bis unter 161,3 cm groß ist, ist schon sehr, sehr klein. Aber selbst dann, wenn alle Balkenflächen gegen null konvergieren, sind Wahrscheinlichkeitsaussagen möglich: Die Wahrscheinlichkeit dafür, dass eine bestimmte Person zwischen 165 bis unter 170 cm groß ist, wird dann gegeben durch die Fläche unter der gekrümmten Kurve im Abszissenbereich zwischen 165 und 170 cm.

Eine Verteilung, wie sie in Abbildung 8.2 dargestellt ist, wird *Dichtefunktion* genannt und mit f(x) bezeichnet.

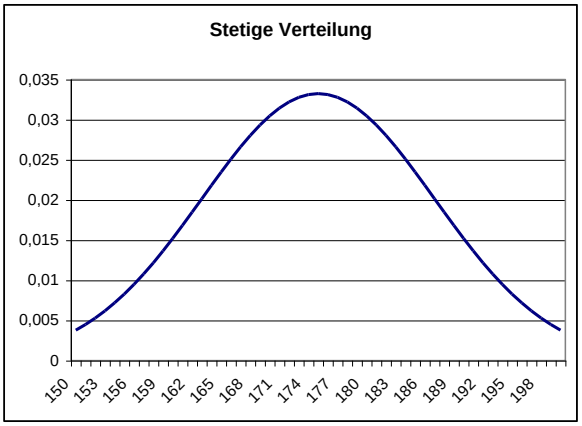


Abb. 8.2 : Stetige Wahrscheinlichkeitsverteilung

8.4 Die Normalverteilung

Eine der wichtigsten stetigen Verteilungen, die Ihnen im Folgenden noch oft begegnen wird, ist die Gauß’sche Normalverteilung. Sie spielt in der Statistik eine außerordentlich bedeutsame Rolle und sieht in der grafischen Darstellung so aus, wie es die Abbildung 8.2 zeigt. Ihre Gestalt kann folgendermaßen beschrieben werden:

1. Die Verteilung ist symmetrisch.
2. Sie nähert sich rechts und links asymptotisch der X-Achse, d.h. sie überdeckt einen Bereich von -unendlich bis +unendlich (auch wenn es keine solchen Körpergrößen in der Realität gibt).
3. Sie hat ein Maximum im Punkt auf der X-Achse, der durch die Symmetrieachse gegeben ist (=Mittelwert der Verteilung = E(X) =μ)
4. Sie hat zwei Wendepunkte in gleicher Entfernung links und rechts von der Symmetrieachse (bei μ-σ und bei μ+σ), wobei mit σ die Standardabweichung bezeichnet wird.

Die Lage einer bestimmten Normalverteilung im Achsenkreuz hängt von zwei Größen ab, nämlich von ihrem arithmetischen Mittel μ und von ihrer Standardabweichung σ. In der Grafik der Abbildung 8.2 liegt eine Normalverteilung vor mit dem Mittelwert 175 (cm) und der Standardabweichung 10 cm.

Es wird deutlich, dass es unendlich viele verschiedene Normalverteilungen gibt, je nachdem, welcher Mittelwert und welche Standardabweichung gegeben ist. Die spezielle Normalverteilung mit dem Mittelwert 0 und der Standardabweichung 1 wird *Standardnormalverteilung* genannt.

Jede normalverteilte Zufallsvariable kann in eine Standardnormalvariable transformiert (standardisiert) werden, wenn man von jedem interessierenden X-Wert den Mittelwert abzieht und diese Differenz durch die Standardabweichung dividiert. So führt z.B. die Körpergröße 180 cm zu folgendem standardisierten Wert:

$$k = (180 - 175)/10 = 0{,}5$$

Ist die Variable X normalverteilt, so ist die Variable K standardnormalverteilt. Die Fläche unter der Normalverteilung zwischen $x_1 = 180$ und $x_2 = 190$ cm entspricht der Fläche unter der Standardnormalverteilung zwischen $k_1 = 0{,}5$ und $k_2 = 1{,}5$. Somit können Wahrscheinlichkeiten für eine normalverteilte Zufallsvariable, gleichgültig, welche Normalverteilung im Einzelfall vorliegt, über die Standardnormalverteilung, deren Flächenbereiche tabelliert vorliegen, bestimmt werden.

Wichtig ist, dass bei einer stetigen Wahrscheinlichkeitsverteilung, so auch bei der Normalverteilung, wie schon erwähnt, Wahrscheinlichkeiten durch Flächenbereiche angegeben werden; die Ordinaten unter der Dichtekurve sind hingegen keine Wahrscheinlichkeiten, sondern sie werden als *Wahrscheinlichkeitsdichten* bezeichnet.

Wie man Flächenbereiche (Wahrscheinlichkeiten) der Normalverteilung berechnet, bzw. wie man die Tabellen der Standardnormalverteilung nutzt, braucht Sie nicht zu kümmern, denn SPSS berechnet die entsprechenden Wahrscheinlichkeiten bei Bedarf automatisch.

Die überragende praktische Bedeutung der Normalverteilung hat mit dem Konzept der Zufallsstichproben zu tun und lässt sich aufgrund der folgenden Überlegung illustrieren:

Stellen Sie sich vor, Sie ziehen eine Zufallsstichprobe vom Umfang n = 100, befragen also z.B. 100 zufällig ausgewählte erwachsene Personen nach ihrer Körpergröße, und berechnen aus den Werten dieser Zufallsstichprobe das arithmetische Mittel und die Standardabweichung. Es möge sich ergeben:

$\bar{x}$ = Arithmetisches Mittel = 175 (cm)

s = Standardabweichung = 10 (cm)

Hinweis:

In Zufallsstichproben wird das arithmetische Mittel mit $\bar{x}$ (zu lesen als „x-quer“), die Standardabweichung mit s bezeichnet, während die entsprechenden Grundgesamtheitsparameter mit μ und σ bezeichnet werden.

Offenkundig ist nun das arithmetische Mittel Ausprägung einer Zufallsvariablen, denn es hängt ja vom Zufall ab, welche 100 Personen in die Stichprobe gelangen; damit hängen die 100 Körpergrößen, und damit auch das daraus berechnete arithmetische Mittel (und auch die Standardabweichung), vom Zufall ab. Demnach könnten Sie, schon bevor die Stichprobe tatsächlich gezogen wird, eine Frage formulieren, etwa der folgenden Art:

Wie groß ist die Wahrscheinlichkeit, dass in einer Zufallsstichprobe vom Umfang n = 100 ein arithmetisches Mittel (Durchschnitt der erfragten Körpergrößen) auftaucht, das zwischen 174 und 176 cm liegt? Diese Frage kann mit der zuständigen Wahrscheinlichkeitsverteilung beantwortet werden – und diese ist die Gauß'sche Normalverteilung. Zur Begründung verweisen wir auf das *Zentrale Grenzwerttheorem von Laplace und Liapunoff* (einer der wichtigsten Lehrsätze der mathematischen Statistik), das folgende Aussage macht:

Das arithmetische Mittel aus einer Zufallsstichprobe, in seiner Eigenschaft als Zufallsvariable, folgt näherungsweise der Gauß'schen Normalverteilung – und zwar praktisch unabhängig davon, wie die interessierende Untersuchungsvariable (Variable „Körpergröße") in der Grundgesamtheit selbst verteilt ist.

Diese Näherung ist umso besser, je größer die Zufallsstichprobe ist, um die es geht. Allerdings ist schon ab n = 30 die Näherung so gut, dass tatsächlich mit der Normalverteilung gearbeitet werden kann. Aber mit welcher der unendlich vielen Normalverteilungen? Auch diese Frage beantwortet die mathematische Statistik: Zuständig ist die Normalverteilung, deren Mittelwert dem der Grundgesamtheit, aus der die Zufallsstichprobe gezogen wurde, entspricht; ihre Standardabweichung ergibt sich aus derjenigen der Grundgesamtheit, geteilt durch die Wurzel aus dem Stichprobenumfang (das ist der *Standardfehler* oder *Stichprobenfehler*, dem Sie schon an anderer Stelle begegnet sind). Da die Stichprobenstandardabweichung, so wie sie von SPSS berechnet wird, eine erwartungstreue Schätzung für die Grundgesamtheitsstandardabweichung ist (auch darüber wurde schon gesprochen), kann sie benutzt werden, wenn letztere unbekannt ist, was in der Regel der Fall sein dürfte.

8.5 Hypothesentest

Wir hatten oben behauptet, dass am Anfang jeder empirischen Untersuchung eine Hypothese steht. Eine solche Hypothese (sie wird von den Statistikern *Nullhypothese* genannt) könnte folgendermaßen lauten:

Die Durchschnittsgröße der erwachsenen Deutschen liegt bei 173 cm.

Um diese Hypothese im Licht empirischer Befunde zu überprüfen, ziehen Sie eine Zufallsstichprobe z. B. vom Umfang n = 100, in der sich ein Mittelwert von 175 cm bei einer Standardabweichung von 10 cm ergibt. Widerspricht dieser Befund der Nullhypothese, oder ist er mit ihr (noch) vereinbar? Anders gefragt: Ist die Abweichung von 2 cm zwischen Stichprobenbefund und Nullhypothesenwert noch als im Rahmen des Zufälligen erlaubt anzusehen, oder ist sie, wie man sagt, *signifikant*? Mit dem Stichwort „signifikant" meint man eine Abweichung, die, wenn die Nullhypothese wirklich zutreffen würde (was Sie aber leider nicht wissen), eigentlich gar nicht auftreten dürfte.

Ist die Abweichung signifikant, verwerfen Sie die Nullhypothese (Sie lehnen sie ab), ist sie zufällig, gilt die Nullhypothese als bestätigt.

Die Entscheidung über Bestätigung oder Verwerfung der Nullhypothese kommt nun aufgrund der folgenden Überlegung zustande: Man fragt nach der Wahrscheinlichkeit dafür, dass der beobachtete Stichprobenmittelwert, oder ein noch weiter von der Nullhypothese abweichenderer Wert, zu erwarten ist – und diese Wahrscheinlichkeit berechnet man mit der zuständigen Wahrscheinlichkeitsverteilung, die gemäß des Zentralen Grenzwerttheorems eine Gauß'sche Normalverteilung ist. Ist die zu berechnende Wahrscheinlichkeit sehr klein, nehmen Sie dieses „unwahrscheinliche" Ergebnis zum Anlass, die Nullhypothese zu verwerfen. Ist der Stichprobenbefund hingegen hochwahrscheinlich, gilt die Nullhypothese als bestätigt.

Offensichtlich ist es erforderlich, zunächst darüber zu entscheiden, was eine kleine, und was eine hohe Wahrscheinlichkeit ist. In der Praxis der Stichprobenstatistik hat man sich angewöhnt, als kleine Wahrscheinlichkeit z.B. 5% vorzugeben. Man nennt diese „Grenzwahrscheinlichkeit" das *Signifikanzniveau* (nicht selten wird auch mit 10% oder 1% oder mit 0,5% gearbeitet).

Kehren Sie zu der obigen Frage zurück: Wie groß ist die Wahrscheinlichkeit dafür, dass der beobachtete Stichprobenmittelwert, oder ein noch weiter von der Nullhypothese abweichenderer Wert, zu erwarten ist? Diese Wahrscheinlichkeit nennt man *Überschreitungswahrscheinlichkeit* oder *rechnerische Signifikanz*, und diese wird Ihnen in vielen der folgenden SPSS-Ergebnisausgaben unter dem Stichwort „*Signifikanz*“ begegnen. Wie groß ist die Überschreitungswahrscheinlichkeit in unserem Beispiel?

Die Antwort gibt die zuständige Wahrscheinlichkeitsverteilung (oder SPSS), also die Normalverteilung mit dem Mittelwert 173 und der Standardabweichung, die sich aus der der Stichprobe geteilt durch die Wurzel aus dem Stichprobenumfang ergibt (diese Berechnung führt zu dem Wert 1). Berechnet man die Fläche unter dieser speziellen Normalverteilung rechts vom Punkt 175, hat man die gesuchte Überschreitungswahrscheinlichkeit. Wir wollen die notwendigen Rechenschritte hier nicht vorführen (SPSS macht das ja später automatisch), sondern geben nur das Ergebnis bekannt: Die Überschreitungswahrscheinlichkeit beträgt 0,023 oder 2,3%. Dieser Wert besagt: Es besteht eine Wahrscheinlichkeit von nur 2,3% dafür, dass in einer Zufallsstichprobe vom Umfang n = 100 ein Mittelwert von 175 cm oder größer auftritt, wenn in Wahrheit der Grundgesamtheitsmittelwert, wie die Nullhypothese behauptet, bei 173 cm liegen sollte. Wenn man von einem Signifikanzniveau von 5% ausgeht, bedeutet dieser Befund, dass die Überschreitungswahrscheinlichkeit klein genug ist, um die Nullhypothese zu verwerfen.

Das Verwerfen der Nullhypothese ist natürlich mit dem Risiko verbunden, zu Unrecht zu verwerfen. Diese Fehlentscheidungswahrscheinlichkeit ist genauso groß wie das vorgegebene Signifikanzniveau. Nicht zuletzt deshalb wird dafür eine möglichst kleine Wahrscheinlichkeit vorgegeben.

Nun gilt aber der folgende Zusammenhang: Je kleiner das Signifikanzniveau ist, also je kleiner die Wahrscheinlichkeit des Fehlers vom Typ I ist (siehe Abschnitt 8.1), desto größer muss eine beobachtete Abweichung des Stichprobenbefundes von der Behauptung der Nullhypothese ausfallen, bevor diese tatsächlich verworfen werden kann. Logischerweise wächst damit die Gefahr des Fehlers vom Typ II (Bestätigung einer an sich falschen Nullhypothese) – beide Fehlertypen stehen in einem (indirekten) wechselseitigen Verhältnis zueinander (die Wahrscheinlichkeit eines Fehlers vom Typ II kann übrigens nur berechnet werden, wenn der Nullhypothese eine Alternativhypothese gegenübergestellt wird; wir wollen darauf hier nicht eingehen und verweisen auf die statistische Fachliteratur (siehe z.B. *Tiede/Voß: Schließen mit Statistik – Verstehen, Verlag Oldenbourg, 2000*).

Es gibt allerdings einen Weg, **beide** Fehlerwahrscheinlichkeiten zu verringern, nämlich die Erhöhung des Stichprobenumfangs n.

Was wir hier am Beispiel des Stichprobenmittelwertes erläutert haben, gilt auch für andere Stichprobenparameter, z.B. für einen Stichprobenanteilswert.

Auch dazu ein Beispiel: Die Nullhypothese lautet, dass der SPD-Wähleranteil bei allen Wahlberechtigten bei 42% liegt. In einer Zufallsstichprobe vom Umfang n = 203 ergibt sich ein SPD-Wähleranteil von 39%. Kann aufgrund dieses Befundes die Nullhypothese verworfen werden?

Zur Beantwortung dieser Frage berechnen Sie die Überschreitungswahrscheinlichkeit für 39% oder weniger – Gültigkeit der Nullhypothese vorausgesetzt.

Zuständig ist wieder eine Normalverteilung, denn der Zufallsstichprobenanteilswert ist Ausprägung einer ebenfalls näherungsweise normalverteilten Zufallsvariablen, mit dem Mittelwert 42 (%) und der Standardabweichung 3,46 (%). Diese Standardabweichung ergibt sich, wie man den Lehrbüchern der Wahrscheinlichkeitsstatistik entnehmen kann, als Wurzel aus (42*(100-42)/203) = 3,46. Die Überschreitungswahrscheinlichkeit für 39%

ergibt sich bei einer solchen Normalverteilung zu 19,3% (die konkrete Berechnung dieses Wertes führen wir auch hier nicht vor). Dies bedeutet, dass bei einem Signifikanzniveau von 5% die Nullhypothese nicht verworfen wird; der Stichprobenbefund (39%) weicht nur zufällig vom Wert ab, den die Nullhypothese behauptet (42%).

Entsprechendes gilt auch für andere Stichprobenparameter (man nennt alle diese Hypothesentestverfahren deshalb „*parametrische Signifikanztests*"), wobei allerdings in Einzelfällen auch mit anderen als mit der Gauß'schen Normalverteilung gearbeitet werden muss. Wichtig sind in diesem Zusammenhang beispielsweise die t-Verteilung oder die F-Verteilung, die Sie in folgenden Kapiteln kennen lernen werden. Nur am Rande sei darauf aufmerksam gemacht, dass bei großen Zufallsstichproben die t- und die F-Verteilung wiederum durch die Gauß'sche Normalverteilung ersetzt werden können. Auch dies ist wieder ein Beleg für die überaus große Bedeutung dieser speziellen Wahrscheinlichkeitsverteilung.

Insbesondere dann, wenn die Stichprobenumfänge klein sind, kann nicht mehr die Gauß'sche Normalverteilung verwendet werden, weil dann die Aussage des Zentralen Grenzwerttheorems nicht mehr zutrifft, sondern dann setzt SPSS spezielle Verteilungen ein, auf die wir, wenn es so weit ist, gesondert aufmerksam machen.

8.6 Konfidenzintervalle

Mit dem gleichen Instrumentarium, mit dem der parametrische Hypothesentest durchgeführt wurde, können auch sog. Konfidenzintervalle bestimmt werden. Betrachten Sie dazu das folgende Beispiel:

In einer Zufallsstichprobe vom Umfang n = 400 möge sich als Mittelwert erfasster Körpergrößen der Wert 176 cm ergeben haben, bei einer Standardabweichung von 12 cm. In welchem Bereich kann der wahre Mittelwert der Grundgesamtheit, aus der diese Zufallsstichprobe gezogen wurde, erwartet werden?

Um diese Frage zu beantworten, gehen Sie wieder von der Wahrscheinlichkeitsverteilung aus, der der Zufallsstichprobenmittelwert in seiner Eigenschaft als Zufallsvariable folgt. Dies ist wieder eine Normalverteilung, deren Streuung sich ungefähr aus der Stichprobenstreuung s geteilt durch die Wurzel aus dem Stichprobenumfang n ergibt, was im gegebenen Zahlenbeispiel zum Wert 0,6 (cm) führt.

Wenn wir Ihnen nun sagen, dass 95% der Fläche sich unter einer Normalverteilung im Bereich

$$\mu - 1{,}96 * \text{Streuung und } \mu + 1{,}96 * \text{Streuung}$$

finden, können Sie den folgenden Schluss ziehen:

Der Stichprobenmittelwert (176 cm) liegt so, dass das darum aufgebaute Intervall von 176-1,96*0,6 bis 176+1,96*0,6, also von 174,834 bis 177,176 den wahren Grundgesamtheitsmittelwert mit einer Wahrscheinlichkeit von 95% erfasst. Anders formuliert: Ausgehend vom Stichprobenbefund kann der unbekannte Mittelwert der Grundgesamtheit mit einem Vertrauen von 95% zwischen 174,834 und 177,176 cm erwartet werden.

Ein solches Schätzintervall (man spricht in diesem Zusammenhang von der sog. *Intervallschätzung*) wird auch *Konfidenzintervall* (*Vertrauensintervall*) genannt. Es wird Ihnen später noch des Öfteren begegnen.

8.7 Nichtparametrischer Test

Im Abschnitt 8.5 wurde das Grundmuster des parametrischen Signifikanztests besprochen. Es gibt aber auch Aufgabenstellungen in der angewandten Statistik, die zu den sog. nichtparametrischen Tests führen. Ein häufig auftauchendes Beispiel soll hier vorgestellt werden, der *Chi-Quadrat-Unabhängigkeitstest*.

Erinnern Sie sich an das Beispiel aus Kapitel 7, wo die geschlechtsspezifischen Unterschiede in der Parteienpräferenz dadurch illustriert wurden, dass von SPSS eine bivariate Verteilung erzeugt wurde, in der die beiden Variablen „Geschlecht“ und „bevorzugte politische Partei“ auftauchten. Diese bivariate Verteilung haben wir in Abbildung 8.3 noch einmal vorgestellt.

Partei * Geschlecht Kreuztabelle

			Geschlecht		
			männlich	weiblich	Gesamt
Partei	CDU/CSU	Anzahl	43	40	83
		Erwartete Anzahl	41,1	41,9	83,0
	SPD	Anzahl	44	34	78
		Erwartete Anzahl	38,6	39,4	78,0
	F.D.P.	Anzahl	5	6	11
		Erwartete Anzahl	5,4	5,6	11,0
	Die Grünen	Anzahl	6	15	21
		Erwartete Anzahl	10,4	10,6	21,0
	Sonstige	Anzahl	1	6	7
		Erwartete Anzahl	3,5	3,5	7,0
Gesamt		Anzahl	99	101	200
		Erwartete Anzahl	99,0	101,0	200,0

Abb. 8.3: Geschlecht und bevorzugte politische Partei

In dieser Tabelle wurden auch die Erwartungswerte ausgegeben, d.h. diejenigen Zellenbesetzungen, die zu erwarten wären, wenn Unabhängigkeit zwischen „Geschlecht“ und „bevorzugter politischer Partei“ bestünde (wie man dazu gelangt, wurde in Abschnitt 7.2 besprochen). Damit verfügen Sie über die Ausgangsdaten, um folgende Nullhypothese überprüfen zu können:

Zwischen beiden Variablen besteht Unabhängigkeit.

Der nun anstehende Hypothesentest läuft gemäß der folgenden Überlegungen ab: Je weiter die beobachteten Häufigkeiten von denen abweichen, die bei Gültigkeit der Nullhypothese, also bei Unabhängigkeit der beiden Variablen voneinander, zu erwarten wären, desto eher ist man geneigt, die Nullhypothese zu verwerfen.

Um zur Hypothesenentscheidung zu gelangen, sind jetzt die folgenden Rechenschritte erforderlich (SPSS wird das später alles automatisch erledigen):

1. Man notiert alle Abweichungen zwischen beobachteten und erwarteten Werten.
2. Diese Abweichungen werden quadriert, um zu verhindern, dass sich positive und negative Abweichungen gegenseitig ausgleichen.
3. Die quadrierten Abweichungen werden durch die Erwartungswerte dividiert, um zu erreichen, dass eine bestimmte Abweichung bei kleinem Erwartungswert „ernster“ genommen wird, als bei einem großen Erwartungswert. Sie werden gewissermaßen

relativiert. Zudem wird somit erreicht, dass die Zahl der Beobachtungen keinen direkten Einfluss auf die Testentscheidung ausübt.

4. Alle diese quadrierten, relativierten Abweichungen werden addiert, so dass sich eine einzige Zahl ergibt.

Diese sich zuletzt ergebende Zahl wird „*Pearson'sche Prüfgröße*" genannt und üblicherweise mit U bezeichnet. Sie wird umso größer sein, je größer die ursprünglichen Abweichungen zwischen beobachteten und theoretisch zu erwartenden Häufigkeiten waren. Je größer also U ist, desto eher werden Sie die Nullhypothese der Unabhängigkeit zwischen beiden Untersuchungsvariablen verwerfen.

Die Hypothesenentscheidung kommt nun aufgrund der gleichen Überlegung zustande, wie wir sie im Abschnitt 8.5 vergleichsweise ausführlich vorgestellt haben: Man fragt danach, wie groß die Wahrscheinlichkeit für den betrachteten U-Wert oder für einen noch weiter von der Nullhypothese abweichenderen U-Wert, d.h. für einen noch größeren U-Wert, ist. Dazu benötigen Sie die zuständige Wahrscheinlichkeitsverteilung – und dies ist nicht die Normalverteilung, sondern die sog. *Chi-Quadrat-Verteilung* (daher der Name für diesen Test). Die Gestalt der zuständigen Chi-Quadrat-Verteilung (auch davon gibt es unendlich viele), hängt von der Zahl ihrer sog. *Freiheitsgrade* ab. Diese Anzahl der Freiheitsgrade entspricht der Anzahl der „frei beweglichen" (variablen) Summanden, die die Prüfgröße U bilden. Auf den Begriff der Freiheitsgrade kommen wir noch einmal ausführlicher in einem späteren Kapitel zu sprechen.

Um die Berechnung einmal vorzuführen, betrachten wir ein kleineres Beispiel: Es soll die Hypothese geprüft werden, dass es keinen Zusammenhang zwischen Geschlecht und der Antwort auf die Frage gibt, ob man Kanzlerin Merkel für kompetent halte. Der Ausgangsdatenbestand, der auf der Grundlage einer Zufallsstichprobe vom Umfang n = 200 gewonnen wurde, möge sich wie folgt darstellen:

Geschlecht	männlich	weiblich	Summe
Kanzlerin Merkel			
kompetent	55	65	120
nicht kompetent	50	30	80
Summe	105	95	200

Bei Unabhängigkeit der beiden Variablen voneinander (Aussage der Nullhypothese), müssten sich die 120 Personen, die die Kanzlerin für kompetent halten, im Verhältnis 105:95 auf die beiden Geschlechter aufteilen.

Entsprechendes ergibt sich für die 80 Personen, die sie nicht für kompetent halten. Auf diese Weise erhalten wir die Erwartungswerte der Spalte 2 der folgenden Tabelle.

Stellt man nun beobachtete und erwartete Häufigkeiten in der folgenden Arbeitstabelle einander gegenüber, kann U ausgerechnet werden:

Beobachtet B	*Erwartet E*	*B-E*	$(B-E)^2$	$(B-E)^2/E$
55	63	-8	64	1,016
65	57	+8	64	1,103
50	42	+8	64	1,524
30	38	-8	64	1,684
200	200	0		U = 5,327

Die Testentscheidung reduziert sich jetzt auf die Frage, wie wahrscheinlich ein U-Wert von 5,327 oder größer ist (Überschreitungswahrscheinlichkeit). Diese Frage beantwortet die Chi-Quadrat-Verteilung mit (2-1)*(2-1) = 1 Freiheitsgrad (nur einer der vier Summanden,

aus denen U gebildet wird, ist „frei beweglich“; liegt er fest, sind die drei anderen Summanden, wegen der gegebenen Randverteilungen, nicht mehr „frei beweglich“). Entsprechenden Tabellen kann man den Rückweisungspunkt entnehmen, der in diesem Fall bei einem Signifikanzniveau von 5% bei 3,84 liegt. Es muss also die Nullhypothese der Unabhängigkeit zwischen den beiden betrachteten Variablen auf der Grundlage der erhobenen Zufallsstichprobe verworfen werden, weil U=5,327 > 3,84. Die Wahrscheinlichkeit, dass der gefundene U-Wert überschritten wird, ist also kleiner als 5%.

Es gibt noch eine ganze Reihe weiterer nichtparametrischer Testverfahren, die zum Teil auf der Chi-Quadrat-Verteilung aufbauen, zum Teil aber auchandere Verteilungen nutzen. Wir kommen darauf bei den konkreten SPSS-Anwendungen der folgenden Kapitel zu sprechen.

9 Regressionsrechnung

9.1 Zielsetzungen und statistische Methoden

In diesem Kapitel wenden wir uns wieder konkreten statistischen Anwendungen zu und knüpfen an der Besprechung bivariater Verteilungen des Kapitels 7 an. Dort wurde schon darauf aufmerksam gemacht, dass uns nicht nur die tabellarische oder grafische Präsentation einer bivariaten Verteilung interessiert, sondern auch die Quantifizierung der Zusammenhänge zwischen beiden betrachteten Variablen mit Hilfe geeigneter Maßzahlen. Dabei müssen zwei Aufgaben voneinander unterschieden werden:

1. Die zusammenfassend beschreibende Charakterisierung des Zusammenhangs.
2. Die Bemessung der Stärke des Zusammenhangs.

Für die erste Aufgabe ist die Regressionsrechnung zuständig, mit der wir uns in diesem Kapitel befassen, für die zweite die Korrelationsrechnung, die im folgenden Kapitel betrachtet wird.

> **Hinweis:**
>
> Die Regressionsrechnung ist ein außerordentlich wichtiger Bereich der statistischen Methoden, weil sie Grundlage für viele weiterführende und anspruchsvollere Analyseverfahren ist. Deshalb sollte diesem Kapitel besondere Aufmerksamkeit gewidmet werden.

Zur Demonstration der Vorgehensweise bei der Regressionsrechnung greifen wir auf das Beispiel aus Kapitel 7 zurück, wo für zehn zufällig ausgewählte Schüler die Fehler in einem Deutschdiktat und in einer Mathematik-Klassenarbeit erfasst wurden. Dieser Datenbestand ist in Abbildung 9.1 noch einmal vorgestellt. Er soll als Ausgangsbasis für die weiteren Darlegungen benutzt werden. Bedenken Sie aber in diesem Zusammenhang, dass ein Datenbestand mit nur zehn Beobachtungen zur einer recht „dürftigen“ Punktwolke führt. Für die Erörterung der Methoden, um die es hier geht, bleibt es sich aber gleich, ob man mit zehn, mit hundert oder mit tausend Beobachtungen arbeitet.

	deutsch	mathe
1	5	4
2	2	7
3	7	2
4	0	7
5	3	5
6	6	3
7	8	2
8	2	5
9	6	1
10	3	6

Abb. 9.1: Ausgangsdatenbestand

Grafisch stellt sich dieser kleine Datenbestand als Punktwolke wie in Abbildung 9.2 dar.

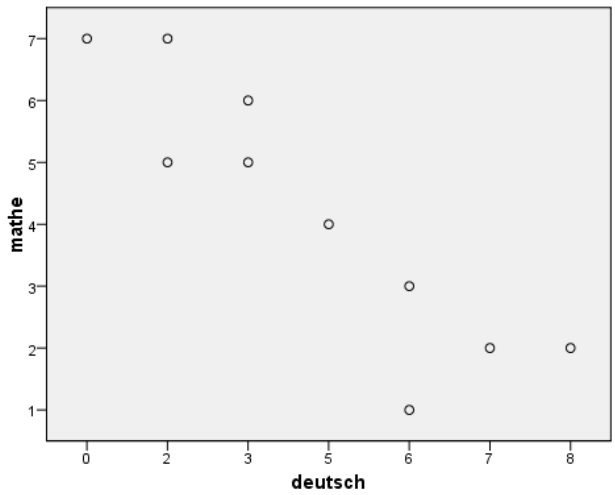


Abb. 9.2: Punktwolke

Die zusammenfassende Beschreibung einer solchen Punktwolke geht so vor sich, dass in die Punktwolke eine mathematische Funktion hineingelegt wird, die sich diesen Punkten möglichst gut anpasst soll. Im einfachsten (und häufigsten) Fall ist diese Funktion eine Gerade, deren Lage im Achsenkreuz bekanntlich vom Ordinatenabschnitt a und vom Steigungswinkel b abhängt. Dies verdeutlicht die allgemein gehaltene Skizze der Abbildung 9.3.

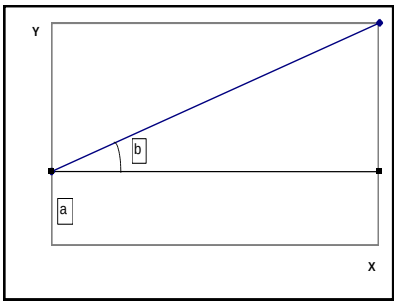


Abb. 9.3: Lineare Funktion

Wenn eine solche Gerade in eine Punktwolke hineingelegt wird, dann gehören zu jedem X-Wert (Fehleranzahl in Deutsch), d.h. zu jedem Schüler, zwei Y-Werte, nämlich der beobachtete Y-Wert (Fehleranzahl in Mathematik) und ein auf der Geraden liegender Yt-Wert (theoretischer Y-Wert). Er gibt an, welche Fehleranzahl der betreffende Schüler in Mathematik erwarten lässt, wenn die Regressionsfunktion den interessierenden Zusammenhang zutreffend beschreibt.

Die Aufgabe, die Regressionsgerade optimal in die Punktwolke hineinzulegen, also ihre Parameter a und b so zu bestimmen, dass die Gerade optimal durch die gegebenen Punkte hindurchläuft, kann nun umformuliert werden: Die Parameter a und b (siehe Abbildung 9.3) sind so zu bestimmen, dass die Summe der quadrierten Abweichungen zwischen den Y-Werten und den Yt-Werten minimiert wird. Diese Rechenvorschrift nennt man die „*Me-*

thode der kleinsten Quadrate“ (*LS-Methode* = Least-Squares-Methode), oder besser ‚*Methode der kleinsten Quadratsumme*‘. Sie führt zu zwei Berechnungsformeln, eine für b und eine für a.

In unserem Beispiel ergibt sich als Ordinatenabschnitt a=7,4 und als Steigung b=-0,76. Die Regressionsgerade lautet also:

$$yt_i = 7{,}4 - 0{,}76 * x_i$$

Sie ist ein zusammenfassend charakterisierender Ausdruck für den gegebenen bivariaten Datenbestand und besagt folgendes:

Jemand, der in Deutsch (Variable X) keinen Fehler macht (x = 0), lässt einen Y-Wert von a = 7,4 erwarten, also 7,4 Fehler in Mathematik. Wer z.B. 5 Fehler in Deutsch macht, lässt 7,4-0,76*5 = 7,4-3,8 = 3,6 Fehler in Mathematik erwarten. Bei einer Zunahme der Fehleranzahl in Deutsch um einen Fehler, geht tendenziell die Fehlerzahl in der Mathematikarbeit um 0,76 Fehler zurück. Die zur Beschreibung dienende Regressionsgerade dient also auch prognostischen Zwecken. Sie zeigt, welcher Y-Wert bei gegebenem X-Wert zu erwarten ist, wenn die Regressionsgerade den interessierenden Zusammenhang zutreffend beschreibt.

9.2 Lineare Regression

Das kleine Schülerbeispiel der Abbildung 9.1 soll jetzt dazu verwendet werden, um mit SPSS die Parameter der linearen Regressionsfunktion auszurechnen. Dazu ist folgendermaßen vorzugehen:

1. Wählen Sie nach Eingabe der Ausgangsdaten Menü ANALYSIEREN/REGRESSION/ LINEAR...

Sie gelangen ins Dialogfenster der Abbildung 9.4.

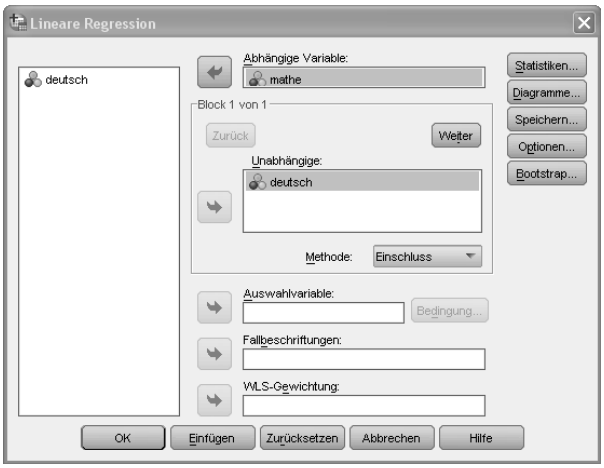


Abb. 9.4: Menü ANALYSIEREN/REGRESSION/LINEAR...

Im Fenster der Abbildung 9.4 tätigen Sie die folgenden Einstellungen:

2. Übertragen Sie die Variable „Mathe“ ins Feld unter ABHÄNGIGE VARIABLE:.

3. Übertragen Sie die Variable „Deutsch“ ins Feld unter UNABHÄNGIGE VARIABLE(N):.
4. Klicken Sie auf OK.

In der ersten Tabelle der Abbildung 9.5 sehen Sie, mit welchen Variablen SPSS gearbeitet hat. Als abhängige Variable wurde „Mathe“ verwendet, als unabhängige „Deutsch“.

In der zweiten Tabelle (MODELLZUSAMMENFASSUNG) werden Maßzahlen dargestellt, auf die wir im folgenden Kapitel eingehen werden.

Unter der Überschrift „ANOVA“ (diese Abkürzung steht für die Prozedur „Varianzanalyse“, auf die im Detail in Kapitel 12 (Abschnitt 12.5) eingegangen wird) stellt SPSS weitere Ergebnisse bereit. Auch auf diese Tabelle werden wir erst an anderer Stelle eingehen.

Wichtig für dieses Kapitel ist die Tabelle in Abbildung 9.6 mit der Überschrift „KOEFFIZIENTEN“ (siehe Abbildung 9.6).

Aufgenommene/Entfernte Variablen[a]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	deutsch[b]	.	Einschluß

a. Abhängige Variable: mathe

b. Alle gewünschten Variablen wurden eingegeben.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,912[a]	,831	,810	,937

a. Einflußvariablen : (Konstante), deutsch

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	34,583	1	34,583	39,429	,000[b]
	Nicht standardisierte Residuen	7,017	8	,877		
	Gesamt	41,600	9			

a. Abhängige Variable: mathe

b. Einflußvariablen : (Konstante), deutsch

Abb. 9.5: Regressionsrechnung, Teil 1

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	7,399	,589		12,556	,000
	deutsch	-,762	,121	-,912	-6,279	,000

a. Abhängige Variable: mathe

Abb. 9.6: Regressionsrechnung, Teil 2

Unter RegressionskoeffizientB finden Sie mit dem Wert -0,762 die Steigung der linearen Regressionsfunktion. Darüber steht der Ordinatenabschnitt (7,399). Unter dem Stichwort STANDARDFEHLER finden Sie die beiden Streuungen der Zufallsvariablen.

Erinnern Sie sich bitte an die Ausführungen im Kapitel 8. Dort hatten wir erläutert, dass beispielsweise der Zufallsstichprobenmittelwert Ausprägung einer Zufallsvariablen ist, die ihrerseits einen Mittelwert und eine Standardabweichung (Streuung) hat. Diese Streuung wurde schon in Kapitel 6 (Abschnitt 6.3) mit dem Begriff „Standardfehler" bezeichnet. Was nun für den Zufallsstichprobenmittelwert gilt, gilt auch für andere Parameter, also auch z.B. für die Steigung b und für den Ordinatenabschnitt a, die sich aus einem Zufallsstichprobendatenbestand ergeben. Und deren Standardfehler sind hier angegeben.

Allerdings folgen a und b in ihrer Eigenschaft als Zufallsvariablen nicht der Gauß'schen Normalverteilung, wie etwa der Stichprobenmittelwert, sondern der t-Verteilung (diese braucht aber im Detail nicht besprochen zu werden). Die entsprechenden Werte der t-Variablen sind unter T oben angegeben.

Wenn Sie prüfen wollen, ob die Steigung b = -0,762 signifikant von null abweicht, müssen Sie die zuständige Verteilung, also die t-Verteilung, verwenden, um folgende Frage zu beantworten (diese Frage kennen Sie schon): Wie wahrscheinlich ist es, Gültigkeit der Nullhypothese vorausgesetzt (keine Steigung in der Grundgesamtheit, d.h. b = 0), dass in einer Zufallsstichprobe vom Umfang n = 10 ein b-Wert auftaucht, der -0,762 ist oder noch weiter von null abweicht? Dies ist wieder die Frage nach der Überschreitungswahrscheinlichkeit, und sie wird von SPSS unter dem Stichwort SIGNIFIKANZ mit 0,000 beantwortet. Die Überschreitungswahrscheinlichkeit ist also außerordentlich klein (sie ist nicht null; da aber SPSS nur drei Dezimalstellen ausgibt, sieht es so aus, als ob sie null wäre), kleiner als jedes übliche Signifikanzniveau. Deshalb wird die Hypothese, in der Grundgesamtheit sei die Steigung null, aufgrund des Stichprobenbefundes (b = -0,762) verworfen.

Entsprechendes gilt auch für den Ordinatenabschnitt a, wo die Überschreitungswahrscheinlichkeit ebenfalls mit 0,000 angegeben wird. Die Hypothese, dass der Ordinatenabschnitt in der Grundgesamtheit null sei, wird also ebenfalls verworfen.

Der im mittleren Bereich der oben vorgestellten Ausgabe unter BETA (standardisierte Koeffizienten) auftauchende Wert (-0,912) ist die Steigung, berechnet aus standardisierten Werten. Was unter „*Standardisierung*" zu verstehen ist, wurde schon an anderer Stelle besprochen. Dieser standardisierte Regressionskoeffizient wird später wichtig, weil in dem Fall, dass mehrere Regressionskoeffizienten (mehrere Steigungen) aus Variablen bestimmt werden, die sich auf ganz unterschiedlichem numerischen Niveau bewegen, nur so der Bedeutungsvergleich zwischen mehreren Steigungsangaben möglich wird (siehe dazu Kapitel 11).

Besonders illustrativ ist es, wenn in das Streudiagramm die mit der Methode der kleinsten Quadrate bestimmte lineare Regressionsfunktion eingezeichnet wird. Dies ist mit SPSS auch möglich, wenn Sie das Diagramm über das Menü DIAGRAMME/VERALTETE DIALOGFELDER/STREU-/PUNKT-DIAGRAMM… erzeugt haben. Es sieht dann genauso aus wie in der Abbildung 9.2.

Führen Sie auf diesem Diagramm einen Doppelklick aus, öffnet sich der sog. Diagramm-Editor. Klicken Sie dann mit der rechten Maustaste einen der Punkte an, öffnet sich ein Kontextmenü, in dem Sie die Position HINZUFÜGEN ANPASSUNGSLINIE BEI GESAMTWERT auswählen. Es ergibt sich dann das Diagramm der Abbildung 9.7.

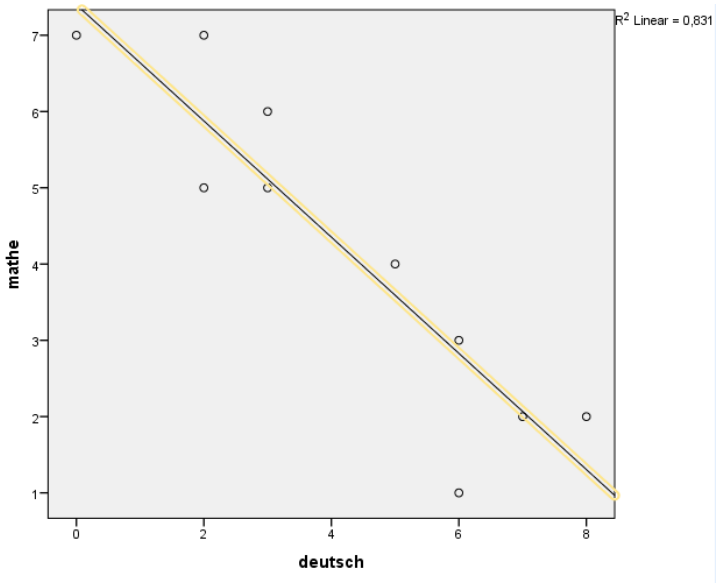


Abb. 9.7: Lineare Regressionsfunktion

Diese Regressionsfunktion zeigt anschaulich den Zusammenhang zwischen den beiden betrachteten Untersuchungsvariablen. Sie erkennen sofort, dass mit zunehmenden Werten der Variablen „Deutsch“ (zunehmende Fehleranzahl), die Fehleranzahlen in Mathematik tendenziell abnehmen. Somit wird auf optischem Wege verdeutlicht, mit welcher Zusammenhangsrichtung – ausgehend von dem gegebenen kleinen Datenbestand – zu rechnen ist.

Natürlich streuen die einzelnen Beobachtungen (die Punkte der Punktwolke) um die Gerade, und es dürfte einleuchten, dass der Zusammenhang zwischen den beiden Untersuchungsvariablen um so stärker ist, je enger die Punkte um die Regressionsgerade herum streuen. Einzelheiten dazu finden sich im folgenden Kapitel, wo über die statistische Zusammenhangsrechnung gesprochen wird.

Hinweis:

Im Diagramm wird jetzt auch rechts oben ausgegeben: R^2 Linear = 0,831. Dieser Wert ist Ihnen schon in der Ausgabe der Abbildung 9.5 unter „Modellzusammenfassung“ begegnet. Wir werden im folgenden Kapitel darauf zu sprechen kommen.

9.3 Vertrauensbereiche

Immer dann, wenn aus Zufallsstichprobendaten Parameterwerte ausgerechnet werden (etwa ein Stichprobenmittelwert oder, wie hier, ein Stichprobenregressionskoeffizient, also z.B. die Steigung der Regressionsgeraden in der Punktwolke der Stichprobendaten), stellt sich die Frage, in welchen Grenzen der entsprechende (unbekannte) Wert der Grundgesamtheit erwartet werden kann. In Kapitel 8 wurde in Abschnitt 8.6 dargelegt, dass mit Hilfe der zuständigen Wahrscheinlichkeitsverteilung solche Bereiche bestimmt werden können. Beim Beispiel zur Regressionsrechnung, von dem wir hier ausgehen, sind dazu die folgenden Schritte erforderlich:

Führen Sie die Regressionsrechnung erneut durch, wie es weiter oben besprochen wurde. ABER: Vor dem abschließenden OK klicken Sie auf die Schaltfläche STATISTIKEN… und Sorgen Sie für Häkchen bei SCHÄTZER und bei KONFIDENZINTERVALLE. Dann WEITER und OK.

SPSS erzeugt jetzt wieder die gleichen Ausgaben, wie sie weiter oben schon vorgestellt wurden. Zusätzlich erhalten Sie jetzt aber unter dem Begriff „95% KONFIDENZINTERVALL FÜR B“ (95%-Vertrauensbereich für B) die folgenden Angaben (gerundet):

6,040	8,758
-1,041	-0,482

Diese Angaben besagen, dass die Regressionssteigung in der Grundgesamtheit mit einem Vertrauen von 95% zwischen -1,04 und -0,48, der Ordinatenabschnitt zwischen 6,04 und 8,76 zu erwarten ist. Dies sind vergleichsweise große Schwankungsbereiche, was damit zu tun hat, dass in unserem Beispiel der Stichprobenumfang mit n = 10 nicht sehr groß ist. Je größer die Stichprobe ist, von der ausgegangen werden kann, desto enger werden derartige Schätzbereiche, d.h. desto präziser werden Ihre Hochrechnungen vom Stichprobenbefund auf die unbekannte Grundgesamtheit.

Wichtiger Hinweis:

Die lineare Regressionsrechnung, wie sie jetzt beschrieben wurde, kann nur bei Vorliegen metrischer Daten zu sinnvollen Ergebnissen führen.

Diese Voraussetzung ist nicht immer erfüllt. Es gibt aber einen Ausweg: Wenn nichtmetrische Variablen in eine Regressionsrechnung eingehen sollen, so ist dies auch möglich, sofern die entsprechenden Variablen dichotom sind, also nur zwei Ausprägungen aufweisen, wie beispielsweise die Variable Geschlecht.

Betrachten Sie einmal den folgenden kleinen Datenbestand:

	Einkommen	Geschlecht
1	3000	0
2	2500	0
3	4000	0
4	3200	0
5	2800	1
6	2500	1
7	2200	1
8	2300	1
9		

Abb. 9.8: Bivariater Datenbestand mit einer nichtmetrischen, dichotomen Variablen

Hier ist nach Einkommen und Geschlecht (0 = männlich; 1 = weiblich) gefragt worden. Eine Regressionsrechnung führt zu den folgenden Befunden:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	3175,000	239,574		13,253	,000
	Geschlecht	-725,000	338,809	-,658	-2,140	,076

a. Abhängige Variable: Einkommen

Abb. 9.9: Ergebnisse der Regressionsrechnung

Wir erhalten hier einen Ordinatenabschnitt von 3175. Dies ist nichts anderes als das Durchschnittseinkommen der männlichen Befragten. Der Steigungswinkel ist -725, und das ist nichts anderes, als der Unterschied zwischen dem Durchschnittseinkommen der Männer und dem der Frauen, wie Sie leicht anhand der Ausgangsdaten überprüfen könne.

Obwohl die Regressionsrechnung eigentlich nicht zulässig war, da eine der beiden Variablen („Geschlecht“) nicht metrisch ist, sind die Rechenergebnisse gleichwohl sinnvoll interpretierbar – und diese Sinnhaftigkeit ist ein Beleg dafür, dass der Verstoß gegen die Voraussetzung für Regressionsrechnungen, nämlich metrische Variablen, quasi „geheilt“ ist.

Generell gilt: Wenn wir eine Variable haben, die mit 0 und 1 kodiert werden kann, dann gibt der Regressionskoeffizient b (der Steigungswinkel) den Unterschied in den zu erwartenden y-Werten (Einkommen) der Kategorie mit dem Wert 1 (Frauen) gegenüber der sog. Referenzkategorie (Wert 0 = Männer) an.

9.4 Nichtlineare Regression

Bisher haben wir ausschließlich lineare Regressionsbeziehungen betrachtet. Es kann nun aber durchaus auch sein, dass Sie bei einer gegebenen Aufgabenstellung von der Überlegung ausgehen müssen, dass der interessierende Zusammenhang – es möge der Überschaubarkeit halber wieder ein bivariater Zusammenhang sein – nichtlinear ist.

Schauen Sie sich beispielsweise die Daten der Tabelle in Abbildung 9.10 an. Hier wurden zehn zufällig ausgewählte Spitzenleichtathleten danach gefragt, wie viele Stunden sie im Tagesdurchschnitt trainieren und wie derzeit ihre Zeit beim 100-Meter-Sprint ist (auch hier handelt es sich um einen vergleichsweise dürftigen Datenbestand, was aber im Hinblick auf die Besprechung der einzusetzenden Methoden keine Rolle spielt; die entsprechenden Überlegungen gelten selbstverständlich auch dann, wenn mehr als zehn Beobachtungen vorliegen).

	training	zeitsec
1	4,50	10,80
2	2,50	11,80
3	3,00	11,30
4	3,50	10,90
5	1,60	13,50
6	2,00	12,50
7	2,50	12,00
8	3,00	11,40
9	4,00	10,90
10	3,50	11,00
11		

Abb. 9.10: Training und 100-m-Zeit

Diese Daten stellen sich grafisch so dar, wie es Abbildung 9.11 zeigt. Diese Abbildung macht deutlich, dass zunehmender Trainingsfleiß tendenziell mit einer Verbesserung der 100-m-Zeit einhergeht, allerdings werden diese Erfolgszuwächse tendenziell kleiner, wenn die Trainingsleistung immer weiter erhöht wird.

Würde man nun versuchen, in diese gekrümmte Punktwolke eine lineare Regressionsfunktion hineinzulegen, dann würde diese erstens die empirischen Punkte nicht besonders gut treffen und zweitens an der Erkenntnis vorbeigehen, dass die erzielbaren Erfolgszuwächse immer bescheidener werden. Wäre dem nicht so, dann müsste ja ein Läufer, der im Tagesdurchschnitt vielleicht 8 Stunden trainiert, 100-m-Zeiten erreichen können, die unter sieben Sekunden liegen. Deshalb ist hier eine gekrümmte Regressionsfunktion, also eine nichtlineare Funktion angemessener.

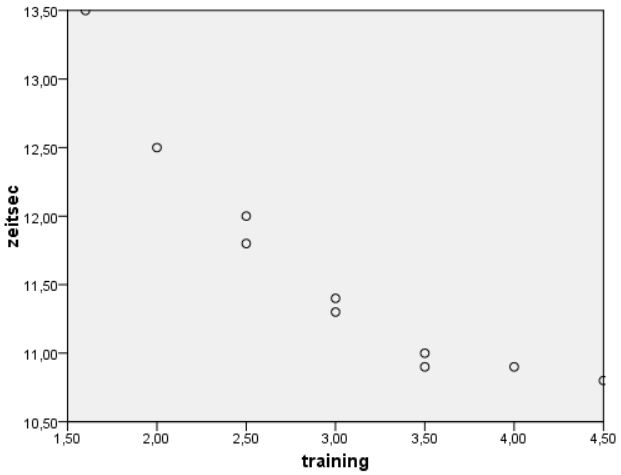


Abb. 9.11: Gekrümmte Punktwolke

Um nun mit SPSS eine solche nichtlineare Funktion zu bestimmen, muss man die Funktion zunächst mathematisch formulieren. Wie der Blick auf Abbildung 9.11 zeigt, könnte eine Parabel geeignet sein. Allgemein sieht eine solche Parabel so aus, wie es die folgende Formel beschreibt:

$$y = b_0 + b_1x + b_2x^2$$

Zu bestimmen sind die Parameter b_0, b_1 und b_2 so, dass sich die Parabel möglichst gut den empirischen Punkten anpasst. SPSS bietet als einfachsten Weg den über das Menü ANALYSIEREN/REGRESSION/KURVENANPASSUNG... an.

1. Greifen Sie auf die Ausgangsdaten zu (siehe Abbildung 9.10).
2. Öffnen Sie das Menü ANALYSIEREN/REGRESSION/KURVENANPASSUNG...

Sie gelangen zu einem Dialogfenster, wo Sie Ihre Variablen eingeben und im Feld MODELLE bei QUADRATISCH anklicken. Es ergibt sich dann das Bild der Abbildung 9.12.

SPSS zeigt auch im rechten Teil der folgenden Tabelle (Abbildung 9.13) die Parameterschätzer, wobei mit „Konstante“ den Ordinatenabschnitt bezeichnet, der oben in der Parabelgleichung b_0 genannt wurde.

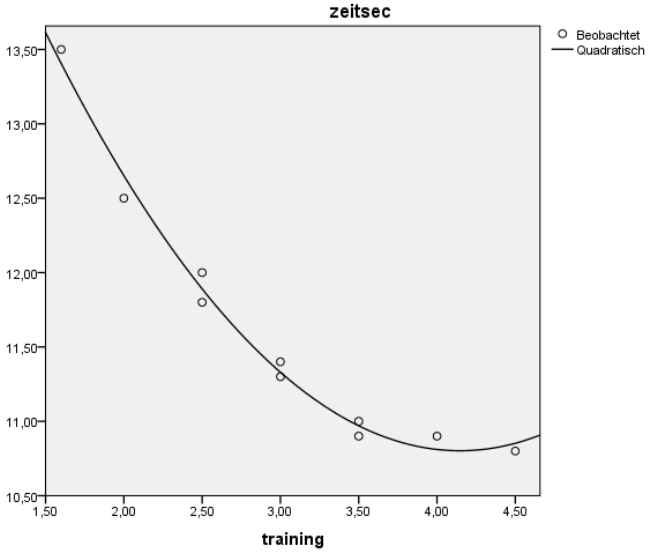


Abb. 9.12: Kurvenanpassung

Modellzusammenfassung und Parameterschätzer

Abhängige Variable: zeitsec

	Modellzusammenfassung					Parameterschätzer		
Gleichung	R-Quadrat	F	Freiheitsgrade 1	Freiheitsgrade 2	Sig.	Konstante	b1	b2
Quadratisch	,989	312,699	2	7	,000	17,705	-3,330	,402

Die unabhängige Variable ist training.

Abb. 9.13: Parameter der nichtlinearen Regressionsfunktion

10 Zusammenhangsrechnung

10.1 Zielsetzungen

Im vorangegangenen Kapitel haben Sie gesehen, wie man eine bivariate Häufigkeitsverteilung mit Hilfe einer Regressionsfunktion zusammenfassend beschreiben kann. Diese Regressionsfunktion konnte auch für prognostische Aufgaben verwendet werden.

Es leuchtet unmittelbar ein, dass die Qualität einer solchen prognostischen Aussage von einer Reihe von Bedingungen abhängt:

1. Wie viele Merkmalsträger sind untersucht worden? Je größer der Stichprobenumfang, desto besser abgesichert sind prognostische Aussagen.
2. Entspricht der gewählte Funktionstyp, also z.B. die lineare Regressionsfunktion dem tatsächlichen Zusammenhang? Wenn nicht, kann eine Prognose nicht sonderlich treffsicher sein.
3. Wie stark ist der Zusammenhang zwischen den beiden betrachteten Variablen? Je deutlicher der statistische Zusammenhang ist, je „schlanker“ also die Ausgangspunktwolke ist, desto treffsicherer ist die Prognose.

Damit sind wir beim Thema dieses Kapitels. Wie kann die Stärke des statistischen Zusammenhangs zwischen zwei Variablen quantifiziert werden? Die Beantwortung dieser Frage führt zu den statistischen Zusammenhangsmaßen, von denen mehrere zur Verfügung stehen. Bei ihrer Besprechung unterscheidet man zweckmäßigerweise nach der Skalenqualität der interessierenden Variablen. Bei metrischen Daten verwendet man den Korrelationskoeffizienten von Bravais/Pearson, bei Ordinaldaten den Rangkorrelationskoeffizienten von Spearman, bei Nominaldaten verwendet man Chi-Quadrat-basierte Maße.

10.2 Korrelationskoeffizient für metrische Variablen

Ausgangspunkt der *Korrelationsrechnung,* wie die *Zusammenhangsrechnung* auch genannt wird, ist die Frage, warum eine interessierende Untersuchungsvariable streut. Auf diese erkenntnisleitende Frage kann man nur dann eine (erste) Antwort finden, wenn man diese Untersuchungsvariable mit anderen in Verbindung bringt, wie das ja im Kapitel 9 mit der Regressionsrechnung schon geschehen ist. Auch im Kapitel über grafische Darstellungen (Kapitel 4) wurde schon gezeigt, dass die gemeinsame Betrachtung von zwei Variablen zu interessanten Befunden führen kann. Sie können dem Streudiagramm beispielsweise entnehmen, dass mit zunehmenden Werten einer Variablen X die Werte einer interessierenden Variablen Y tendenziell zunehmen.

Wenn Sie also die Frage interessiert, warum Körpergewichte zufällig ausgewählter Erwachsener variieren (streuen), dann können Sie in bivariater Betrachtung zu dem folgenden Ergebnis kommen: Wenn die Körpergrößen (X) befragter Personen zunehmen, nimmt tendenziell auch ihr Gewicht zu. Damit haben Sie auf der Grundlage einer bivariaten statistischen Auswertung eine erste Idee darüber gewonnen, welche beeinflussende Größe es sein könnte, mit der die interessierenden unterschiedlichen Körpergewichte zusammenhängen.

Wir nannten für die Bemessung der Stärke des Zusammenhangs zwischen zwei metrischen Variablen den Korrelationskoeffizienten von Bravais/Pearson. Es wird die sog. Kovarianz durch die Wurzel aus dem Produkt der beiden Einzelvarianzen dividiert. Betrachten wir das folgende Beispiel (Fehleranzahlen in zwei Klassenarbeiten):

X=Deutsch	Y=Mathematik
0	5
4	1
3	2
2	2
3	1
Summe: 12	Summe: 11

Ohne SPSS zu bemühen, können Sie rasch ausrechnen, dass der Mittelwert der Deutschfehler bei 12/5 = 2,4 und der der Mathematikfehler bei 11/5 = 2,2 liegt. Die Berechnung der Varianzen für beide Variablen führt zu folgenden Werten:

Varianz der X-Werte = 1,84; Varianz der Y-Werte = 2,16

Die Kovarianz ist die durch n dividierte Summe der Produkte der Abweichungen der X-Werte von ihrem arithmetischen Mittel und der Abweichungen der Y-Werte von ihrem arithmetischen Mittel. Es ergibt sich folgender Wert:

Kovarianz zwischen X und Y = -1,88

Mit diesen Ausgangswerten können Sie jetzt den Korrelationskoeffizienten von Bravais/ Pearson berechnen:

$$r = -1{,}88/\sqrt{1{,}84 * 2{,}16} = -0{,}943$$

Dieser Koeffizient, mit r abgekürzt, ist im Wertebereich zwischen -1 und +1 definiert. Bei r = +1 liegt ein maximal starker gleichgerichteter Zusammenhang vor (steigen die Werte der Variablen X, steigen auch die der Variablen Y; alle empirischen Beobachtungen liegen im Achsenkreuz auf einer ansteigenden Geraden), bei r = -1 liegt ein maximal starker gegenläufiger Zusammenhang vor (steigt X, fällt Y; alle Punkte liegen auf einer fallenden Geraden), bei r = 0 liegt kein statistischer Zusammenhang vor – Zwischenwerte für r können entsprechend interpretiert werden (es wird dabei jeweils Linearität der eventuellen Beziehung zwischen X und Y unterstellt). In unserem Beispiel hat sich also ein sehr starker gegenläufiger Zusammenhang zwischen der Anzahl der Fehler im Deutschdiktat und der Fehleranzahl in der Mathematikarbeit ergeben. Mit steigender Anzahl der Diktatfehler nimmt mit sehr hoher Wahrscheinlichkeit die Fehlerzahl in der Mathematikarbeit ab.

Es gibt nun unangenehmerweise ein paar Fallstricke bei der Interpretation berechneter Korrelationskoeffizienten. Stellen Sie sich beispielsweise vor, dass für eine Reihe betrachteter Länder die Geburtenziffer erfasst wird und die Zahl der Störche. Sie werden dann einen deutlichen gleichgerichteten Zusammenhang erkennen und vielleicht mit r = 0,8 quantifizieren können. Dies dürfte allerdings ein „vorgetäuschter“ Zusammenhang sein. Er kommt durch eine gemeinsam wirkende Drittvariable zustande, deren Einfluss auspartialisiert werden muss, wenn man den „wahren“ Zusammenhang zwischen Geburten und Störchen entdecken will. Darüber werden wir aber erst in Kapitel 11 sprechen.

10.3 Determinationskoeffizient

Eine wichtige weitere Maßzahl in diesem Zusammenhang ist der Determinationskoeffizient, der als das Quadrat des Korrelationskoeffizienten definiert ist. In obigem Zahlenbeispiel ergibt sich also $-0{,}943^2 = 0{,}889$. Dieser Determinationskoeffizient hat bemerkenswerte Eigenschaften, was deutlich wird, wenn Sie sich noch einmal die Varianz der Y-Werte anschauen. Sie hatte sich zu 2,16 ergeben. Wenn man nun in die kleine Punktwolke des obigen Beispiels eine lineare Regressionsfunktion hineinlegt (siehe dazu Kapitel 9), so hat diese die Parameter:

$a = 4{,}652 \qquad b = -1{,}022$

Berechnet man mit diesen Parametern die theoretischen Y-Werte (die Mathematikfehleranzahl, die bei gegebenen X-Werten zu erwarten ist), so ergeben sich die Werte der folgenden Arbeitstabelle:

X=Deutsch	Y=Mathematik	theor. Y-Werte
0	5	4,65
4	1	0,57
3	2	1,59
2	2	2,61
3	1	1,59

Berechnen Sie jetzt auch noch die Varianz der theoretischen Y-Werte, ergibt sich der Wert 1,92. Beziehen Sie diese Varianz auf die Varianz der Y-Werte, erhalten Sie 1,92/2,16 = 0,889.

Dies ist wieder der *Determinationskoeffizient*, den Sie oben als das Quadrat des Korrelationskoeffizienten ($r^2 = -0{,}943^2 = 0{,}889$) kennen gelernt haben. Sie sehen, dass er zum Ausdruck bringt, wie groß die Streuung der theoretischen, zu erwartenden Y-Werte in Bezug auf die Gesamtstreuung der beobachteten Y-Werte ist. Man spricht vom „*erklärten Streuungsanteil*“. 88,9% der Streuung der interessierenden Variablen Y (Anzahl der Fehler in der Mathematikarbeit) wird statistisch erklärt über die Variation der X-Werte (Anzahl der Fehler im Deutschdiktat) unter Nutzung der Hypothese, dass der mit der linearen Regressionsfunktion zum Ausdruck gebrachte Zusammenhang zwischen den beiden interessierenden Variablen den wahren Zusammenhang zwischen beiden zutreffend beschreibt.

Zwei zusätzliche Hinweise sind an dieser Stelle, ausgehend von dem obigen Zahlenbeispiel, angebracht:

1. Die Varianz der theoretischen Y-Werte ist mit 1,92 kleiner als die Varianz der beobachteten Werte (2,16). Nur wenn alle beobachteten Punkte genau auf der Regressionsgeraden liegen würden (Sie haben schon weiter oben erkannt, dass dann ein maximal starker Zusammenhang zwischen X und Y vorliegen würde), wären die beiden genannten Varianzen gleich groß. Der Determinationskoeffizient wäre dann 1.
2. Die Varianz der Reste (Y-Yt) ergänzt sich zusammen mit der Varianz der theoretischen Werte zur Gesamtvarianz der beobachteten Werte. Es gilt also – und dies generell, ohne dass das hier mathematisch bewiesen werden soll –, das sogenannte *Prinzip der Varianzzerlegung*:

   Varianz der Y-Werte = Varianz der Yt-Werte + Varianz der Reste (Y-Yt)

10.4 Rangkorrelationskoeffizient für Ordinaldaten

Wenn keine metrischen Daten vorliegen, kann der Korrelationskoeffizient von Bravais/Pearson als Maß der Stärke des Zusammenhangs zwischen zwei Untersuchungsvariablen X und Y nicht eingesetzt werden (über eine wichtige Ausnahme wird in Abschnitt 10.6 gesprochen). Deshalb hat man für ordinalskalierte und nominalskalierte Variablen andere Maßzahlen entwickelt. Für Ordinaldaten steht u.a. der *Rangkorrelationskoeffizient von Spearman* zur Verfügung, der in diesem Abschnitt betrachtet werden soll.

Der Koeffizient von Spearman (ρ) ist aus dem Korrelationskoeffizienten von Bravais/Pearson hergeleitet worden. Er geht aber nicht von den Merkmalswerten der beiden Variablen X und Y aus, sondern benutzt nur deren Rangpositionen. Die entsprechende Berechnung überlassen wir SPSS (siehe Abschnitt 10.7).

Auch dieser Koeffizient ist im Wertebereich zwischen -1 und +1 definiert, so dass konkrete Werte, entsprechend wie beim Korrelationskoeffizienten r, interpretiert werden können.

10.5 Zusammenhangsmaße für Nominaldaten

Wenn die vorliegenden Daten nur nominalskaliert sind, ist zur Berechnung der Stärke des statistischen Zusammenhangs auf andere Maßzahlen zurückzugreifen. Davon gibt es mehrere, von denen hier die beiden wichtigsten betrachtet werden sollen, der sogenannte Vierfelder-Koeffizient (Phi) und das Kontingenzmaß von Pearson (C).

Vierfelder-Koeffizient

Stellen Sie sich zwei nominalskalierte Variablen vor, die jeweils nur zwei Ausprägungen aufweisen (man nennt Variablen mit nur zwei Ausprägungen *dichotome Variablen*), also etwa die Variable „Geschlecht“ mit den beiden Ausprägungen „männlich“ und „weiblich“ und eine zweite Untersuchungsvariable, etwa die Antworten auf die Frage „Stimmen Sie der Forderung zu, eine generelle Geschwindigkeitsbegrenzung auf Autobahnen einzuführen?“ Diese Frage kann nur mit Ja oder Nein beantwortet werden, so dass es sich also auch dabei um eine dichotome Variable handelt. Zudem sind beide nominalskaliert. Ein entsprechender Datenbestand könnte also zum Beispiel folgendermaßen aussehen:

	männlich	weiblich	Summe
ja	15	28	43
nein	25	12	37
Summe	40	40	80

Es interessiert jetzt die Frage: Wie stark ist der statistische Zusammenhang zwischen den beiden betrachteten Variablen, also zwischen dem Geschlecht und dem Antwortverhalten auf die gestellte Frage?

In der folgenden Tabelle wurde eine allgemeine Notation für derartige Vierfelder-Tabellen vereinbart:

	0	1	Summe
0	a	b	S1
1	c	d	S2
Summe	S3	S4	n

Rechnet man nun a*d - b*c und bezieht diese Differenz auf die Wurzel aus dem Produkt aller Randsummen, ergibt sich der sog. *Vierfelder-Phi-Koeffizient.*

Für unser Zahlenbeispiel ergibt sich der Wert Phi = -0,326. Bei der Interpretation eines solchen Wertes ist zu beachten, dass auch diese Maßzahl im Wertebereich zwischen -1 und +1 definiert ist. Allerdings hat das Vorzeichen keine Bedeutung. Würde man nämlich in der obigen Tabelle die Ja-Zeile einfach mit der Nein-Zeile vertauschen, was natürlich bei einer nominalskalierten Variablen erlaubt ist, würde der Koeffizient das Vorzeichen wechseln. Es gibt hier keine Richtung des Zusammenhangs, es gibt nur einen mehr oder weniger starken Zusammenhang, bzw., wenn der Phi-Koeffizient den Wert 0 ergibt, keinen Zusammenhang.

Kontingenzmaß von Pearson

Wenn die beiden Variablen nicht dichotom sind, oder wenn wenigstens eine von beiden mehr als zwei Ausprägungen hat, dann kann dieser Phi-Koeffizient nicht verwendet werden. In diesem Fall bietet sich der *Kontingenzkoeffizient von Pearson* an, zu dessen Herleitung wir auf das Zahlenbeispiel aus Kapitel 7 zurückgreifen, wo es um die Gegenüberstellung der Variablen „Geschlecht“ und „bevorzugte politische Partei“ ging (siehe Abbildung 10.1).

Partei * Geschlecht Kreuztabelle

			Geschlecht		
			männlich	weiblich	Gesamt
Partei	CDU/CSU	Anzahl	43	40	83
		Erwartete Anzahl	41,1	41,9	83,0
	SPD	Anzahl	44	34	78
		Erwartete Anzahl	38,6	39,4	78,0
	F.D.P.	Anzahl	5	6	11
		Erwartete Anzahl	5,4	5,6	11,0
	Die Grünen	Anzahl	6	15	21
		Erwartete Anzahl	10,4	10,6	21,0
	Sonstige	Anzahl	1	6	7
		Erwartete Anzahl	3,5	3,5	7,0
Gesamt		Anzahl	99	101	200
		Erwartete Anzahl	99,0	101,0	200,0

Abb. 10.1: Kreuztabelle mit Erwartungswerten

In dieser Abbildung wurden von SPSS auch schon die *Erwartungswerte* berechnet (Erwartungswert = Zeilensumme * Spaltensumme / Gesamtzahl der Beobachtungen), d.h. diejenigen Häufigkeiten, die in den einzelnen Zellen der Verteilung zu erwarten wären, wenn Unabhängigkeit zwischen beiden Variablen bestehen würde.

Es bietet sich nun an, die Differenzen zwischen beobachteten Häufigkeiten und den bei Unabhängigkeit zu erwartenden Häufigkeiten zum Ausgangspunkt eines Zusammenhangsmaßes zu machen, denn offenbar müssen diese Differenzen um so größer werden, je stärker der interessierende Zusammenhang ist. Diese Differenzen haben wir in der drittletzten Spalte der folgenden Arbeitstabelle notiert. Um wieder zu verhindern, dass sich negative und positive Differenzen gegenseitig ausgleichen, wurden sie in der vorletzten Spalte dieser

Arbeitstabelle quadriert. Diese quadrierten Differenzen wurden in der letzten Spalte durch die Erwartungswerte dividiert (darüber wurde schon an anderer Stelle gesprochen). Die Summe, die rechts unten berechnet wurde, wird *Pearson'sche Prüfgröße U* genannt.

	beobachtet B	***erwartet E***	***Diff (B - E)***	***(B-E)²***	***(B-E)²/E***
CDU/CSU - m	43	41,4	1,6	2,56	0,062
CDU/CSU - w	40	41,9	-1,9	3,61	0,086
SPD - m	44	38,6	5,4	29,16	0,755
SPD - w	34	39,4	-5,4	29,16	0,740
FDP - m	5	5,4	-0,4	0,16	0,030
FDP - w	6	5,6	0,4	0,16	0,029
Die Grünen - m	6	10,4	-4,4	19,36	1,862
Die Grünen - w	15	10,6	4,4	19,36	1,826
Sonstige - m	1	3,5	-2,5	6,25	1,786
Sonstige - w	6	3,5	2,5	6,25	1,786
				U =	**8,961**

Pearson hat nun auf Basis dieser Größe U ein Zusammenhangsmaß C entwickelt, indem er U durch U+n dividiert und aus diesem Quotienten die Wurzel zieht. Hier ergibt sich also:

$$C = \sqrt{8{,}961 / (8{,}961 + 200)} = 0{,}62$$

Die Maßzahl C wird *Kontingenzkoeffizient* genannt. Er ist im Wertebereich zwischen 0 und 1 definiert. Allerdings wird der Maximalwert C = 1 nur erreicht bei bivariaten Tabellen (*Kontingenztabellen*), die unendlich viele Spalten und unendlich viele Zeilen haben. In allen anderen Fällen ergibt sich der Maximalwert, den C erreichen kann, indem der Mittelwert gebildet wird aus

$$\sqrt{(z-1)/z} \text{ und } \sqrt{(s-1)/s}$$

wobei mit z und s die Zeilenzahl und die Spaltenzahl der Kontingenztabelle gemeint sind. Hier erhalten Sie also als Maximalwert:

$$(\sqrt{(5-1)/5} + \sqrt{(2-1)/2})/2 = 0{,}8 \text{ (gerundet)}$$

10.6 Der Alleskönner

Wir kommen in diesem Abschnitt noch einmal auf den Korrelationskoeffizienten von Bravais/Pearson zurück (siehe Abschnitt 10.2), von dem festgestellt wurde, dass er nur bei metrischen Daten berechnet werden kann. Es kann nun allerdings gezeigt werden, dass diese Einschränkung eigentlich nur theoretischer Natur ist. Betrachten Sie dazu noch einmal das Beispiel zur Berechnung des Vierfelder-Phi-Koeffizienten:

	männlich	weiblich	Summe
ja	15	28	43
nein	25	12	37
Summe	40	40	80

Hier hatte sich Phi = -0,326 ergeben, wobei das Vorzeichen, wie erwähnt wurde, unbeachtlich ist.

Sie können nun in dieser Tabelle die dichotomen Ausprägungen der beiden Variablen jeweils mit 0 und 1 kodieren, also „männlich“ = 0, „weiblich“ = 1, „ja“ = 0 und „nein“ = 1. Wenn man nun die so kodierten Daten in eine neue Arbeitstabelle einträgt, in der ersichtlich ist, dass zum Beispiel die Merkmalswertkombination (0,0) 15 Mal aufgetreten ist, erhält man die folgende Tabelle:

Merkmalskombination	X	Y	Häufigkeit
0, 0 (männlich - ja)	0	0	15
0, 1 (weiblich - ja)	1	0	28
0, 1 (männlich - nein)	0	1	25
1, 1 (weiblich - nein)	1	1	12

Wenn man nun aus diesen Angaben den Korrelationskoeffizienten von Bravais/Pearson berechnet, ergibt sich mit r = -0,326 der gleiche Wert wie bei der Berechnung des zuständigen Vierfelder-Phi-Koeffizienten

Daraus ist folgendes zu entnehmen: Wenn dichotome Variablen vorliegen, die (zulässigerweise) mit 0 und 1 kodiert werden, kann man den Vierfelder-Phi-Koeffizienten durch den Korrelationskoeffizienten von Bravais/Pearson ersetzen, obwohl dieser eigentlich wegen der Skalenqualität der Variablen nicht zuständig ist.

Hat man nicht-dichotome Variablen (zum Beispiel die Variable „bevorzugte politische Partei“), ist diese Ersetzung allerdings nicht möglich. Es sei denn, Sie würden die nicht-dichotome Variable künstlich dichotomisieren und wieder mit 0 und 1 kodieren.

Selbstverständlich ist es möglich, die Variable „bevorzugte politische Partei“ nur mit zwei Ausprägungen auszustatten, zum Beispiel mit den Ausprägungen „SPD“ und „Nicht-SPD“. Dabei gehen allerdings Detailinformationen verloren – aber es könnte auch jetzt wieder der Korrelationskoeffizient von Bravais/Pearson eingesetzt werden.

Wenn man den Informationsverlust, der mit der Dichotomisierung einer an sich *polytomen Variablen* (eine Variable mit mehr als zwei Ausprägungen) einhergeht, vermeiden will, gibt es noch einen Ausweg: Eine polytome Variable mit zum Beispiel fünf Ausprägungen kann künstlich in vier dichotome Variablen zerlegt werden, wie das folgende Beispiel zeigt:

Polytome Variable Y		:	Bevorzugte politische Partei
Ausprägungen von Y		:	CDU/CSU, SPD, FDP, Die Grünen, Sonstige
Zerlegung:			
Variable 1	Ausprägungen	:	CDU/CSU - Nicht CDU/CSU
Variable 2	Ausprägungen	:	SPD - Nicht SPD
Variable 3	Ausprägungen	:	FDP - Nicht FDP
Variable 4	Ausprägungen	:	Die Grünen - Nicht Die Grünen

Überlegen Sie, warum Sie keine fünfte künstliche Variable brauchen und dennoch den gesamten Ausgangsinformationsbestand durch vier dichotomisierte, mit 0 und 1 kodierten Variablen zum Ausdruck bringen können. Solche künstlichen Variablen werden als *Dummy-Variablen* bezeichnet.

Da nun, wie oben beschrieben wurde, bei 0/1-kodierten Variablen der Korrelationskoeffizient von Bravais/Pearson eingesetzt werden kann, kann diese Maßzahl als universelle Maßzahl zur Bemessung der Stärke statistischer Zusammenhänge verwendet werden.

10.7 Berechnungen

Metrische Variablen

Ausgehend vom Beispiel in Abschnitt 10.2, wo die Fehleranzahl im Deutschdiktat und die Fehler in einer Mathematikarbeit von fünf zufällig ausgewählten Schülern einer Altersklasse einander gegenübergestellt wurden, soll nun der zuständige Korrelationskoeffizient von Bravais/Pearson mit SPSS berechnet werden. Dazu ist wie folgt vorzugehen: Hier noch einmal die Ausgangsdaten

	deutsch	mathe
1	0	5
2	4	1
3	3	2
4	2	2
5	3	1

Abb. 10.2: Ausgangsdaten für die Korrelationsberechnung

Zuständig für die Korrelationsberechnung ist das Menü ANALYSIEREN/KORRELATION/ BIVARIAT... Im sich öffnenden Dialogfenster übertragen Sie beide Variablen nach rechts und klicken an – sofern das noch erforderlich ist – bei PEARSON. Mit OK gelangen Sie dann zu den folgenden Befunden:

Korrelationen

		deutsch	mathe
deutsch	Korrelation nach Pearson	1	-,943*
	Signifikanz (2-seitig)		,016
	N	5	5
mathe	Korrelation nach Pearson	-,943*	1
	Signifikanz (2-seitig)	,016	
	N	5	5

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Abb. 10.3: Korrelationskoeffizienten

Hinweis:

Die Korrelation einer Variablen mit sich selbst ist immer r = 1; die Korrelation von X mit Y ist genauso groß wie die von Y mit X.

SPSS berechnet den Wert r = -0,943 (den hatten wir weiter oben schon per Hand berechnet) und gibt an, dass dieser Wert signifikant von null verschieden ist. SPSS gibt zudem die Zahl der Fälle (n=5) und die Signifikanz mit 0,016 an. Bei einem üblichen Signifikanzniveau von 5% bedeutet diese Überschreitungswahrscheinlichkeit, dass, ausgehend von dem absolut sehr hohen Wert r = -0,943, die Hypothese der Unabhängigkeit zwischen beiden Variablen in der Grundgesamtheit, aus der die Zufallsstichprobe gezogen wurde, verworfen werden muss.

Wir sind hier von einem zweiseitigen Signifikanzniveau ausgegangen, weil die Hypothese, dass in der Grundgesamtheit kein Zusammenhang vorliegt, sowohl von großen positiven wie auch von zahlenmäßig großen negativen Korrelationskoeffizientenwerten verworfen

werden kann. Die zuständige Wahrscheinlichkeitsverteilung, welcher der Korrelationskoeffizient aus einer Zufallsstichprobe in seiner Eigenschaft als Zufallsvariable folgt, wollen wir hier nicht vorstellen. Welche Verteilung zu verwenden ist, hängt unter anderem von der Stichprobengröße ab (siehe dazu: *Tiede/Voß*, Schließen mit Statistik – Verstehen, Verlag Oldenbourg, 2000).

Ordinalskalierte Variablen

Gegeben sei der folgende Datenbestand: Aufwand in Minuten zur Vorbereitung auf eine Klassenarbeit und die in dieser Arbeit erzielte Zensur:

Min (Zeit in Minuten) X	Zensur Y
100	3
180	1
60	5
120	3
200	2

1. Geben Sie diese Daten in eine neue Tabelle ein.
2. Wählen Sie ANALYSIEREN/KORRELATION/BIVARIAT...
3. Übertragen Sie beide Variablen („Min“ und „Zensur“) in den Bereich VARIABLEN:.
4. Sorgen Sie für ein Häkchen im Bereich KORRELATIONSKOEFFIZIENTEN bei SPEARMAN.
5. Sollte bei PEARSON noch ein Häkchen sein, klicken Sie es weg.
6. Klicken Sie auf die runde Optionsschaltfläche im Bereich TEST AUF SIGNIFIKANZ bei ZWEISEITIG, falls dies noch erforderlich ist.
7. Klicken Sie auf OK und Sie gelangen zur Abbildung 10.4.

Korrelationen

			Vorbereitungszeit in Minuten	Zensur
Spearman-Rho	Vorbereitungszeit in Minuten	Korrelationskoeffizient	1,000	-,872
		Sig. (2-seitig)	.	,054
		N	5	5
	Zensur	Korrelationskoeffizient	-,872	1,000
		Sig. (2-seitig)	,054	.
		N	5	5

Abb. 10.4: Rangkorrelationskoeffizienten von Spearman

Hier ist wieder nur der Wert $\rho = -0{,}872$ von Interesse. Er weist eine Überschreitungswahrscheinlichkeit von 0,054 auf. Dieser Befund besagt (bei einem Signifikanzniveau von 5%), dass trotz des starken Zusammenhangs die Hypothese der Unabhängigkeit zwischen beiden Variablen in der Grundgesamtheit, aus der die Zufallsstichprobe mit den Ausgangsdaten stammt, nicht verworfen werden kann. Das hat natürlich damit zu tun, dass die Stichprobe außerordentlich klein ist.

Nominalskalierte Variablen

Hier greifen wir zur Illustration auf unser erstes Beispiel aus Kapitel 2 zurück, wo die Variablen „Geschlecht“ und „bevorzugte politische Partei“ auftauchten (aus der Datei B00.SAV). Um den Pearson’schen Kontingenzkoeffizienten C zu berechnen, ist folgendermaßen vorzugehen:

1. Öffnen der Tabelle mit den Ausgangsdaten (B00.SAV).
2. Auswählen von ANALYSIEREN/DESKRIPTIVE STATISTIKEN/KREUZTABELLEN...
3. Im sich öffnenden Dialogfenster übertragen Sie die Variablen „Sex“ in den Bereich SPALTEN: und die Variable „Partei“ in den Bereich ZEILEN:.
4. Klicken Sie auf die Schaltfläche STATISTIKEN...

Sie gelangen zu einem Fenster, in dem Sie bei CHI-QUADRAT und bei KONTINGENZ-KOEFFIZIENT anklicken. Mit WEITER und OK gelangen Sie zu der Ausgabe in Abbildung 10.5.
In der Tabelle CHI-QUADRAT-TESTS finden sie beim Stichwort CHI-QUADRAT NACH PEARSON den Wert 8,89. Dies ist die Pearson’sche Prüfgröße U, die Sie weiter oben bei einem anderen Beispiel schon kennen gelernt haben (siehe Abschnitt 10.5). Daneben steht unter DF der Wert 4. Mit dieser Notation wird der Begriff „*degrees of freedom*“ (*Freiheitsgrade*) abgekürzt. Was hat es damit auf sich?

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	8,891[a]	4	,064
Likelihood-Quotient	9,412	4	,052
Zusammenhang linear-mit-linear	5,690	1	,017
Anzahl der gültigen Fälle	200		

a. 2 Zellen (20,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 3,47.

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Kontingenzkoeffizient	,206	,064
Anzahl der gültigen Fälle		200	

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Abb. 10.5: Ergebnisse der Berechnungen

Wir hatten schon an anderer Stelle darauf aufmerksam gemacht, dass U (SPSS bezeichnet dieses U als Chi-Quadrat) Ausprägung einer Zufallsvariablen ist, die näherungsweise der sog. *Chi-Quadrat-Verteilung* folgt (siehe Abschnitt 8.7).

Will man beispielsweise die Hypothese der Unabhängigkeit der beiden betrachteten Variablen in der Grundgesamtheit überprüfen, wird wieder die Überschreitungswahrscheinlichkeit berechnet, die Wahrscheinlichkeit dafür also, dass – Gültigkeit der zu prüfenden Nullhypothese vorausgesetzt – in einer Zufallsstichprobe vom angegebenen Umfang ein U-Wert auftritt, der 8,89 oder noch größer (noch weiter von der Nullhypothese abweichend) ist.

Zur Berechnung dieser Überschreitungswahrscheinlichkeit wird die zuständige Verteilung verwendet, also die Chi-Quadrat-Verteilung (in diesem Beispiel ist aber eine wichtige Einschränkung zu beachten, auf die wir weiter unten zu sprechen kommen). Nun gibt es aber unendlich viele verschiedene Chi-Quadrat-Verteilungen, je nachdem, aus wie vielen Summanden die Prüfgröße U gebildet wurde. Die Zahl der Summanden liegt bei einer Kontingenztabelle mit 5 Zeilen und 2 Spalten bei 2*5 = 10 (Sie erinnern sich: Es werden 10 quadrierte, relativierte Abweichungen zwischen beobachteten Häufigkeiten und den bei Unabhängigkeit zu erwartenden, theoretischen Häufigkeiten addiert, um zu U zu gelangen).

Genau genommen kann jeder Summand als Ausprägung einer eigenen Zufallsvariablen interpretiert werden, weil es ja vom Zufall abhängt, wie jede einzelne der beobachteten Häufigkeiten in den einzelnen Tabellenzellen ausfällt. Damit ist auch zufällig, wie groß die Abweichung, wie groß die quadrierte Abweichung und wie groß die quadrierte, relativierte Abweichung ist.

Nun muss allerdings die folgende Überlegung berücksichtigt werden, die Sie aus anderem Zusammenhang schon kennen: Wenn Sie alle verschiedenen Zufallsstichproben vom Umfang n = 203 ziehen würden, die man aus der gegebenen Grundgesamtheit ziehen könnte, dann sind nur vier der 10 Summanden „echte“ Variablen. Sind nämlich vier Zellen der Kontingenztabelle mit beobachteten Häufigkeiten, die zufällig so ausfallen, wie sie beobachtet werden, besetzt, dann ergeben sich die übrigen Zellenbesetzungen aus der mathematischen Notwendigkeit, dass die Randverteilungen (Geschlechtsverteilung und Verteilung der Variablen „bevorzugte politische Partei“) gegeben sind.

Daraus können Sie entnehmen, dass eine Kontingenztabelle mit z = 5 Zeilen und s = 2 Spalten nur zu (5-1)*(2-1) = 4 „frei beweglichen“ (variablen) Summanden führt. Deshalb sagt man, dass U Ausprägung einer Zufallsvariablen ist, die approximativ einer Chi-Quadrat-Verteilung folgt mit 4 Freiheitsgraden, allgemein mit (z-1)*(s-1) Freiheitsgraden. Dies ist der Wert, den SPSS unter DF ausgibt.

Wichtige Einschränkung:

Die Behauptung, dass U näherungsweise einer Chi-Quadrat-Verteilung folgt, trifft nur dann zu, wenn jeder der Erwartungswerte in den einzelnen Zellen der Kontingenztabelle mindestens den Wert 5 aufweist. Dies ist in unserem Beispiel aber nicht der Fall. SPSS gibt nämlich an (siehe Abbildung 10.5), dass der kleinste Erwartungswert bei 3,47 liegt, und dass es zwei Zellen gibt, die einen Erwartungswert aufweisen, der kleiner als 5 ist. Dies bedeutet, dass in diesem Beispiel die Überschreitungswahrscheinlichkeit nicht mit der Chi-Quadrat-Verteilung berechnet werden kann.

In einem solchen Fall ist es erforderlich, vor den Signifikanzberechnungen Klassen zusammenzulegen nzw. – falls auch dann noch Erwartungswerte kleiner als 5 auftauchen – zu einem anderen Verfahren überzugehen.

Schließlich berechnet SPSS den Kontingenzkoeffizienten von Pearson zu C = 0,206 (siehe Tabelle SYMMETRISCHE MAßE in Abbildung 10.5).

11 Multiple Regression und partielle Korrelation

11.1 Fragestellungen

Es wurde an anderer Stelle am Beispiel des statistischen Zusammenhangs zwischen Geburten und Störchen in verschiedenen Ländern erläutert, dass es sog. *Drittvariableneinflüsse* geben kann, die zunächst überzeugend wirkende Zusammenhänge zwischen je zwei Untersuchungsvariablen gewissermaßen vortäuschen können. Es leuchtet ja unmittelbar ein, dass Geburtenziffern und Storchanzahl deutlich korrelieren müssen, wenn beispielsweise beide betrachteten Variablen, was zu erwarten ist, mit dem wirtschaftlichen Entwicklungsstand der betrachteten Länder zusammenhängen. Wir wissen, dass in industrialisierten Ländern die Geburtenziffern tendenziell im Vergleich etwa zu nichtindustrialisierten Entwicklungsländern niedrig sind. Zugleich ist auch die Zahl der Störche in Industrieländern eher bescheiden, weil diesen Tieren nicht mehr die Umweltbedingungen geboten werden, die es ihnen erlauben würden, in größerer Zahl zu existieren. Es braucht deshalb nicht zu überraschen, dass Geburtenziffern und Storchanzahl positiv miteinander korrelieren.

Sicherlich gibt es aber auch Zusammenhänge zwischen irgendwelchen Variablen X und Y, bei denen der eventuelle, gemeinsam wirkende Drittvariableneinfluss Ihnen nicht so ohne weiteres ein- oder auffällt. Trotzdem mag er aber existieren. Und wenn schon Drittvariableneinflüsse denkbar sind, dann liegt natürlich auch die Idee nahe, dass es auch eine vierte, fünfte oder sechste Variable geben mag, die den eigentlich interessierenden bivariaten Zusammenhang störend beeinflusst.

Ein zweites Beispiel, das Ihnen auch schon begegnet ist, beleuchtet einen zweiten wichtigen Aspekt des Drittvariableneinflusses: An anderer Stelle hatten wir uns mit der Frage befasst, inwieweit der Ernteertrag auf Probefeldern von unterschiedlichen Mengen des Düngemitteleinsatzes beeinflusst wird. Es leuchtet nun unmittelbar ein, dass es auch andere Variablen geben mag (Bodenqualität, Sonnenscheindauer, durchschnittliche Niederschlagsmengen, Einsatz von Unkrautvernichtungsmitteln etc.), die die Variation der abhängigen Variablen Y (Ernteertrag) mit beeinflussen. Will man also eine Prognose über den Ernteertrag auf Probefeldern wagen, dann wird diese Prognose vermutlich besser werden, wenn man nicht nur den Düngemitteleinsatz als unabhängige Variable X berücksichtigt, sondern weitere unabhängige Variablen, also in einem ersten Schritt eine weitere Variable Z, vielleicht den Einsatz an Unkrautvernichtungsmitteln.

In diesem Kapitel wird deshalb der Frage nachgegangen, wie man Drittvariableneinflüssen auf die Spur kommen kann. Wenn man diese Frage beantworten kann, dann bereitet auch die Einbeziehung einer vierten, fünften oder sechsten Variablen keine neuen Probleme mehr. Diese Frage kann, entsprechend den gerade skizzierten beiden Beispielen, in zwei Teilfragen aufgegliedert werden:

1. Wie kann der eventuelle gemeinsame Einfluss einer dritten Variablen auf eine eigentlich interessierende bivariate Beziehung aus dieser „herausgenommen“ werden, um die bivariate Beziehung um diesen Einfluss gewissermaßen zu bereinigen?
2. Wie kann der eventuelle Drittvariableneinfluss genutzt werden, um Prognosen der abhängigen Variablen zu verbessern?

Die erste Teilfrage führt zur Berechnung partieller Korrelationskoeffizienten (siehe Abschnitt 11.3), die zweite Teilfrage führt zur multiplen Regressionsrechnung, der wir uns im folgenden Abschnitt, in der Beschränkung auf den Drei-Variablen-Fall, zuwenden.

11.2 Multiple Regression (Drei-Variablen-Fall)

Betrachten Sie noch einmal das Beispiel des Ernteertrages auf Probefeldern:

Es soll die Hypothese untersucht werden, dass der Ernteertrag auf Probefeldern nicht nur vom Düngemitteleinsatz, sondern auch vom Einsatz von Unkrautvernichtungsmitteln abhängt. Hier liegt also eine abhängige Variable vor (Y = Ernteertrag) und zwei Unabhängige (X = Düngemitteleinsatz und Z = Einsatz von Unkrautvernichtungsmitteln). Die beiden Variablen X und Z werden in diesem gedanklichen Modell des Zusammenwirkens verschiedener Größen als „Unabhängige“ bezeichnet, obwohl sie beide ihrerseits möglicherweise miteinander zusammenhängen können; auf diesen Umstand kommen wir später noch zu sprechen.

Wenn Sie eine entsprechende empirische Untersuchung durchführen, entsteht eine dreidimensionale Punktwolke, die von SPSS auch noch grafisch dargestellt werden kann (spätestens aber bei der Einbeziehung einer vierten oder weiteren Variablen sind grafische Darstellungen nicht mehr möglich). Auch die dreidimensionale Punktwolke kann durch eine lineare (oder nichtlineare) Regressionsfunktion zusammenfassend beschrieben werden. Im klassischen linearen Fall handelt es sich dabei um eine lineare (ebene) Fläche in einem dreidimensionalen Achsenkreuz, deren Lage durch drei Parameter beschrieben und festgelegt wird:

1. Schnittpunkt mit der Y-Achse (Y = Ernteertrag)
2. Steigung in Richtung der X-Achse (X = Düngemitteleinsatz)
3. Steigung in Richtung der Z-Achse (Z = Einsatz von Unkrautvernichtern).

Um zu erkennen, wie SPSS mit einem solchen Beispiel umgeht, gehen wir von den Daten der Abbildung 11.1 aus.

	ernte	duenger	unkraut
1	125	1,50	250
2	120	1,25	200
3	130	1,55	245
4	135	1,50	240
5	145	1,62	245
6	155	1,75	280
7	150	1,65	270
8	130	1,43	225
9	120	1,28	230
10	140	1,45	250

Abb. 11.1: Ernteertrag, Düngemitteleinsatz und Verwendung von Unkrautvernichtern

In Abbildung 11.1 finden Sie die Angaben für 10 Probefelder:

Ernte	= Ernteertrag in Zentnern
Duenger	= Düngemitteleinsatz in kg,
Unkraut	= Unkrautvernichtungsmittel in g.

Zur Bestimmung der linearen Regressionsfläche sind die folgenden Schritte notwendig:

1. Wählen Sie – von der Ausgangstabelle ausgehend – ANALYSIEREN/REGRESSION/ LINEAR...
2. Übertragen Sie die Variable „Ernte“ ins Feld ABHÄNGIGE VARIABLE:.
3. Übertragen Sie die Variable „Duenger“ ins Feld UNABHÄNGIGE VARIABLE(N):.
4. Übertragen Sie auch die Variable „Unkraut“ ins Feld UNABHÄNGIGE VARIABLE(N):.
5. Klicken Sie STATISTIKEN... an. Sie gelangen zum Fenster der Abbildung 11.2.

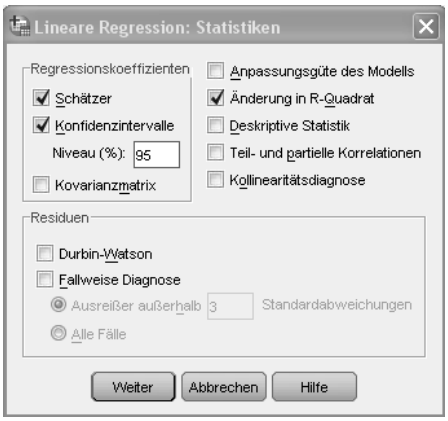


Abb. 11.2: Menü ANALYSIEREN/REGRESSION/LINEAR..., Schaltfläche STATISTIKEN...

6. Sorgen Sie für Häkchen bei SCHÄTZER, bei KONFIDENZINTERVALLE und bei ÄNDERUNG IN R-QUADRAT.
7. Klicken Sie auf WEITER.
8. Klicken Sie im Fenster ANALYSIEREN/REGRESSION/LINEAR... auf OK.

SPSS erzeugt jetzt die Befunde der Abbildung 11.3.

SPSS gibt zunächst in der Tabelle mit der Überschrift MODELLZUSAMMENFASSUNG unter dem Stichwort ÄNDERUNG IN R-QUADRAT einen *multiplen Determinationskoeffizienten aus (0,799),* was einem *multiplen Korrelationskoeffizienten* von r = 0,894 entspricht. Er bringt zum Ausdruck, wie stark der gemeinsame Zusammenhang zwischen den drei betrachteten Variablen ist.

Der multiple Determinationskoeffizient mit $r^2 = 0{,}799$ besagt, dass fast 80% der Streuung der interessierenden Variablen „Ernteertrag“ durch die Variation der beiden Variablen „Düngemitteleinsatz“ und „Einsatz von Unkrautvernichtungsmitteln“ statistisch, unter Nutzung der Hypothese des linearen Zusammenhangs zwischen den Variablen, erklärt wird. Vergleicht man diesen Wert mit dem bivariaten Determinationskoeffizienten zwischen X und Y („Düngemitteleinsatz“ und „Ernteertrag“), der bei $r^2 = 0{,}79$ (gerundet) lag, erkennt man, dass die Einbeziehung der dritten Variablen Z („Einsatz von Unkrautvernichtungs-

mitteln“) einen sehr geringen Zuwachs (aber immerhin) an statistischer Erklärungskraft mit sich brachte.

Modellzusammenfassung

Modell	Änderungsstatistiken				
	Änderung in R-Quadrat	Änderung in F	df1	df2	Sig. Änderung in F
1	,799[a]	13,873	2	7	,004

a. Einflußvariablen : (Konstante), unkraut, duenger

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressionskoeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	25,859	22,583		1,145	,290	-27,541	79,259
	duenger	58,125	29,111	,742	1,997	,086	-10,711	126,961
	unkraut	,091	,202	,166	,448	,668	-,388	,569

a. Abhängige Variable: ernte

Abb. 11.3: Befunde der multiplen Regression

Warum dieser Erklärungszuwachs so gering ausfällt, werden wir weiter unten beim Stichwort „Multikollinearität“ erläutern. Dieser Wert (0,799) ist signifikant von null verschieden, wie beim Stichwort ÄNDERUNG IN SIGNIFIKANZ VON F zu erkennen ist. Die Überschreitungswahrscheinlichkeit wird mit 0,004 angegeben. Dies ist die Überschreitungswahrscheinlichkeit für den F-Wert 13,873 (die zuständige Wahrscheinlichkeitsverteilung ist nicht die Gauß’sche Normalverteilung, sondern die F-Verteilung), d.h. der in der Stichprobe festgestellte multiple Zusammenhang kann auch für die Grundgesamtheit, aus der die Stichprobe stammt, unterstellt werden.

Weiterhin produziert SPSS die Parameter der linearen Regressionsfläche. Sie sehen in Abbildung 11.3, dass SPSS als Steigung zwischen X und Y (Dünger und Ernteertrag) den Wert 58,125 ausgibt, als Steigung zwischen Z und Y (Unkrautvernichtungsmittel und Ernteertrag) den Wert 0,09 (gerundet). Diese Größen werden *partielle Regressionskoeffizienten* genannt. Berücksichtigen Sie aber bei der Interpretation dieser beiden Werte die unterschiedlichen Maßeinheiten.

Wir hatten ja schon an anderer Stelle darauf aufmerksam gemacht, dass die berechneten B-Werte auf unterschiedliche Maßeinheiten der Ausgangsgrößen reagieren. Will man dies verhindern – und gerade dann, wenn, wie hier, zwei B-Werte miteinander verglichen werden könnten (was wirkt stärker auf den Ernteertrag, Dünger oder Unkrautvernichtungsmittel?), empfiehlt es sich, die Regressionskoeffizienten aus standardisierten Daten auszurechnen. SPSS hat das schon geleistet und weist die Rechenergebnisse unter der Überschrift BETA aus (siehe Abbildung 11.3). Dabei ergibt sich für den Düngemitteleinsatz ein standardisierter B-Wert von 0,742, für die Unkrautvernichtungsmittel von 0,166. Diese beiden Werte sind direkt miteinander vergleichbar. Ihr Unterschied besagt, dass der Düngemitteleinsatz das Ernteergebnis deutlicher beeinflusst (etwa viereinhalb Mal so stark) als der Einsatz von Unkrautvernichtungsmitteln.

Der Ordinatenabschnitt wird zu 25,859 angegeben. Dies bedeutet einen zu erwartenden Ernteertrag von knapp 26 Zentnern, wenn keine Dünger (X = 0) und keine Unkrautvernichtungsmittel (Z = 0) eingesetzt werden.

Weiterhin werden wieder die 95%-Vertrauensbereiche für die entsprechenden Koeffizienten der Grundgesamtheit, aus der die Stichprobe entnommen wurde, unter „95%-KON-

FIDENZINTERVALL FÜR B“ ausgegeben. Sie sind, wie Sie erkennen können, außerordentlich groß. Dies hat natürlich wieder damit zu tun, dass unsere Stichprobe mit n = 10 sehr klein ist, und u.a. auch damit, dass die Streuungsverhältnisse der einzelnen Variablen recht groß sind. Es verwundert deshalb nicht, dass unter dem Stichwort SIGNIFIKANZ relativ hohe Überschreitungswahrscheinlichkeiten angegeben werden. Für den ersten B-Wert liegt sie bei 0,086, für den zweiten B-Wert sogar bei 0,668, für den Ordinatenabschnitt, den man auch *Regressionskonstante* nennt, bei 0,290. Unterstellt man ein Signifikanzniveau von 5%, so bedeuten diese Überschreitungswahrscheinlichkeiten, dass keiner der drei Koeffizienten als statistisch signifikant von null verschieden angesehen werden kann. Verwechseln Sie dieses Ergebnis aber nicht mit der Aussage des multiplen Determinationskoeffizienten, der ja, wie weiter oben ausgeführt wurde, signifikant von null verschieden ist, der aber eine Aussage für beide beeinflussenden Variablen zugleich macht.

Statistischer Hinweis:

Weiter oben wurde darauf aufmerksam gemacht, dass die unabhängigen Variablen X und Z möglicherweise miteinander korrelieren, also ihrerseits nicht voneinander unabhängig sind. Wenn Sie beispielsweise, um dies zu kontrollieren, mit ANALYSIEREN/KORRELATION/BIVARIAT... die Korrelationsmatrix für alle drei betrachteten Variablen ausgeben lassen, gelangen Sie zu den Ergebnissen der Abbildung 11.4, die diese Korrelationskoeffizienten präsentiert.

Korrelationen

		ernte	duenger	unkraut
ernte	Korrelation nach Pearson	1	,890**	,827**
	Signifikanz (2-seitig)		,001	,003
	N	10	10	10
duenger	Korrelation nach Pearson	,890**	1	,890**
	Signifikanz (2-seitig)	,001		,001
	N	10	10	10
unkraut	Korrelation nach Pearson	,827**	,890**	1
	Signifikanz (2-seitig)	,003	,001	
	N	10	10	10

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

Abb. 11.4: Matrix der bivariaten Korrelationskoeffizienten

Sie erkennen einen deutlichen Zusammenhang zwischen den Variablen Düngemitteleinsatz (X) und dem Einsatz von Unkrautvernichtungsmitteln (Z), der mit r = 0,890 bemessen wird. In einem solchen Fall der Abhängigkeit zwischen den beeinflussenden Variablen spricht man von *Multikollinearität* oder einfach von *Kollinearität*. Im Fall perfekter Multikollinearität bietet die Regressionsrechnung keine eindeutige Lösung an. Die Schätzung der Regressionskoeffizienten wird umso instabiler, je höher der Grad an Multikollinearität ist.

Es bietet sich deshalb an, bei hoher Multikollinearität zwischen zwei beeinflussenden Variablen X und Z eine der beiden Unabhängigen aus dem Regressionsmodell herauszulassen, weil sie ja, ausweislich des hohen bivariaten Korrelationskoeffizienten, hinreichend durch die andere repräsentiert wird.

SPSS bietet Ihnen die Möglichkeit einer Kollinearitätsdiagnose, auf die wir aber nicht eingehen wollen. Allerdings führen diese Überlegungen zu der Frage, wie man bei der Auswahl der Variablen, die ein Regressionsmodell konstituieren, vorzugehen hat.

Am Anfang steht eine theoretische Überlegung, die an der Frage anknüpft, welches die Variablen sind, die den Ernteertrag beeinflussen. Wir haben in diesem Beispiel die Variablen

„Düngemitteleinsatz“ und „Einsatz von Unkrautvernichtungsmitteln“ ins Spiel gebracht. Sie haben aber gesehen, dass diese beiden Variablen deutlich miteinander korrelieren (hier erweist sich wieder die Zweckmäßigkeit der zuerst durchzuführenden bivariaten Berechnungen), so dass es nahe liegt, auf eine der beiden Unabhängigen zu verzichten und dafür vielleicht eine andere Variable in das gedankliche Modell aufzunehmen (vielleicht die Regenniederschlagsmenge). SPSS leistet bei der Beantwortung der Frage, welche Variablen berücksichtigt werden sollen, bescheidene Hilfestellung: Schauen Sie sich dazu im Dialogfenster der Abbildung 11.5 den Bereich METHODE: an.

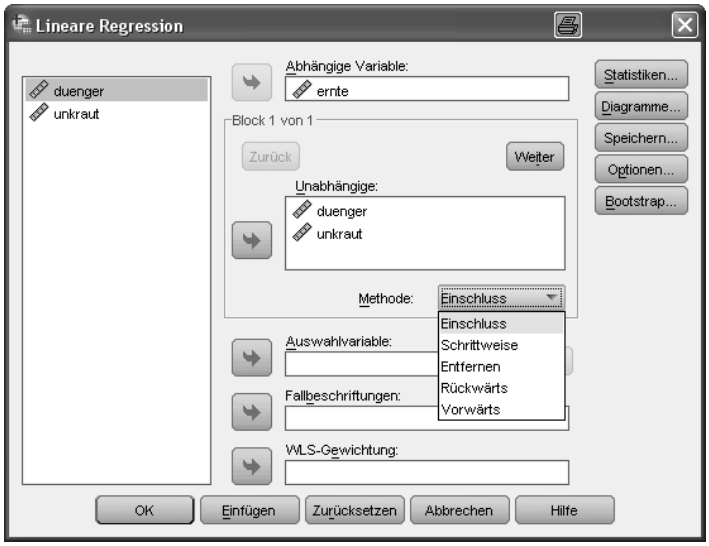


Abb. 11.5: Menü ANALYSIEREN/REGRESSION/LINEAR...,Bereich METHODE:

In diesem Bereich METHODE: der Abbildung 11.5 haben wir per Voreinstellung bisher immer EINSCHLUSS benutzt. Dies ist die übliche Vorgehensweise. Bei ihr werden alle von Ihnen als unabhängige Variablen angegebenen Variablen im Regressionsmodell verwendet.

Bei dieser Methode müssen Sie anhand der Ergebnisse entscheiden, ob eine dritte unabhängige Variable zur Verbesserung der Modellgüte beiträgt oder nicht und deshalb besser wieder aus dem Modell entfernt werden sollte.

SPSS stellt aber, wie Sie der Abbildung 11.5 entnehmen können, auch andere Auswahlverfahren bereit. Bei diesen Vorgehensweisen entscheidet SPSS, welche Variablen aufgenommen werden sollen und welche nicht, wobei sich das Programm an der statistischen Erklärungskraft der einzelnen Variablen orientiert. Wählen Sie beispielsweise die Methode RÜCKWÄRTS, wird SPSS die Variable Z („Einsatz von Unkrautvernichtungsmitteln“) wegen ihrer geringen Erklärungskraft, also wegen ihrer hohen Kollinearität mit der Variablen X („Düngemitteleinsatz“) aus dem Modell ausschließen.

11.3 Partielle Korrelation

Zu Beginn dieses Kapitels hatten wir als Aufgabe der multiplen Berechnungen auch die Bereinigung bivariater Korrelationskoeffizienten um Drittvariableneinflüsse angesprochen. Man spricht vom *Auspartialisieren* des Drittvariableneinflusses durch Berechnung *partiel-*

ler Korrelationskoeffizienten. Schauen Sie sich den Datenbestand der Abbildung 11.6 an. Dort haben wir für 10 ausgewählte Länder drei Variablen erfasst:

Y = Geburtenziffer (Zahl der Geburten pro 100 Einwohner)

X = Störche (Zahl der Störche pro 100 Quadratkilometer)

Z = Industrialisierungsgrad (Anteil der Industrieproduktion am Sozialprodukt).

	land	geburt	storch	indust
1	BRD	,50	,81	55,50
2	Frankreich	,75	,85	42,10
3	England	,65	,72	47,30
4	USA	,91	1,12	49,20
5	Indien	1,72	1,88	22,80
6	Aegypten	1,93	2,21	21,20
7	China	1,66	2,13	23,40
8	Peru	1,35	1,21	36,30
9	Burma	1,57	1,88	27,30
10	Sudan	2,02	2,33	18,10
11				

Abb. 11.6: Länderdaten (fiktive Angaben)

Berechnet man mit SPSS aus diesen Angaben alle bivariaten Korrelationskoeffizienten, gelangt man zur Abbildung 11.7.

Korrelationen

		geburt	storch	indust
geburt	Korrelation nach Pearson	1	,967**	-,971**
	Signifikanz (2-seitig)		,000	,000
	N	10	10	10
storch	Korrelation nach Pearson	,967**	1	-,945**
	Signifikanz (2-seitig)	,000		,000
	N	10	10	10
indust	Korrelation nach Pearson	-,971**	-,945**	1
	Signifikanz (2-seitig)	,000	,000	
	N	10	10	10

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

Abb. 11.7: Bivariate Korrelationskoeffizienten

Es zeigt sich ein sehr deutlicher gleichgerichteter Zusammenhang zwischen Geburten und Störchen (r = 0,972). Aber auch der Zusammenhang zwischen Industrialisierungsgrad und Störchen ist mit r = -0,9420 sehr deutlich – und zwar gegenläufig, ebenso der Zusammenhang zwischen Industrialisierungsgrad und Geburten (r = -0,968). Es taucht deshalb sofort der Verdacht auf, dass der eigentlich interessierende Zusammenhang zwischen Störchen und Geburten eher ein „vorgetäuschter" Zusammenhang sein dürfte. Deshalb soll nun aus diesem Zusammenhang der gemeinsame Einfluss der Variablen „Industrialisierungsgrad" auspartialisiert werden. SPSS leistet dies über ANALYSIEREN/ KORRELATION/PARTIELL... Es öffnet sich daraufhin das Dialogfenster der Abbildung 11.8.

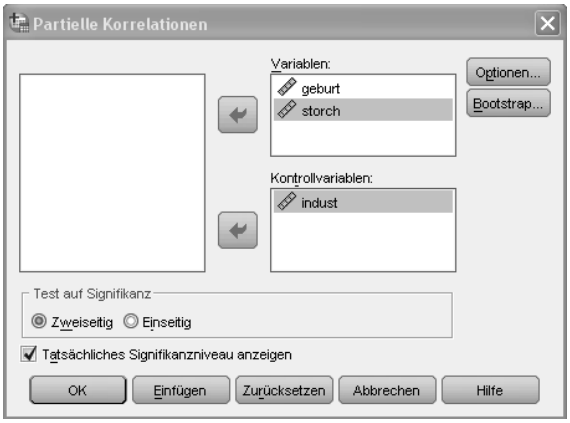


Abb. 11.8: Menü ANALYSIEREN/KORRELATION/ PARTIELL...

Im Fenster der Abbildung 11.8 übertragen Sie die Variablen „Geburt“ und „Storch“ in den Bereich VARIABLEN:, die Variable „Indust“ in den Bereich KONTROLLVARIABLEN:. Klicken Sie dann OK an, erzeugt SPSS die Ergebnisse der Abbildung 11.9.

Korrelationen

Kontrollvariablen			geburt	storch
indust	geburt	Korrelation	1,000	,633
		Signifikanz (zweiseitig)	.	,067
		Freiheitsgrade	0	7
	storch	Korrelation	,633	1,000
		Signifikanz (zweiseitig)	,067	.
		Freiheitsgrade	7	0

Abb. 11.9: Partieller Korrelationskoeffizient

Mit der Formel für den partiellen Korrelationskoeffizienten berechnet SPSS den Wert r = 0,691. Partialisiert man also den gemeinsamen Einfluss der Variablen „Industrialisierungsgrad“ aus dem Zusammenhang zwischen Störchen und Geburten aus, der ursprünglich mit r = 0,972 bemessen wurde, so verringert sich die Stärke des „eigentlichen“ Zusammenhangs deutlich auf den Wert r = 0,691.

12 Statistische Tests für Mittelwerte

12.1 Aufgabenstellung

Das arithmetische Mittel (siehe auch Kapitel 5) ist zweifelsohne die wichtigste Maßzahl der univariaten deskriptiven Statistik. Diese Maßzahl kennzeichnet die zentrale Tendenz der Häufigkeitsverteilung einer metrischen Variablen. Sie kann deshalb, wenn sie für eine Untersuchungsvariable, deren Werte auf der Basis einer Zufallsstichprobe gewonnen wurden, dazu verwendet werden, um *Hypothesen* zu überprüfen, die beispielsweise folgendermaßen lauten können:

1. Das durchschnittliche Nettoeinkommen abhängig Beschäftigter in der Bundesrepublik Deutschland liegt derzeit bei DM 2800 monatlich.
2. Die durchschnittliche Körpergröße erwachsener Männer ist in Bayern nicht anders als in Berlin.
3. In den 16 Ländern der Bundesrepublik zeigen sich keine Unterschiede in der durchschnittlichen Familiengröße.

Beim ersten Beispiel wird dem hypothetisch behaupteten Mittelwert derjenige der Stichprobe gegenübergestellt, um die Hypothesenentscheidung herbeizuführen. Es handelt sich dabei um einen einfachen parametrischen Signifikanztest, wie Sie ihn in theoretischer Erörterung in Kapitel 8 (Abschnitt 8.5) kennen gelernt haben. Das zweite Beispiel erfordert zwei Zufallsstichproben, eine in Bayern und eine in Berlin. Der entsprechende Test nennt sich Mittelwertdifferenzentest. Er testet, genau genommen, die Hypothese, dass die Mittelwerte der beiden Grundgesamtheiten gleich sind, dass in Wahrheit also keine Mittelwertdifferenz vorliegt, bzw. dass es nur eine einzige Grundgesamtheit gibt. Das dritte Beispiel ist ähnlich gelagert, mit dem Unterschied, dass jetzt mehr als zwei Zufallsstichproben vorliegen. Diese Aufgabenstellung führt zur sog. Varianzanalyse.

Der Vollständigkeit halber möchten wir darauf aufmerksam machen, dass in entsprechender Weise, beim Vorliegen nichtmetrischer Daten, Anteilswerttests und Anteilswertdifferenzentests durchgeführt werden können.

12.2 Test des arithmetischen Mittels

Hier geht es jetzt um eine wahrscheinlichkeitsstatistische Absicherung von Stichprobenbefunden oder anders formuliert, am Beginn des entsprechenden Forschungsprozesses steht eine Mittelwerthypothese. Sie könnte beispielsweise lauten:

Das Durchschnittsalter der deutschen wahlberechtigten Bevölkerung liegt bei 45 Jahren.

Um diese Hypothese zu überprüfen, greifen wir auf die Zufallsstichprobe vom Umfang n=203 zurück, also auf den Ausgangsdatenbestand des ersten Beispiels, in der sich ein Durchschnittsalter von 43,07 Jahren bei einer Standardabweichung von 15,52 Jahren ergeben hatte. Es soll nun, ausgehend von den theoretischen Überlegungen des Abschnitts 8.5, die Frage beantwortet werden, ob die Abweichung des Stichprobenmittelwertes von dem Wert, den die Nullhypothese behauptet (45 Jahre), als statistisch signifikant zu klassifizie-

ren ist oder nicht. Im ersten Fall wäre die Nullhypothese verworfen, im zweiten hingegen bestätigt. Vorgegeben sei ein zweiseitiges Signifikanzniveau von 10%.

Wie wird nun mit SPSS diese Testentscheidung herbeigeführt?

1. Wählen Sie, ausgehend von der Datei B00.SAV den Befehl ANALYSIEREN/MITTELWERTE VERGLEICHEN/T-TEST BEI EINER STICHPROBE...

Sie gelangen ins Dialogfenster der Abbildung 12.1.

Abb. 12.1: Menü ANALYSIEREN/MITTELWERTE VERGLEICHEN/T-TEST BEI EINER STICHPROBE...

2. Im Fenster der Abbildung 12.1 übertragen Sie die Variable „Alter“ in den Bereich TESTVARIABLE(N):.
3. In den Bereich TESTWERT: tragen Sie den Wert 45 ein.
4. Klicken Sie dann die Schaltfläche OPTIONEN... an.

Sie gelangen ins Fenster der Abbildung 12.2.

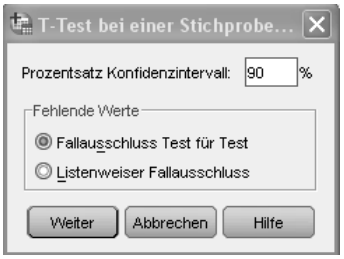


Abb. 12.2: Optionen für den Mittelwerttest

5. Tragen Sie bei KONFIDENZINTERVALL: den Wert 90 ein.
6. Klicken Sie auf WEITER.
7. Klicken Sie im Fenster der Abbildung 12.1 OK an.

SPSS erzeugt jetzt die Ergebnisse der Abbildung 12.3:

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
alter	203	43,07	15,523	1,090

Test bei einer Sichprobe

	Testwert = 45					
					90% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
alter	-1,772	202	,078	-1,931	-3,73	-,13

Abb. 12.3: Signifikanztest des Mittelwerts

Sie erkennen in dieser Abbildung, dass SPSS zunächst noch einmal, unter der Überschrift TEST BEI EINER STICHPROBE, den Wert der Nullhypothese, die getestet werden soll, ausgibt, also den Wert 45. Darunter wird der t-Wert ausgegeben (-1,772), die Zahl der Freiheitsgrade der einzusetzenden t-Verteilung (202) und die Überschreitungswahrscheinlichkeit (SIG. 2-seitig = 0,078). Bei einem zweiseitigen Signifikanzniveau von 10% (dies entspricht einem Konfidenzintervall von 90%, das ja für diesen Test vorgegeben wurde), bedeutet diese Überschreitungswahrscheinlichkeit, dass die Hypothese, der Mittelwert betrage 45, nicht bestätigt werden kann.

Zusätzlich wird die Differenz zwischen Hypothesenwert und Stichprobenbefund angegeben (Mittlere Differenz = -1,93 = 45-43,07, und das 90%-KONFIDENZINTERVALL der Differenz. Dieses Intervall reicht von -3,73 bis -0,13. Hier ist die Testentscheidung wiederum ablesbar: Weil dieser Bereich den Wert 0 nicht mit einschließt, ist die Nullhypothese zu verwerfen.

Zugleich wird Ihnen mit dieser Angabe die Möglichkeit geboten, ausgehend vom Stichprobenbefund (Durchschnittsalter = 43,07 Jahre) den unbekannten Mittelwert der Grundgesamtheit zu schätzen. Mit einem Vertrauen von 90% ist er im Bereich zwischen 45-3,73 = 41,27 und 45-0,13 = 44,87 zu erwarten.

Sie erkennen aus der Abbildung 12.3, dass eine t-Verteilung mit 202 Freiheitsgraden zuständig ist (DF = degrees of freedom = n-1 = 203-1 = 202). Wir hatten weiter oben und auch schon in Kapitel 8 behauptet, dass dieser Test auch mit der Gauß'schen Normalverteilung durchgeführt werden kann. Hier ist nun einschränkend anzumerken, dass das Zentrale Grenzwerttheorem die Gültigkeit der Normalverteilung als Wahrscheinlichkeitsverteilung für den Zufallsstichprobenmittelwert (in seiner Eigenschaft als Zufallsvariable) nur näherungsweise behauptet. Es wurde festgestellt, dass approximativ (näherungsweise) eine Normalverteilung gilt mit einem Mittelwert, der dem der Nullhypothese entspricht (45) und einer Standardabweichung (auch Standardfehler oder Stichprobenfehler genannt), die sich aus der Standardabweichung der Grundgesamtheit, geteilt durch die Wurzel aus dem Stichprobenumfang, ergibt.

Nun ist aber in aller Regel die Standardabweichung der Grundgesamtheit nicht bekannt und wird deshalb durch die der Stichprobe ersetzt. Diese Ersetzung durch den erwartungstreuen Schätzer ist aber nicht unproblematisch, weil ja die Stichprobenstandardabweichung ihrerseits keine Konstante, sondern auch Ausprägung einer Zufallsvariablen ist. Dies ist der Grund dafür, dass SPSS korrekterweise nicht mit der Gauß'schen Normalverteilung, sondern mit der für diesen Fall zuständigen t-Verteilung arbeitet. Die Hypothesenentscheidung fällt aber beim gegebenen Beispiel in beiden Fällen gleich aus. Allerdings kann es bei be-

stimmten Mittelwertdifferenzen so sein, dass bei Verwendung der Normalverteilung die Verwerfung der Nullhypothese, bei Verwendung der t-Verteilung aber (gerade noch) ihre Bestätigung empfohlen wird. Die t-Verteilung schützt ein bisschen länger die Nullhypothese vor der Verwerfung, eben wegen der Unschärfe, die durch die ersatzweise Verwendung der Stichprobenstandardabweichung anstelle derjenigen der Grundgesamtheit bei der Berechnung des Standardfehlers ins Spiel kommt. Insbesondere bei kleineren Stichprobenumfängen muss deshalb mit der t-Verteilung gearbeitet werden.

12.3 Mittelwertdifferenzentest

Hier geht es jetzt um den Fall von zwei unterschiedlichen, voneinander unabhängigen Zufallsstichproben. Stellen Sie sich beispielsweise vor, in Bayern wird eine Zufallsstichprobe gezogen und auch eine in Berlin. In beiden Stichproben werden Körpergrößen erfasst und jeweils die durchschnittliche Körpergröße ausgerechnet. Es soll die Hypothese geprüft werden, dass der eventuell auftauchende Mittelwertunterschied nur zufällig ist, dass also die beiden Grundgesamtheiten den gleichen Mittelwert aufweisen, bzw. dass nur eine einzige Grundgesamtheit vorliegt.

Der Grundgedanke des durchzuführenden Tests unterscheidet sich nicht von den bisher besprochenen Aufgabenstellungen. Man kann die Stichproben-Mittelwertdifferenz als Ausprägung einer Zufallsvariablen auffassen, und es geht jetzt wieder darum, die Überschreitungswahrscheinlichkeit der beobachteten Mittelwertdifferenz zu berechnen, um diese mit dem vorher festzulegenden Signifikanzniveau zu vergleichen. Zur Berechnung der Überschreitungswahrscheinlichkeit benötigen wir die zuständige Wahrscheinlichkeitsverteilung. Bei großen Stichprobenumfängen ist dies wieder approximativ eine Gauß'sche Normalverteilung, während die exakte Verteilung wieder eine t-Verteilung ist.

Ein einfaches Zahlenbeispiel soll die Vorgehensweise erläutern. In der Tabelle der Abbildung 12.4 haben wir Körpergrößen notiert, die aus zwei Zufallsstichproben stammen.

Wie Sie in Abbildung 12.4 sehen, sind die Werte beider Stichproben in eine Spalte eingetragen worden, die Stichprobenzugehörigkeit wurde als zweite Variable definiert (es steht 1 für Bayern, 2 für Berlin).

Zur Durchführung des Tests ist nun wie folgt vorzugehen:

1. Geben Sie die Ausgangsdaten, wie es Abbildung 12.4 zeigt, in eine neue Tabelle ein.
2. Wählen Sie ANALYSIEREN/MITTELWERTE VERGLEICHEN/T-TEST BEI UNABHÄNGIGEN STICHPROBEN... nSie gelangen ins Dialogfenster der Abbildung 12.5.
3. Im Fenster der Abbildung 12.5 übertragen Sie die Variable „Groesse" in den Bereich TESTVARIABLE(N):.
4. Übertragen Sie die Variable „Land" in den Bereich GRUPPENVARIABLE:.
5. Klicken Sie auf die Schaltfläche GRUPPEN DEF.... (Gruppen definieren). Sie gelangen ins Fenster der Abbildung 12.6

	groesse	land
1	178,00	1
2	172,00	1
3	175,50	1
4	183,00	1
5	181,20	1
6	172,30	1
7	174,70	1
8	181,00	1
9	173,90	1
10	172,80	1
11	176,60	2
12	182,30	2
13	185,50	2
14	179,40	2
15	179,30	2
16	177,50	2
17	172,40	2
18	184,50	2

Abb. 12.4: Daten für den Mittelwertdifferenzentest

Abb. 12.5: Menü ANALYSIEREN/MITTELWERTE VERGLEICHEN/T-TEST BEI UNABHÄNGIGEN STICHPROBEN...

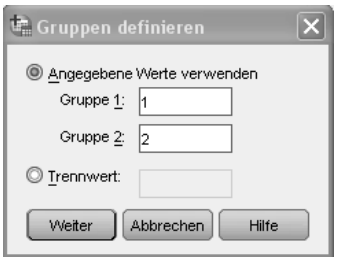


Abb. 12.6: Gruppen definieren

6. Klicken Sie, wenn es erforderlich ist, den Optionsschalter bei ANGEGEBENE WERTE VERWENDEN an.
7. Geben Sie bei GRUPPE 1: den Wert 1 ein.
8. Geben Sie bei GRUPPE 2: den Wert 2 ein.
9. Klicken Sie auf WEITER.
10. Klicken Sie im Fenster der Abbildung 12.5 auf OK.

SPSS erzeugt jetzt die Ergebnisse der Abbildung 12.7.

Gruppenstatistiken

	land	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
groesse	1	10	176,4400	4,07409	1,28834
	2	8	179,6875	4,33307	1,53197

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
groesse	Varianzen sind gleich	,032	,861	-1,634	16	,122	-3,24750	1,98719	-7,46016	,96516
	Varianzen sind nicht gleich			-1,622	14,688	,126	-3,24750	2,00169	-7,52190	1,02690

Abb. 12.7: Mittelwertdifferenzentest

Sie erkennen in dieser Abbildung, dass SPSS zunächst die Zahl der Fälle und die beiden Stichprobenmittelwerte ausgibt (176,4400 und 179,6875). Daneben stehen die Standardabweichungen (4,0741 und 4,3331) und die Standardfehler der Mittelwerte (1,2883 und 1,5320). Im zweiten Teil der SPSS-Ausgabe werden zwei Fälle unterschieden: Zunächst geht SPSS davon aus, dass die Varianzen gleich sind; in der zweiten Zeile wird Ungleichheit der Varianzen unterstellt. Was hat es damit auf sich?

Beim Mittelwertdifferenzentest wird vorausgesetzt, dass die Grundgesamtheitsstreuungen gleich sind. Da sich aber die Stichprobenstandardabweichungen voneinander unterschieden (siehe Ergebnisse in Abbildung 12.7), kann die Gültigkeit dieser Voraussetzung bezweifelt werden.

Deshalb arbeitet SPSS mit beiden Möglichkeiten: Zum einen wird die Gleichheit der Varianzen unterstellt, zum anderen wird mit den ungleichen Stichprobenvarianzen gearbeitet. Dabei wird im ersten Fall ein Test vorgeschaltet, der die Behauptung der Gleichheit der Grundgesamtheitsstreuungen prüft.

Dieser vorgeschaltete Test ist der *Levene-Test auf Varianzgleicheit*, der zu einer hohen Überschreitungswahrscheinlichkeit (zuständig ist die F-Verteilung) von 0,861 (=86,1%) führt. Die Hypothese der Gleichheit der Grundgesamtheitsstreuungen kann also verwendet werden, d.h. zur Interpretation der Ergebnisse des Mittelwertdifferenzentests können wir uns auf die erste Zeile der Tabelle in Abbildung 12.7 konzentrieren, andernfalls wären die Ergebnisse der zweiten Zeile interessant.

Die eigentlich zu überprüfende Nullhypothese (es gibt nur eine einzige Grundgesamtheit) impliziert Gleichheit der Standardabweichungen. Wenn nun der Standardfehler (Streuung der für die Berechnung der Überschreitungswahrscheinlichkeit zuständigen Wahrscheinlichkeitsverteilung) aus den Stichprobenstandardabweichungen geschätzt wird (das ist in der Regel, so auch hier, notwendig, weil die Streuungsverhältnisse der Grundgesamtheit(en) unbekannt sind), wird nicht selten ein gewogenes arithmetisches Mittel aus beiden Stichprobenstandardabweichungen verwendet (sog. *gepoolte Standardabweichung*). Dies führt dann zu den Ergebnissen, die SPSS in der zweiten Zeile der Tabelle in Abbildung 12.7 ausweist.

In beiden Fällen gelangt man zu Überschreitungswahrscheinlichkeiten (0,122 und 0,126), die zur Entscheidung führen, dass die zu testende Nullhypothese zu bestätigen ist. Die Hypothese der Gleichheit der Grundgesamtheitsmittelwerte kann aufgrund der Stichprobenbefunde nicht verworfen werden.

12.4 Varianzanalyse einfacher Klassifikation

Steht man vor der Aufgabe, die Mittelwertunterschiede aus mehr als zwei voneinander unabhängigen Stichproben zu untersuchen, sind die einfachen Vergleiche von je zwei Mittelwerten nicht tauglich, allein schon deshalb, weil z.B. bei 16 Stichproben aus den 16 Bundesländern 120 Mittelwertvergleiche durchgeführt werden müssten. Hier ist ein Verfahren einzusetzen, dass unter dem Namen „Varianzanalyse“ bekannt geworden ist. Dies ist ein etwas irreführender Name, weil es eigentlich nicht um Varianzen (Streuungen) sondern um Mittelwerte geht; Sie werden aber gleich erkennen, warum dieser Name gewählt wurde.

Zur Illustration der Vorgehensweise gleich wieder ein Beispiel, das am Ausgangsdatenbestand des ersten Beispiels (siehe Kapitel 2) anknüpft. Dort wurde in einer Stichprobe vom Umfang n = 203 u.a. das Alter zufällig ausgewählter wahlberechtigter Personen und das Bundesland, in dem sie wohnen, erfasst. Mit diesen Daten können Sie die Hypothese überprüfen, dass das Durchschnittsalter in allen Bundesländern gleich ist, dass also die 16 Mittelwerte, wenn überhaupt, dann nur zufällig voneinander abweichen, oder dass nur eine einzige Grundgesamtheit vorliegt.

Um diese Hypothese mit SPSS zu überprüfen, gehen Sie folgendermaßen vor:

1. Wählen Sie ANALYSIEREN/MITTELWERTE VERGLEICHEN/EINFAKTORIELLE ANOVA...

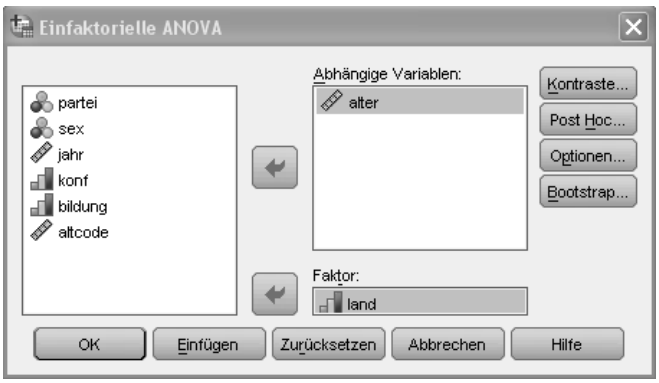


Abb. 12.8: Menü ANALYSIEREN/MITTELWERTE VERGLEICHEN/EINFAKTORIELLE ANOVA...

Hinweis:

Der Name ANOVA für dieses Verfahren kommt von der Abkürzung „ANalysis OF VAriance“ = Varianzanalyse.

2. Übertragen Sie im Fenster der Abbildung 12.8 die Variable „Alter“ in den Bereich ABHÄNGIGE VARIABLEN:.
3. Übertragen Sie die Variable „Land“ in den Bereich FAKTOR:.
4. Klicken Sie auf OK und Sie gelangen zur Abbildung 12.9.

Einfaktorielle ANOVA

alter

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	6872,227	15	458,148	2,049	,014
Innerhalb der Gruppen	41804,808	187	223,555		
Gesamt	48677,034	202			

Abb. 12.9: Ergebnisse der einfaktoriellen Varianzanalyse

In der Abbildung 12.9 finden Sie zunächst drei Quadratsummen. Die Quadratsumme ZWISCHEN DEN GRUPPEN liegt bei 6872,227 bei 15 Freiheitsgraden (DF = Zahl der Bundesländer bzw. Zahl der Einzelstichproben minus 1). Dann finden Sie die Quadratsumme INNERHALB DER GRUPPEN (41804,808 bei 187 Freiheitsgraden (=Gesamtstichprobenumfang minus Zahl der Teilstichproben = 203-16). Schließlich finden Sie die Quadratsumme GESAMT (48677,034 bei 202 Freiheitsgraden = n-1).

Zu diesen Quadratsummen sind einige Erläuterungen erforderlich: Wenn Sie aus allen n=203 Altersangaben die Varianz berechnen wollen (siehe dazu Kapitel 6), müssen Sie zunächst die Summe aller quadrierten Abweichungen der Merkmalswerte vom Gesamtmittelwert berechnen. Dies ergibt die Quadratsumme, die beim Stichwort GESAMT ausgewiesen ist.

Zusätzlich interessieren Sie sich natürlich dafür, wie sich die 16 Teilstichprobenmittelwerte (die Gruppenmittelwerte, wie man auch sagt) voneinander unterscheiden. Es ist zu erwarten, dass diese 16 Bundesland-Mittelwerte voneinander abweichen. Sind diese Abweichungen zu groß, werden sie als nicht mehr zufällig angesehen, und die Hypothese der gemeinsamen Grundgesamtheit wäre dann zu verwerfen. Deshalb ist es zweckmäßig, auch die Streuung dieser 16 Gruppenmittelwerte mit der Varianz dieser Werte zu bemessen – und auch dafür benötigen Sie zunächst eine Summe quadrierter Abweichungen, nämlich der Abweichungen der Gruppenmittelwerte vom Gesamtmittel. Diese Quadratsumme wird bei dem Stichwort ZWISCHEN DEN GRUPPEN ausgewiesen.

Schließlich kann auch über alle 16 Bundesländer hinweg untersucht werden, wie weit die einzelnen Merkmalswerte in den Gruppen vom jeweiligen Gruppenmittelwert abweichen. Auch dazu ist die Varianz geeignet, und es wird deshalb zunächst wieder, jetzt eine mit den Teilstichprobenumfängen gewichtete, Quadratsumme gebildet (Abweichungen der Merkmalswerte innerhalb der Gruppen vom jeweiligen Gruppenmittelwert – und dies über alle 16 Gruppen addiert), die beim Stichwort INNERHALB DER GRUPPEN ausgewiesen wird.

Aus diesen drei Quadratsummen können dann drei Varianzen berechnet werden. Die ersten beiden weist SPSS unter dem Stichwort MITTEL DER QUADRATE aus:

Varianz zwischen den Gruppen: 458,148

Varianz innerhalb der Gruppen: 223,555

Die Gesamtvarianz liegt übrigens bei 240,98.

Bevor wir nun auf die Testentscheidung zu sprechen kommen, möchten wir auf folgenden Umstand aufmerksam machen: Die drei berechneten Quadratsummen hängen direkt mathematisch miteinander zusammen, und zwar gilt:

Gesamte Quadratsumme =

= Quadratsumme zwischen den Gruppen + Quadratsumme innerhalb der Gruppen

48677,0345 = 6872,2269 + 41804,8076

Man spricht in diesem Zusammenhang vom Prinzip der *Quadratsummenzerlegung*.

Wenn nun die Nullhypothese zutrifft, wenn es also nur eine einzige Grundgesamtheit gibt, dann müssten die gerade berechneten drei Varianzen, jede für sich, erwartungstreue Schätzer für die unbekannte Grundgesamtheitsvarianz sein, anders formuliert, die Varianz zwischen den Gruppen (458,148) dürfte nur zufällig und nicht signifikant von der Varianz innerhalb der Gruppen (223,555) abweichen.

Es bietet sich deshalb an, die Nullhypothese über den Vergleich dieser beiden Varianzen zu prüfen – daher der Name des Verfahrens „Varianzanalyse". Wenn diese beiden Varianzen nur zufällig voneinander abweichen, dann dürfte der Quotient aus beiden nur zufällig vom Wert 1 abweichen (üblicherweise setzt man bei dieser Quotientenberechnung die größere der beiden Varianzen in den Zähler, die kleinere in den Nenner). Hier erhalten wir als Varianzquotient den Wert, den SPSS unter F ausweist (siehe Abbildung 12.9), also den Wert 2,049.

Die Hypothesenentscheidung reduziert sich jetzt auf folgende Frage: Wie wahrscheinlich ist das Ergebnis 2,049 oder ein noch weiter von der Nullhypothese abweichenderer Wert (also noch weiter vom Quotient = 1 abweichender)? Diese Überschreitungswahrscheinlichkeit wird wieder mit dem vorgegebenen Signifikanzniveau verglichen. Sie wird mit der zuständigen Wahrscheinlichkeitsverteilung (dies ist die F-Verteilung) bestimmt und ergibt sich hier zu 0,014 (= 1,4%), wie SPSS unter dem Stichwort SIGNIFIKANZ ausweist. Diese Überschreitungswahrscheinlichkeit ist kleiner als die üblichen Signifikanzniveaus, so dass die Nullhypothese verworfen wird. Es kann nicht unterstellt werden, dass die 16 Teilstichproben aus einer gemeinsamen Grundgesamtheit stammen, d.h. die Hypothese, dass die Durchschnittsalter in den 16 Bundesländern gleich seien, ist zu verwerfen.

Der Vollständigkeit halber möchten wir an dieser Stelle darauf aufmerksam machen, dass auch mehrfaktorielle Untersuchungsansätze denkbar sind. Stellen Sie sich beispielsweise vor, man hätte nicht 16 Bundesländerstichproben, sondern 32, nämlich je zwei in jedem Bundesland, wobei die eine Stichprobe pro Bundesland nur Männer erfasst, die andere nur Frauen. Mit einer zweifaktoriellen Varianzanalyse könnte dann untersucht werden, ob die beiden Faktoren („Bundesland" und „Geschlecht") Mittelwertunterschiede in der beispielsweise interessierenden Variablen „Alter" begründen können.

13 Anpassungstests

13.1 Aufgabenstellung

Im Gegensatz zu den bisher besprochenen Signifikanztestverfahren, die als parametrische Tests bezeichnet werden können, sollen nun nichtparametrische Tests besprochen werden, und zwar eine bestimmte Gruppe dieser Verfahren, die Anpassungstests. Diese Verfahren behandeln Verteilungshypothesen. Es geht dabei um die Frage, ob sich eine empirisch beobachtete Häufigkeitsverteilung, die aus einer Zufallsstichprobe stammt, mit hinreichender Güte an eine theoretische Verteilung anpassen lässt. Stellen Sie sich beispielsweise die folgende Nullhypothese vor:

> Die Augenzahlen beim einfachen Würfelwurf sind gleichverteilt.

Um diese Hypothese zu prüfen, könnten Sie 60 Mal würfeln und die Häufigkeiten der Augenzahlen notieren. Zu erwarten wären bei 60 Würfen jeweils die Häufigkeiten 10; die beobachteten Häufigkeiten werden aber mehr oder weniger deutlich davon abweichen. Sind diese Abweichungen zufällig, wird die Nullhypothese bestätigt, sind sie hingegen signifikant, wird sie verworfen.

Es geht also darum, an die theoretische Verteilung mit den theoretischen Häufigkeiten 10, 10, 10 usw. die empirische Verteilung anzupassen. Gelingt die Anpassung, so sind die Abweichungen zwischen beobachteten und theoretisch zu erwartenden Häufigkeiten nur zufällig, gelingt die Anpassung nicht, sind die Abweichungen signifikant. Im ersten Fall wird die Nullhypothese bestätigt, im zweiten Fall wird sie verworfen.

Sehr häufig in der statistischen Praxis geht es um Aufgabenstellungen, bei denen an eine empirische Häufigkeitsverteilung eine Gauß'sche Normalverteilung anzupassen ist, wie etwa bei der folgenden Nullhypothese:

> Intelligenzquotienten sind normalverteilt.

Eine solche Normalverteilungshypothese ist u.a. deshalb so wichtig, weil viele statistische Verfahren die Normalverteilung der interessierenden Untersuchungsvariablen voraussetzen. In diesem Fall ist es erforderlich, ausgehend von den Daten einer Zufallsstichprobe, zunächst diese Hypothese zu überprüfen.

13.2 Chi-Quadrat-Anpassungstest

In der Tabelle der Abbildung 13.1 finden Sie die Augenzahlen von 60 Würfelwürfen.

> **Wichtiger Hinweis:**
>
> Bei allen bisherigen Beispielen sind wir davon ausgegangen, dass Daten einer sog. Urliste vorliegen. Wenn also beispielsweise nur eine Variable betrachtet wird (z.B. Alter befragter Personen), dann war jede befragte Person mit einem einzigen Wert vertreten. Wir könnten deshalb das Würfelbeispiel auch so anlegen, dass tatsächlich 60 einzelne Augenzahlen angegeben werden (jeder Fall, jeder Wurf weist eine Augenzahl auf).

Wir können aber auch von einer schon fertigen Häufigkeitsverteilung ausgehen – und eine solche zeigt Abbildung 13.1. Es wird ja in der praktischen statistischen Arbeit nicht selten der Fall sein, dass man keine eigenen Erhebungen durchführt, sondern auf schon vorhandene Daten, die etwa in Form einer Häufigkeitsverteilung vorliegen, zurückgreift.

	Augenzahl	Häufigkeit
1	1	9
2	2	11
3	3	12
4	4	10
5	5	10
6	6	8

Abb. 13.1: Häufigkeitsverteilung über 60 Würfelwürfe

Liegt eine schon fertige Häufigkeitsverteilung vor, muss den statistischen Auswertungsverfahren ein Schritt vorangestellt werden, der Gewichtung genannt wird. In unserem Beispiel bedeutet dies, dass die einzelnen Augenzahlen 1 bis 6 zunächst mit ihren Häufigkeiten gewichtet werden müssen.

Dafür zuständig ist das Menü DATEN/FÄLLE GEWICHTEN… Dieses führt in das Fenster der Abbildung 13.2:

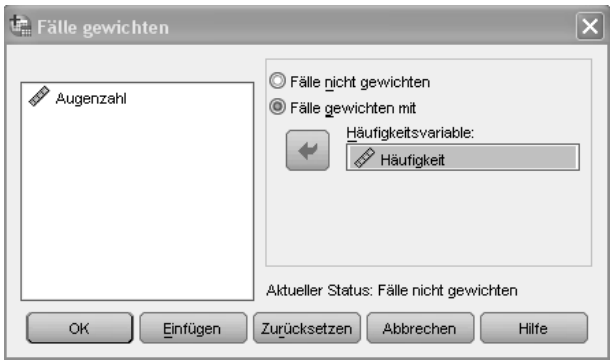


Abb. 13.2: Menü DATEN/FÄLLE GEWICHTEN…

Hier klicken Sie an bei FÄLLE GEWICHTEN MIT und übertragen die Variable „Häufigkeit“ in das Feld HÄUFIGKEITSVARIABLE:, gefolgt von OK.

Die Häufigkeitsverteilung der Abbildung 13.1 zeigt, dass nicht alle Augenzahlen tatsächlich gleich häufig aufgetreten sind. Kann also die Hypothese der Gleichverteilung aufrechterhalten bleiben? Um diese Frage mit SPSS zu beantworten, gehen Sie folgendermaßen vor:

1. Wählen Sie ANALYSIEREN/NICHTPARAMETRISCHE TESTS/ALTE DIALOGFELDER/ CHI-QUADRAT...

Sie gelangen ins Fenster der Abbildung 13.3.

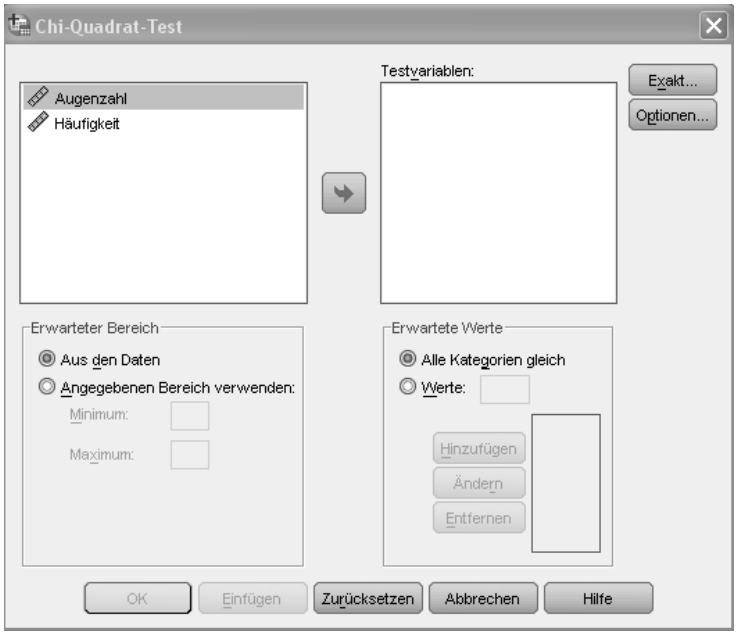


Abb. 13.3: Menü ANALYSIEREN/NICHTPARAMETRISCHE TESTS/ALTE DIALOGFELDER/CHI-QUADRAT...

2. Im Fenster der Abbildung 13.3 übertragen Sie die Variable „Augenzahl“ in den Bereich TESTVARIABLEN:.
3. Im Bereich ERWARTETE WERTE muss der Optionsschalter bei ALLE KATEGORIEN GLEICH angeklickt sein.
4. Klicken Sie auf OK und Sie gelangen zur Abbildung 13.4.

Augenzahl

	Beobachtetes N	Erwartete Anzahl	Residuum
1	9	10,0	-1,0
2	11	10,0	1,0
3	12	10,0	2,0
4	10	10,0	,0
5	10	10,0	,0
6	8	10,0	-2,0
Gesamt	60		

Statistik für Test

	Augenzahl
Chi-Quadrat	1,000[a]
df	5
Asymptotische Signifikanz	,963

a. Bei 0 Zellen (0,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 10,0.

Abb. 13.4: Anpassungstest

Sie sehen, es werden zunächst die beobachteten, die erwarteten Häufigkeiten und die Abweichungen (Residuen) ausgegeben. Dann wird unter CHI-QUADRAT die Pearson'sche Prüfgröße U ausgerechnet (U = 1,000), die Sie schon kennen gelernt haben. Diese Größe folgt näherungsweise einer Chi-Quadrat-Verteilung, wobei die Zahl ihrer Freiheitsgrade sich aus der Zahl der Summanden minus 1, also zu 6-1 = 5 ergibt.

Die Testentscheidung kommt somit in der Weise zustande, dass wir nach der Überschreitungswahrscheinlichkeit für U fragen. Mit der zuständigen Chi-Quadrat-Verteilung ergibt sich diese unter dem Stichwort ASYMPTOTISCHE SIGNIFIKANZ zu 0,963 (96,3%). Die Nullhypothese der Gleichverteilung kann also nicht verworfen werden.

Hinweis:

Schon in Kapitel 10 haben wir darauf aufmerksam gemacht, dass die Verwendung der Chi-Quadrat-Verteilung voraussetzt, dass die Erwartungswerte alle mindestens größer als 5 sein müssen. Dies ist hier auch der Fall. Hätten wir aber beispielsweise nur 20 statt 60 Würfelwürfe beobachtet, so könnte die Chi-Quadrat-Verteilung zur Durchführung dieses Tests nicht mehr verwendet werden. Es müssten dann zuerst Klassen zusammengelegt werden, z.B. alle geraden Augenzahlen in eine, alle ungeraden in eine zweite Klasse. Dabei würden allerdings Detailinformationen verloren gehen, was ein durchaus unerwünschter Nebeneffekt einer solchen Zusammenlegung wäre.

Werden auch durch das Zusammenlegen die Erwartungswerte nicht größer als 5, zum Beispiel bei nur 8 Würfelwürfen wäre dies zu erwarten, muss alternativ zu einem anderen Testverfahren gegriffen werden. Glücklicherweise stehen derartige Verfahren zur Verfügung. Eines davon wird in Abschnitt 13.5 besprochen. Allerdings ist generell anzumerken, dass die Zurückweisung einer Hypothese der oben formulierten Art um so weniger zu erwarten ist, je kleiner die Stichprobe ist, auf deren Basis die Testentscheidung herbeigeführt wird. Je kleiner die Stichprobe ist, desto augenfälliger und deutlicher müssen zum Beispiel die Abweichungen von der Gleichverteilung (Würfelbeispiel) sein, damit die Chance besteht, die Hypothese der Gleichverteilung verwerfen zu können.

13.3 Test auf Normalverteilung

Hier geht es um die Hypothese, dass eine bestimmte Untersuchungsvariable normalverteilt sei. SPSS kann dies grafisch überprüfen. Betrachten Sie dazu das folgende Beispiel:

Bei 20 zufällig ausgewählten Schülern einer Altersklasse werden die Fehler in einem Deutschdiktat notiert. Es ergibt sich die Tabelle in Abbildung 13.5.

	fehler
1	5
2	0
3	1
4	5
5	4
6	7
7	5
8	4
9	6
10	3
11	4
12	4
13	5
14	3
15	2
16	7
17	5
18	4
19	5
20	6

Abb. 13.5: Fehleranzahlen

Wenn man aus diesen Daten der Abbildung 13.5 mit SPSS eine Häufigkeitstabelle erstellt, ergibt sich das Bild der Abbildung 13.6.

fehler

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	1	5,0	5,0	5,0
	1	1	5,0	5,0	10,0
	2	1	5,0	5,0	15,0
	3	2	10,0	10,0	25,0
	4	5	25,0	25,0	50,0
	5	6	30,0	30,0	80,0
	6	2	10,0	10,0	90,0
	7	2	10,0	10,0	100,0
	Gesamt	20	100,0	100,0	

Abb. 13.6: Häufigkeitsverteilung

Man kann jetzt die folgende Überlegung anstellen: Wenn die Fehleranzahlen, wie die zu prüfende Nullhypothese behauptet, normalverteilt wären, müsste sich als Häufigkeitsverteilung näherungsweise die bekannte Glockenkurve ergeben. Betrachten Sie dazu Abbildung 13.7, der entnommen werden kann, dass von einer „glockenförmigen“ Verteilung (also von einer Normalverteilung) nicht gesprochen werden kann.

Die Testentscheidung selbst kann über einen Kolmogorov-Smirnov-Test herbeigeführt werden, den wir im nächsten Abschnitt besprechen. Allerdings kann auch der oben angesprochene Chi-Quadrat-Anpassungstest verwendet werden, wobei aber zunächst „per Hand“ die erwarteten Häufigkeiten bestimmt werden müssten. Das ersparen wir uns aber.

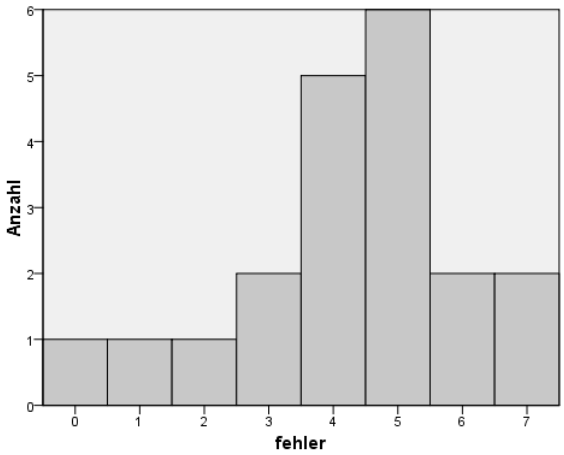


Abb. 13.7: Histogramm der Fehleranzahlen

13.4 Kolmogorov-Smirnov-Test

Der Kolmogorov-Smirnov-Test dient dazu, eine Verteilungshypothese zu testen, wobei die Verteilungsfunktionen der beobachteten und der theoretischen Häufigkeiten miteinander verglichen werden (unter dem Begriff der *Verteilungsfunktion* versteht man die aufwärts kumulierten relativen Häufigkeiten). Der Grundgedanke dieses Tests setzt an der maximalen Differenz zwischen den Werten der beiden Verteilungsfunktionen an, die nicht zu groß werden darf, wenn die Nullhypothese bestätigt werden soll.

Vorausgesetzt wird bei diesem Test eine metrische Untersuchungsvariable (die Variable „Fehleranzahl im Deutschdiktat“ erfüllt diese Bedingung). Auch bei gruppierten Daten kann das Verfahren verwendet werden, wenn die Faustregel eingehalten wird, dass n>20 (da sind wir mit meinem Beispiel gerade an der Grenze, denn n = 20), und wenn die Anzahl der Klassen größer als 5 ist.

Um diesen Test mit SPSS einzusetzen, gehen Sie folgendermaßen vor:

1. Geben Sie die Ausgangsdaten in eine neue Tabelle ein.
2. Wählen Sie ANALYSIEREN/NICHTPARAMETRISCHE TESTS/ALTE DIALOGFELDER/K-S BEI EINER STICHPROBE... Sie gelangen ins Fenster der Abbildung 13.8.
3. Im Fenster der Abbildung 13.8 übertragen Sie die Variable „Fehler“ in den Bereich TESTVARIABLEN:.
4. Im Bereich TESTVERTEILUNG sorgen Sie für ein Häkchen bei NORMAL.
5. Klicken Sie OK an.

Abb. 13.8: Menü ANALYSIEREN/NICHTPARAMETRISCHE TESTS/ALTE DIALOGFELDER/K-S BEI EINER STICHPROBE...

SPSS erzeugt jetzt die Ergebnisse der Abbildung 13.9.

Kolmogorov-Smirnov-Anpassungstest

		fehler
N		20
Parameter der Normalverteilung[a,b]	Mittelwert	4,25
	Standardabweichung	1,803
Extremste Differenzen	Absolut	,195
	Positiv	,139
	Negativ	-,195
Kolmogorov-Smirnov-Z		,871
Asymptotische Signifikanz (2-seitig)		,433

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Abb. 13.9: Kolmogorov-Smirnov-Test

In dieser Abbildung wird unter dem Stichwort ASYMPTOTISCHE SIGNIFIKANZ die Überschreitungswahrscheinlichkeit ausgegeben. Sie bemisst sich zu 0,433 (43,3%) und ist somit größer als ein übliches Signifikanzniveau. Die Hypothese, dass die Fehleranzahlen normalverteilt seien, kann somit bestätigt werden.

Dieses Ergebnis mag Sie überraschen. Es hat damit zu tun, dass der Stichprobenumfang in diesem Beispiel mit n=20 relativ klein ist. Erst bei größeren Stichprobenumfängen würden die beobachteten Abweichungen der erhobenen von den theoretisch zu erwartenden Häufigkeiten zu einer Verwerfung der Normalverteilungshypothese führen. Über dieses Problem, das im Zusammenhang mit kleinen Stichprobenumfängen auftaucht, wurde schon etwas ausführlicher am Ende des Abschnitts 13.2 gesprochen, so dass sich an dieser Stelle weitere Erörterungen erübrigen.

13.5 Binomialtest

Hat man es mit einer dichotomen Variablen zu tun, oder hat man, wegen zu kleiner Erwartungswerte beim Chi-Quadrat-Anpassungstest (siehe Abschnitt 13.2) die Klassenzahl auf zwei reduziert, die ursprünglich polytome Variable also künstlich dichotomisiert, ist der Binomialtest einzusetzen. Betrachten Sie folgendes Beispiel:

Die Nullhypothese, die getestet werden soll möge lauten, dass 50% aller Menschen weiblichen Geschlechts sind. In einer Zufallsstichprobe vom Umfang n=8 finden sich nur zwei Frauen. Ist dieser Befund mit der Nullhypothese vereinbar, oder muss diese verworfen werden?

Es ist auf den ersten Blick nicht einsichtig, dass es sich hier um eine Fragestellung handelt, die einen Anpassungstest erfordert. Schließlich handelt es sich doch offensichtlich um einen Anteilswerttest, also um einen parametrischen Test, der bei hinreichend großem Stichprobenumfang (Faustregel: n>30) mit Hilfe der Normalverteilung durchgeführt werden könnte – allerdings nicht hier, wo der Stichprobenumfang nur n=8 beträgt.

Diese Fragestellung kann aber auch als Verteilungshypothese aufgefasst werden: Kann die beobachtete (dichotome) Verteilung (2 Frauen, 6 Männer) mit hinreichender Güte an die theoretische (dichotome) Verteilung (4 Frauen, 4 Männer) angepasst werden? Diese Art der Fragestellung haben Sie schon in Abschnitt 13.2 kennengelernt, wo der Chi-Quadrat-Anpassungstest (Würfelbeispiel) besprochen wurde. Dieses Verfahren kann jetzt aber nicht verwendet werden, weil die Erwartungswerte (jeweils 4) kleiner als 5 sind, womit gegen die Faustregel für den Einsatz der Chi-Quadrat-Verteilung verstoßen wird.

Hier hilft nun die *Binomialverteilung* weiter. Es handelt sich dabei um eine diskrete Wahrscheinlichkeitsverteilung, mit deren Hilfe die folgende Frage beantwortet werden kann: Wie wahrscheinlich ist es, dass bei n voneinander unabhängigen Versuchen, bei denen jeweils nur zwei Ereignisse möglich sind, x_i-mal das im Sinn der Fragestellung günstige Ereignis auftritt, wenn bei jedem Versuch das günstige Ereignis die Wahrscheinlichkeit π hat.

Um die Testentscheidung bei dem gegebenen Beispiel herbeizuführen, ist also nach der Überschreitungswahrscheinlichkeit zu fragen: Wie wahrscheinlich ist es, dass bei n=8 zufällig ausgewählten Personen 2 oder weniger Frauen auftauchen (2 oder noch weiter von der Nullhypothese abweichend), wenn bei jedem „Versuch“ die Wahrscheinlichkeit dafür, eine Frau auszuwählen, 50% beträgt, wenn also bei n=8 Personen 4 Frauen zu erwarten sind?

Um dieses Beispiel mit SPSS durchzurechnen, haben wir in eine neue Tabelle zwei Einsen (1=weiblich) und 6 Nullen (0=männlich) eingegeben (siehe Abbildung 13.10). Dies ist ein vergleichsweise bescheidener Datenbestand, aber gerade dafür ist ja, wie schon ausgeführt wurde, der Binomialtest geeignet.

	sex
1	1
2	1
3	0
4	0
5	0
6	0
7	0
8	0

Abb. 13.10: Ausgangsdaten für den Binomialtest

Um diesen Test durchzuführen, bedienen wir uns der neuen Möglichkeiten, die SPSS 20 bietet, indem wir in der Variablenansicht für die Variable „sex“ im Bereich MESSNIVEAU die Kategorie NOMINAL einstellen.

Dann wählen wir Menü ANALYSIEREN/NICHTPARAMETRISCHE TESTS/EINE STICHPROBE…; beachten, dass angeklickt ist bei BEOBACHTETE UND HYPOTHETISCHE DATEN AUTOMATISCH VERGLEICHEN, und klicken auf AUSFÜHREN. Dies führt zum Befund der Abbildung 13.11:

Übersicht über Hypothesentest

	Nullhypothese	Test	Sig.	Entscheidung
1	Die durch sex = 1,00 und 0,00 definierten Kategorien treten mit den Wahrscheinlichkeiten 0,5 und 0,5 auf.	Test auf Binomialverteilung einer Stichprobe	,289[1]	Nullhypothese behalten.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist , 05.

[1]Für diesen Test wird die exakte Signifikanz angezeigt.

Abb. 13.11: Ergebnis des Binomialtests

Unter dem Stichwort SIG. wird die gesuchte Überschreitungswahrscheinlichkeit zu 0,289 (28,9%) ausgegeben. Sie ist größer als die üblichen Signifikanzniveaus, so dass die Hypothese der Gleichverteilung von Männern und Frauen nicht verworfen werden kann. Auch das hat wieder mit dem sehr kleinen Stichprobenumfang zu tun.

13.6 Exkurs: Der Chi-Quadrat-Unabhängigkeitstest

Der *Unabhängigkeitstest* wendet sich einer anderen Frage zu: Hier werden zwei Variablen in einer Zufallsstichprobe gemeinsam betrachtet, und es soll darüber entschieden werden, ob diese beiden Variablen voneinander unabhängig sind oder nicht. Ein entsprechendes Beispiel hatten Sie schon in Abschnitt 10.7 kennen gelernt: Gibt es einen Zusammenhang zwischen „Geschlecht“ und „bevorzugter politischer Partei“? Oder anders gefragt: Kann die Hypothese der Unabhängigkeit dieser beiden Variablen verworfen werden, oder ist sie zu bestätigen?

Genau genommen handelt es sich hier nicht um einen Anpassungstest – nicht umsonst wird dieses Testverfahren deshalb Unabhängigkeitstest genannt –, aber zwei Gründe sprechen dafür, dieses Verfahren hier anzusprechen: Zum einen entspricht das einzusetzende Instrumentarium dem, das Sie in Abschnitt 13.2 kennen gelernt haben, zum anderen kann die Frage der Unabhängigkeit zweier Variablen in einer Stichprobe auch so formuliert werden: Kann die bivariate Häufigkeitsverteilung, die beobachtet wurde, mit hinreichender Güte an eine bivariate theoretische Verteilung angepasst werden, nämlich an die Verteilung, die die bei Unabhängigkeit der beiden Variablen voneinander zu erwartenden Häufigkeiten präsentiert?

Ein Beispiel für den Chi-Quadrat-Unabhängigkeitstest erübrigt sich an dieser Stelle. Wir verweisen auf Abschnitt 10.7, wo sich eine entsprechende Aufgabe findet.

14 Nichtparametrische Tests

14.1 Aufgabenstellung

In Kapitel 13 haben Sie Anpassungstestverfahren kennen gelernt, die man zu den nichtparametrischen Tests zählt. Unter dieser Überschrift der nichtparametrischen Tests versammeln sich aber noch eine große Zahl weiterer Testverfahren, von denen hier einige besonders wichtige angesprochen werden sollen.

Die Sequenzanalyse untersucht, ob eine Folge dichotomer Werte zufällig ist oder einem nicht zufälligen Muster folgt. Mit diesem Test kann also beispielsweise die Frage geprüft werden, ob bei der zufälligen Auswahl von zu befragenden Personen die Abfolge von Männern und Frauen zufällig ist oder nicht, also, ob unter diesem Gesichtspunkt des Geschlechts die Zufälligkeit der Auswahl gewährleistet ist. Dieses Beispiel verdeutlicht, dass dieser Test schon bei nominalskalierten Daten verwendet werden kann, wobei allerdings vorausgesetzt werden muss, dass die Untersuchungsvariable, die betrachtet wird, vom dichotomen Typ ist, also nur zwei Ausprägungen aufweist.

Der Mann-Whitney-U-Test prüft die Frage, ob zwei Stichproben aus der gleichen Grundgesamtheit stammen oder nicht. Für diesen Test werden die Rangpositionen der Variablenausprägungen verwendet, d.h. dieser Test ist bei Ordinalskalenniveau der Untersuchungsvariablen einsetzbar, den nur bei Variablen dieser (oder einer höheren) Skalenqualität können Rangpositionen bestimmt werden.

Der Kruskall-Wallis-H-Test untersucht, ob mehrere Stichproben (mehr als zwei) der gleichen Grundgesamtheit entstammen. Die Fragestellung entspricht also derjenigen des Mann-Whitney-U-Tests, wird jetzt aber auf den in der Praxis nicht unwichtigen Fall ausgedehnt, dass mehr als zwei Stichproben vorliegen. Auch hier genügt Ordinalskalenqualität der Untersuchungsvariablen.

Der Median-Test prüft, ob mehrere Stichproben aus Grundgesamtheiten mit gleichem Median entstammen. Auch hier genügt demnach Ordinalskalenqualität der Variablen, denn der Median ist ja ein Mittelwert, der sich aus der Rangordnung der Ausgangswerte ergibt.

Der McNemar-Test ist ein Testverfahren für zwei verbundene Stichproben (siehe dazu Abschnitt 14.6), wobei untersucht wird, ob sich signifikante Änderungen in den Ausprägungen einer dichotomen Variablen beim Vergleich der einen mit der anderen Stichprobe ergeben.

Diese knappen Erläuterungen verdeutlichen, wie wichtig diese (und andere, hier nicht besprochene) nichtparametrische Testverfahren für die statistische Praxis sind. Da metrisches Skalenniveau nicht vorausgesetzt wird, sind sie auch bei Daten mit schwächerer Skalenqualität, wie sie ja etwa im Bereich der empirischen Sozialforschung nicht selten sind, einsetzbar.

14.2 Sequenzanalyse

Betrachten Sie das folgende Beispiel: 20 Personen werden zufällig ausgewählt, wobei unter anderem das Geschlecht erfasst wird. Wenn wir die Ausprägung „männlich“ mit 0 und

„weiblich“ mit 1 kodieren, könnte sich eine Folge von Nullen und Einsen ergeben, die sich wie folgt darstellt:

1 1 0 0 0 1 0 1 0 1 1 1 1 0 0 1 0 0 1 0

Ist diese Abfolge zufällig oder nicht? Die hier zu prüfende Nullhypothese geht also davon aus, dass Einsen und Nullen im Wechsel auftreten müssten, wobei unterstellt wird, dass es gleichwahrscheinlich ist, ob die nächste befragte Person männlichen oder weiblichen Geschlechts ist. Um die Sequenzanalyse mit SPSS durchzuführen, gehen Sie folgendermaßen vor:

1. Geben Sie die obigen Daten in die erste Spalte einer neuen Tabelle ein und legen Sie das Messniveau auf NOMINAL fest..
2. Wählen Sie ANALYSIEREN/NICHTPARAMAMETRISCHE TESTS/EINE STICHPROBE...
3. Klicken Sie an bei SEQUENZ AUF ZUFÄLLIGKEIT ÜBERPRÜFEN.
4. Klicken Sie auf AUSFÜHREN.

Übersicht über Hypothesentest

	Nullhypothese	Test	Sig.	Entscheidung
1	Die durch sex = (1,00) und (0,00) definierte Wertesequenz ist zufällig.	Sequenztest einer Stichprobe	,818	Nullhypothese behalten.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Abb. 14.1: Ergebnis der Sequenzanalyse

In Abbildung 14.1 ergibt sich als Überschreitungswahrscheinlichkeit 0,818. Dies bedeutet, dass die Hypothese der Zufälligkeit der Abfolge von Frauen und Männern nicht verworfen werden kann.

14.3 Mann-Whitney-U-Test

Dieser Test eignet sich für Hypothesen, die sich auf die Daten von zwei voneinander unabhängigen Stichproben stützen. Die beiden Stichproben werden aus dem Ausgangsdatenbestand durch zwei Gruppen einer Kontrollvariablen gebildet. Erinnern Sie sich an das Ausgangsbeispiel aus Kapitel 2. Dort wurde unter anderem nach dem Geschlecht gefragt, so dass die Gesamtstichprobe unter Nutzung dieser Variablen in zwei Teilstichproben zerlegt werden kann, eine „Frauen-Stichprobe“ und eine „Männer-Stichprobe“. Diese beiden Teilstichproben sind voneinander unabhängig, weil die einzelnen Personen ja nach dem Zufallsprinzip in die Gesamtstichprobe aufgenommen wurden.

Die Nullhypothese, die jetzt geprüft werden kann, lautet:

Beide Teilstichproben stammen aus nur einer Grundgesamtheit, bzw. die „Frauen-Grundgesamtheit“ unterscheidet sich nicht von der „Männer-Grundgesamtheit“.

Beispielsweise können Sie jetzt also untersuchen, ob die Verteilung der Bildungsabschlüsse (Variable „Bildung“) bei Frauen und Männern gleich ist. Der Test selbst orientiert sich an

der folgenden Überlegung: Die Ausprägungen der Variablen „Bildung“ werden in beiden Teilstichproben (Frauen und Männer) zunächst aufsteigend geordnet (Sie sehen, dass Ordinalskalenqualität der Untersuchungsvariablen, hier also der Variablen „Bildung“, erforderlich ist). Die Hypothesenentscheidung kommt so zustande, dass die Summen der Rangzahlen, bzw. die beiden Rangzahlen-Durchschnitte, in beiden Gruppen miteinander verglichen werden. Je weiter sie voneinander abweichen, desto eher ist die Nullhypothese zu verwerfen.

1. Legen Sie das Messniveau auf der Variablen „Bildung“ aus der Datei B00.SAV auf SKALA fest..
2. Wählen Sie ANALYSIEREN/NICHTPARAMAMETRISCHE TESTS/UNABHÄNGIGE STICHPROBEN...
3. Klicken Sie auf FELDER und übertragen Sie die Variable „bildung“ in den Bereich TESTFELDER:
4. Übertragen Sie die Variable „sex“ in den Bereich GRUPPEN:
5. Klicken Sie auf AUSFÜHREN.

Übersicht über Hypothesentest

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von Letzter Bildungsabschluß ist über Kategorien von Geschlecht gleich.	Mann-Whitney-U-Test unabhängiger Stichproben	,139	Nullhypothese behalten.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Abb. 14.2: Ergebnis des Mann-Whitney-U-Tests

Sie sehen in Abbildung 14.2 wieder die interessierende Überschreitungswahrscheinlichkeit, die sich hier zu 0,139 (13,9%) ergibt. Bei einem üblichen Signifikanzniveau von 5% kann deshalb die Hypothese, dass sich die Bildungsabschlüsse bei Männern und Frauen nicht unterscheiden, nicht verworfen werden.

14.4 Kruskall-Wallis-H-Test

Dieser Test wendet sich der gleichen Aufgabe zu wie der gerade besprochene Mann-Whitney-U-Test, dehnt die Betrachtung jetzt auf mehr als zwei voneinander unabhängige Stichproben aus. Sie könnten also beispielsweise die Frage aufgreifen, ob die Verteilungen der Bildungsabschlüsse gleich sind, wenn man die Gesamtstichprobe mittels der Variablen „Konfession“ in vier Teilstichproben zerlegt (1=katholisch, 2=evangelisch, 3=sonstiges, 4=konfessionslos). Wenn Sie es richtig bedenken, entspricht dies der Hypothese, dass die beiden Variablen „Konfession“ und „zuletzt erreichter Bildungsabschluss“ nichts miteinander zu tun haben, was auch mit einem Chi-Quadrat-Unabhängigkeitstest geprüft werden könnte, vorausgesetzt, die bei Unabhängigkeit der beiden Variablen voneinander zu erwartenden Häufigkeiten sind alle größer als 5. Der Kruskall-Wallis-H-Test prüft diese Hypothese wieder über den Vergleich der mittleren Rangzahlen. Beim Einsatz von SPSS ist wie folgt vorzugehen:

1. Wählen Sie ANALYSIEREN/NICHTPARAMAMETRISCHE TESTS/UNABHÄNGIGE STICHPROBEN...
2. Klicken Sie auf ZIEL und wählen Sie VERTEILUNGEN ZWISCHEN GRUPPEN AUTOMATISCH VERGLEICHEN.
3. Klicken Sie auf FELDER und übertragen Sie die Variable „bildung“ in den Bereich TESTFELDER:
4. Übertragen Sie die Variable „konf“ in den Bereich GRUPPEN:
5. Klicken Sie auf AUSFÜHREN.

Übersicht über Hypothesentest

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von Letzter Bildungsabschluß ist über Kategorien von Konfession gleich.	Kruskal-Wallis-Test unabhängiger Stichproben	,628	Nullhypothese behalten.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Abb. 14.3: Ergebnis des Kruskal-Wallis-Tests

Der Blick auf die Überschreitungswahrscheinlichkeit zeigt den Wert von 0,628 (62,8%), so dass die Hypothese der Gleichheit der vier Teil-Grundgesamtheiten nicht verworfen werden kann.

14.5 Median-Test

Der Median-Test ist für Aufgabenstellungen geeignet, die denen des Mann-Whitney-U-Tests und des Kruskall-Wallis-H-Tests entsprechen. Bei der Konstruktion der Testgröße geht man jetzt aber nicht von den Rangzahlen der Untersuchungsvariablen aus, sondern der Median-Test geht von den Medianen der einzelnen Gruppen aus. Geprüft wird also die Hypothese, dass sich die Mediane (es wird also wieder eine mindestens ordinalskalierte Untersuchungsvariable vorausgesetzt) in den Grundgesamtheiten, aus denen Teilstichproben entnommen wurden, nicht signifikant voneinander unterscheiden.

Betrachten Sie noch einmal die „Männer-“ und die „Frauen-Stichprobe“. Es soll die Hypothese geprüft werden, dass die Zentralwerte (Mediane) der Variablen „Alter“ sich nicht signifikant voneinander unterscheiden. Da die Variable „Alter“ metrisch ist, hätte eine entsprechende Hypothese sich auch auf Mittelwertunterschiede beziehen können, wie sie in Abschnitt 12.4 schon behandelt wurde.

1. Wählen Sie ANALYSIEREN/NICHTPARAMAMETRISCHE TESTS/UNABHÄNGIGE STICHPROBEN...
2. Klicken Sie auf ZIEL und wählen Sie MEDIANE ZWISCHEN GRUPPEN VERGLEICHEN.
3. Klicken Sie auf FELDER und übertragen Sie die Variable „alter“ in den Bereich TESTFELDER:
4. Übertragen Sie die Variable „sex“ in den Bereich GRUPPEN:
5. Klicken Sie auf AUSFÜHREN.

Übersicht über Hypothesentest

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Medianwerte von alter sind über Kategorien von Geschlecht gleich.	Mediantest unabhängiger Stichproben	,943	Nullhypothese behalten.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Abb. 14.4: Ergebnis des Median-Tests

Anhand der hohen Überschreitungswahrscheinlichkeit in Abbildung 14.4 erkennen Sie, dass es keinen signifikanten Unterschied der Mediane des Alters zwischen männlichen und weiblichen Befragten gibt.

14.6 McNemar-Test

Dieser Test behandelt ein anderes Problem als alle bisher besprochenen Testverfahren, nämlich das Problem *verbundener Stichproben*. Dem Zwei- oder Mehrstichprobenfall sind Sie schon mehrfach begegnet, aber dabei handelte es sich immer um voneinander unabhängige Stichproben. Dieses Konzept wird nun aufgegeben, denn es werden zwei Stichproben betrachtet, die sich gegenseitig beeinflussen. Das klassische Beispiel für eine solche Aufgabenstellung ist die zweimalige Befragung der gleichen Personengruppe.

Stellen Sie sich folgende Situation vor: 20 zufällig ausgewählte Hausfrauen werden im Zuge einer Produktanalyse danach befragt, ob ihnen die Verpackung des Waschpulvers „Grellweiß“ gefällt. Es mögen 6 dieser Hausfrauen diese Frage bejahen, 14 verneinen sie. Daraufhin zaubert der Produktmanager aus einer Schublade eine zweite Verpackung hervor, die ganz anders gestylt ist, und fragt die gleichen 20 Hausfrauen erneut nach ihrer Meinung. Möglicherweise sind jetzt 13 Hausfrauen für diese Packung, 7 dagegen. Es interessiert jetzt natürlich die Frage, ob der beobachtete Anteilswertunterschied (Anteile der Befürworterinnen) statistisch signifikant ist. Bei der ersten Befragung lag dieser Anteil bei 30%, in der zweiten Befragung bei 65%.

Mit den bisher besprochenen Verfahren lässt sich diese Frage nicht beantworten, weil sie ja alle auf dem Konzept voneinander unabhängiger Stichproben beruhten. Die beiden Befragungen, die hier durchgeführt wurden, sind aber im höchsten Grade voneinander abhängig, schlicht und einfach deshalb, weil beide Male die gleichen Personen befragt wurden.

Hier kommt nun der McNemar-Test zum Zuge, der von der Anzahl derjenigen Befragten ausgeht, die von der ersten zur zweiten Befragung ihre Meinung geändert haben. Der Einsatz dieses Tests setzt also voraus, dass nicht nur die Anteile der Ja- und Nein-Sagerinnen, sondern auch die Anzahlen der sog. „Wechsler“ erfasst werden. Diese Befunde können schematisch in einer Tabelle erfasst werden, die folgendermaßen aussehen könnte:

erste Befragung	ja	nein	Summe
zweite Befragung			
ja	5	*8*	13
nein	*1*	6	7
Summe	6	14	

Die Befragungsergebnisse, die einleitend zu diesem Beispiel genannt wurden, finden sich in dieser Tabelle in der Summenzeile (erste Befragung) bzw. in der Summenspalte (zweite Befragung). Innerhalb der Tabelle finden Sie:

1. Zahl derjenigen, die beim 1. und beim 2. Mal „ja“ gesagt haben: 5
2. Zahl derjenigen, die beim 1. und beim 2. Mal „nein“ gesagt haben: 6
3. Zahl derjenigen, die beim 1. Mal „ja“ und beim 2. Mal „nein“ gesagt haben: 1
4. Zahl derjenigen, die beim 1. Mal „nein“ und beim 2. Mal „ja“ gesagt haben: 8

Hier interessieren nun die Anzahlen der „Wechsler“ (1 und 8). Würde die Nullhypothese zutreffen (die Verpackungsart hat keinen Einfluss auf den „Ja“-Anteil), dann wäre zu erwarten, dass die Zahl der Wechsler in die eine Richtung mit derjenigen in die andere Richtung übereinstimmt. Je weiter die beiden Wechslerzahlen voneinander abweichen, desto eher wird die Nullhypothese verworfen.

Da wir insgesamt 9 Wechsler haben, wäre bei Gültigkeit der Nullhypothese zu erwarten, dass 4,5 in die eine, 4,5 in die andere Richtung wechseln. Zur Testentscheidung muss deshalb die Frage beantwortet werden, wie wahrscheinlich es ist, dass nur eine Person (oder weniger, d.h. oder noch weiter von der Nullhypothese abweichend) von 9 Personen in die eine Richtung wechselt, wenn die Wahrscheinlichkeit für einen Wechsel in die eine oder in die andere Richtung bei einer Person bei 0,5 liegt.

Zuständig für die Beantwortung dieser Frage ist die Binomialverteilung, die der McNemar-Test auch nutzt, und die Sie schon im Zusammenhang mit dem Binomialtest in Abschnitt 13.5 kennen gelernt haben. Bei hinreichend großen Fallzahlen wird ersatzweise von SPSS ein Chi-Quadrat-Test genutzt.

Ausgehend von dem obigen Beispiel kann nun SPSS eingesetzt werden, wobei zwei Variablen definiert werden („vorher“ und „nachher“), die der Einfachheit halber mit 0 (für „ja“) und 1 (für „nein“) kodiert werden. Dies führt zur Ausgangstabelle der Abbildung 14.5.

Um den McNemar-Test durchzuführen, sind die folgenden Schritte erforderlich:

1. Wählen Sie ANALYSIEREN/NICHTPARAMAMETRISCHE TESTS/VERBUNDENE STICHPROBEN...
2. Klicken Sie auf ZIEL und wählen Sie BEOBACHTETE UND HYPOTHETISCHE DATEN AUTOMATISCH VERGLEICHEN.
3. Klicken Sie auf FELDER und übertragen Sie die Variablen „vorher“ und „nachher“ in den Bereich TESTFELDER:
4. Klicken Sie auf AUSFÜHREN.

	vorher	nachher
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	1	0
7	1	0
8	1	0
9	1	0
10	1	0
11	1	0
12	1	0
13	1	0
14	0	1
15	1	1
16	1	1
17	1	1
18	1	1
19	1	1
20	1	1

Abb. 14.5: Ausgangsdaten für den McNemar-Test

Übersicht über Hypothesentest

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilungen von verschiedenen Werten über vorher und nachher sind gleich wahrscheinlich.	McNemar-Test verbundener Stichproben	,039[1]	Nullhypothese ablehnen.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist , 05.

[1]Für diesen Test wird die exakte Signifikanz angezeigt.

Abb. 14.6: Ergebnis des McNemar-Tests

Die Überschreitungswahrscheinlichkeit liegt in diesem Beispiel bei nur 0,039 (3,9%). Dies bedeutet, dass die Nullhypothese, dass die Wechsleranzahlen zwischen den beiden Befragungen gleich sind, dass also die Änderung der Waschmittelverpackung zwischen den beiden Befragungen keinen signifikanten Einfluss gehabt hätte, zu verwerfen ist.

Der Vollständigkeit halber sei darauf aufmerksam gemacht, dass derartige Fragestellungen auch über einen *Wilcoxon-Test* oder über einen *Vorzeichentest* bearbeitet werden können (siehe dazu *Tiede/Voß*: Schließen mit Statistik – Verstehen, Verlag Oldenbourg 2000).

15 Zeitreihenstatistik

15.1 Aufgabenstellung

Die Betrachtung und Analyse statistischer Zeitreihen ist ein weiteres wesentliches Einsatzgebiet der statistischen Methoden. Von allen bisher besprochenen Aufgaben unterscheidet sich diese in der Weise, dass nun nicht Werte einer oder mehrerer Variablen, die bestimmten Merkmalsträgern (z.B. befragten Personen) zugeordnet werden, zu betrachten sind, sondern Zeitreihenwerte.

Unter einer *Zeitreihe* versteht der Statistiker die Zuordnung der Werte einer interessierenden Variablen zu Zeitpunkten oder zu Zeitintervallen. Beispielsweise entsteht eine Zeitreihe dann, wenn Sie jeden Morgen auf die Badezimmerwaage steigen, um Ihr Gewicht zu notieren, oder wenn Sie jeden Mittag um 12 Uhr die Temperatur auf Ihrem Balkon messen, oder wenn Sie jede Stunde die Temperatur eines fiebernden Kindes messen. In diesen Beispielen werden zeitpunktsbezogene Werte erfasst. Anders ist es bei vielen wirtschaftsstatistischen Daten, z.B. beim Bruttosozialprodukt (Wert aller pro Jahr geschaffenen Güter und Dienstleistungen in einer Volkswirtschaft). Übernehmen Sie beispielsweise aus dem Statistischen Jahrbuch für Deutschland die Sozialproduktsangaben für 1950 bis 2002, so liegen zeitintervallbezogene Daten („pro Jahr“) vor.

Die Zeitreihenstatistik widmet sich nun in erster Linie den folgenden Aufgaben:

1. Grafische Präsentation von Zeitreihen
2. Bestimmung von Trendfunktionen
3. Trendprognosen

Diese und einige weitere interessierende Stichworte werden in den folgenden Abschnitten näher beleuchtet.

15.2 Zeitreihendiagramme

Im Folgenden soll das Beispiel der Bevölkerungsentwicklung in Deutschland von 1970 bis 2009 betrachtet werden. Die Ausgangsdaten finden sich in der folgenden Tabelle.

Eine solche tabellarische Darstellung, wie sie Abbildung 15.1 zeigt, ist wesentlich weniger aussagekräftig als ein Diagramm. Deshalb soll nun mit SPSS diese Zeitreihe grafisch dargestellt werden. Dazu gehen Sie wie folgt vor:

Jahr	Bev.		Jahr	Bev.		Jahr	Bev.
1970	78,1		1985	77,7		2000	82,3
1971	78,6		1986	77,8		2001	82,4
1972	78,8		1987	77,9		2002	82,5
1973	79,1		1988	78,4		2003	82,5
1974	78,9		1989	79,1		2004	82,5
1975	78,5		1990	79,8		2005	82,4
1976	78,2		1991	80,3		2006	82,3
1977	78,1		1992	81,0		2007	82,2
1978	78,1		1993	81,3		2008	82,0
1979	78,2		1994	81,5		2009	81,8
1980	78,4		1995	81,8			
1981	78,4		1996	82,0			
1982	78,2		1997	82,1			
1983	78,0		1998	82,0			
1984	77,7		1999	82,2			

1. Geben Sie die Zeitreihendaten (Jahresangaben und Variable „Bev“) in zwei Spalten einer SPSS-Tabelle untereinander ein.
2. Wählen Sie DIAGRAMME/GRAFIKTAFERL-VORAUSWAHL...
3. Klicken Sie auf Jahr und (mit der Strg-Taste) auf Bevölkerung.
4. Wählen Sie die Diagrammoption LINIE
5. Klicken Sie auf OK.

Abb. 15.1: Zeitreihendiagramm

15.3 Trendberechnung

Eine besonders wichtige Aufgabe der statistischen Zeitreihenanalyse besteht darin, aus den gegebenen Zeitreihenwerten den Trend auszurechnen. Unter einem *Trend* versteht der Sta-

tistiker die langfristige Entwicklungstendenz einer Zeitreihe, wobei von kurzfristigen, mehr oder weniger regelmäßigen oder auch zufälligen Schwankungen abgesehen wird.

Bei der Trendberechnung geht es darum, in das grafische Bild einer Zeitreihe eine Glättungslinie hineinzulegen, die die gegebene Zeitreihe in ihrer langfristigen Entwicklungsrichtung charakterisiert. Diese Aufgabe erinnert an die Bestimmung einer Regressionsfunktion (siehe Kapitel 9), und in der Tat sind die einzusetzenden Methoden im Prinzip die gleichen. Auch hier wird unterschieden zwischen dem einfachen Fall der Bestimmung einer linearen Trendfunktion und der etwas aufwendigeren Bestimmung einer nichtlinearen Trendfunktion, wobei im zweiten Fall zunächst eine Entscheidung darüber herbeigeführt werden muss, von welchem mathematischen Typ die nichtlineare Funktion sein soll. Diese Entscheidung – und auch die grundsätzliche Entscheidung, ob linear oder nichtlinear – kommt aufgrund theoretisch-inhaltlicher, sachbezogener Überlegungen und mit Blick auf das optische Bild der gegebenen Zeitreihe zustande.

Zur Illustration der Vorgehensweise greifen wir auf das obige Beispiel der deutschen Bevölkerungsentwicklung zurück. Diese wurde in Abbildung 15.1 vorgestellt.

Zur Bestimmung der Parameter einer linearen Trendfunktion verwenden wir die gleichen Schritte wie bei der Berechnung einer linearen Regressionsfunktion. Dabei ergibt sich:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		Regressionskoeffizient B	Standardfehler	Beta		
1	(Konstante)	-199,852	25,499		-7,838	,000
	Jahr	,141	,013	,872	10,978	,000

a. Abhängige Variable: Bevölkerung in Millionen

Abb. 15.2: Parameter der linearen Trendfunktion

Sie erkennen, dass der Ordinatenabschnitt mit a = -199,852 weit im negativen Bereich liegt, was darauf zurückzuführen ist, dass der Ursprung des Achsenkreuzes, das hier verwendet wird, im Jahr 0 (Christi Geburt) liegt. Mithin kann dieser a-Wert nicht sinnvoll interpretiert werden. Wichtiger ist der Steigungswinkel b = 0,141, der besagt, dass im Schnitt von Jahr zu Jahr die Bevölkerungszahl um 0,141 Millionen = 141.000 angestiegen ist.

Die mathematische Funktion yt = -199.852 + 0,141*Jahr kann nun für prognostische Zwecke verwendet werden. Wenn wir z.B. an der Stelle „Jahr“ den Wert 2015 eingeben, erhalten wir eine Prognose für das Jahr 2015. Es ergibt sich der Wert 84,263. Im Jahr 2015 ist also mit mehr als 84 Millionen Menschen zu rechnen – vorausgesetzt, der lineare Trend beschreibt die vergangene Entwicklung angemessen und gilt auch in Zukunft. Der Blick auf Abbildung 15.1 zeigt aber, dass die Angemessenheit der linearen Funktion durchaus bezweifelt werden kann.

Deshalb werden wir im folgenden Schritt die Parameter einer kubischen Parabel bestimmen, die dem grafischen Bild besser entspricht. Zur besseren Veranschaulichung werden wir aber zuvor die Variable „Jahr“ umrechnen, indem wir 1970 auf Null setzen. Dies wird möglich mit dem Menü TRANSFORMIEREN/VARIABLE BERECHNEN…, das ins Fenster der Abbildung 15.3 führt:

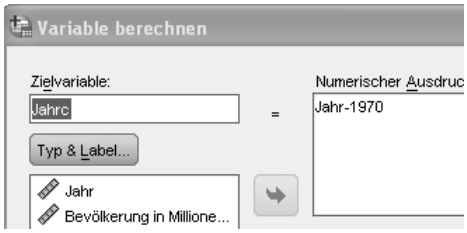


Abb. 15.3: Menü TRANSFORMIEREN/VARIABLE BERECHNEN… (Ausschnitt)

Wählen Sie dann Menü ANALYSIEREN/REGRESSION/KURVENANPASSUNG, öffnet sich das folgende Fenster:

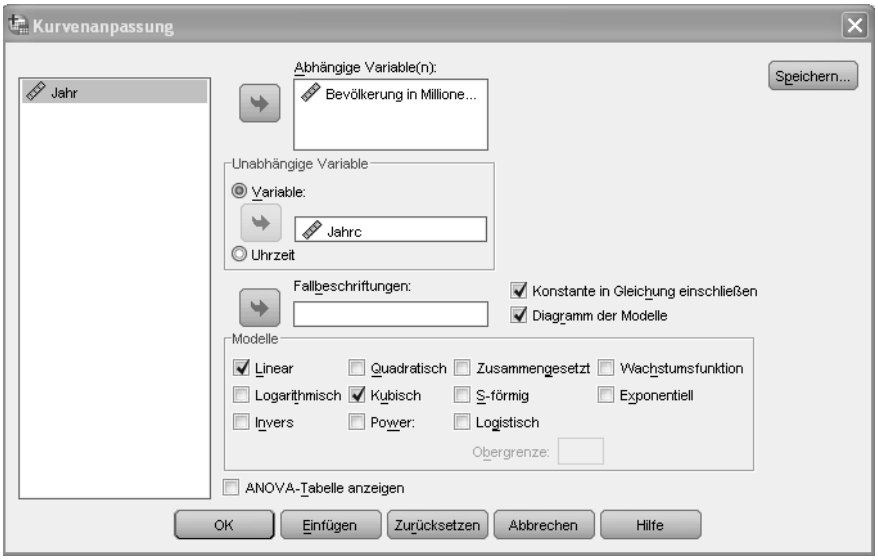


Abb. 15.4: Menü ANALYSIEREN/REGRESSION/KURVENANPASSUNG

Nehmen Sie die Eingaben so vor, wie es dieses Fenster zeigt und klicken Sie auf OK. Sie erhalten dann u.a. das folgende Diagramm, dem Sie entnehmen können, dass sich die kubische Parabel den empirischen Werten sehr viel besser anpasst als die lineare Funktion.

Die Parameter der kubischen Parabel werden ebenfalls ausgegeben, wie es Abb. 15.6 zeigt.

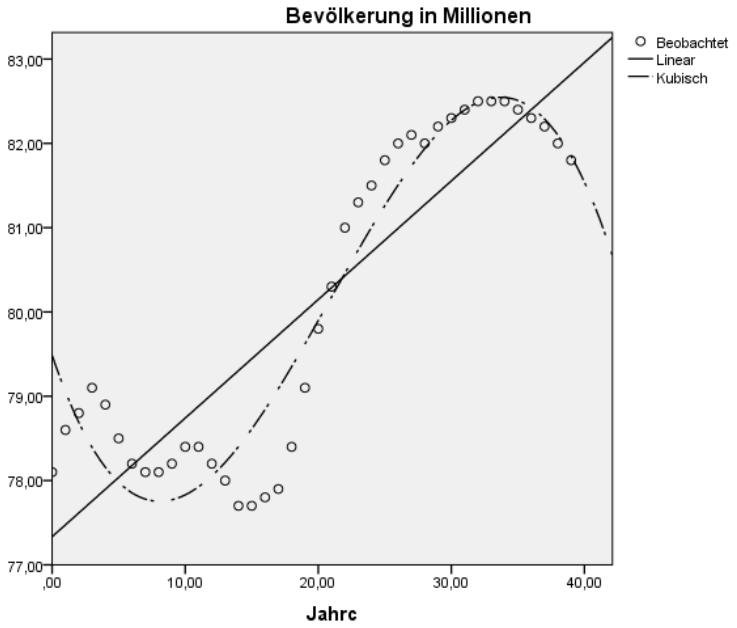


Abb. 15.5: Lineare und kubische Funktion

Parameterschätzer			
Konstante	b1	b2	b3
77,334	,141		
79,485	-,465604	,035704	-,000569

Abb. 15.6: Parameter der kubischen Funktion

Hinweis:

Wir haben in dieser Ausgabetabelle die Zahl der Dezimalstellen von standardmäßig 3 auf 6 erhöht. Dazu müssen Sie die Ausgabetabelle doppelt anklicken, dann die betreffenden Zellen markieren, danach im Menü FORMAT/ZELLENEIGENSCHAFTEN, Register FORMATWERT die Zahl der Dezimalstellen erhöhen; danach Schaltfläche ANWENDEN anklicken und OK.

Wenn Sie nun eine Prognose für 2015 mit dieser Funktion durchführen wollen, müssen Sie beachten, dass das Jahr 2015 nun durch den Wert 45 (1970 = 0) gekennzeichnet ist. Sie müssen also rechnen:

$$Yt = 79{,}485 - 0{,}465604*45 + 0{,}035704*45^2 - 0{,}000569*45^3 = 78.98$$

Es versteht sich, dass Sie nach dem gleichen Muster, wie es eben besprochen wurde, auch andere nichtlineare Funktionen erproben können, wobei besondere Aufmerksamkeit der Frage gewidmet werden muss, welcher Funktionstyp eingesetzt werden soll. Zur Beantwortung dieser Frage sind inhaltlich-theoretische Überlegungen erforderlich, die die Auswahl einer bestimmten mathematischen Funktion in hinreichender Weise begründen können.

16 Faktorenanalyse

16.1 Aufgabenstellung

In Kapitel 11 haben Sie unter den Stichworten „multiple Regression“ und „partielle Korrelation“ eine erste Idee davon gewonnen, welches die zentralen Fragestellungen der multiplen Verfahren sind, bei denen immer mehr als zwei statistische Untersuchungsvariablen gemeinsam betrachtet werden. Letztlich ging es darum – und dies wird auch in diesem und im folgenden Kapitel so sein –, durch Einbeziehung dritter, vierter, fünfter ... Untersuchungsvariablen zu informativeren Ergebnissen zu gelangen, als dies mit den Methoden der univariaten oder der bivariaten Statistik möglich ist. Es geht also um die Gewinnung zusätzlicher Erkenntnisse.

Es ist im Bereich der multivariaten Statistik ein sehr leistungsfähiges Instrumentarium, ein Bündel unterschiedlicher Verfahren, entwickelt worden, wobei insbesondere die Faktorenanalyse ganz besonders wichtig geworden ist. Während viele der bisher besprochenen Methoden notfalls noch „per Hand“ eingesetzt werden können, ist jetzt der Rechnereinsatz unumgänglich, weil die mathematischen Schwierigkeiten und insbesondere der erforderliche Zeitaufwand alle vernünftigen Grenzen sprengen würden, wollte man per Hand rechnen. Es verwundert deshalb nicht, dass diesen Verfahren der Faktorenanalyse der Durchbruch erst gelang, als die entsprechenden Rechnerkapazitäten bereitgestellt werden konnten. Hinzu kommt, dass Software benötigt wird, wie z.B. SPSS.

Zur Erläuterung der Aufgaben und der Vorgehensweise der Faktorenanalyse soll zunächst ein sehr einfaches und überschaubares Beispiel vorgestellt werden: Stellen Sie sich eine Schulklasse vor, in der mit verschiedenen Leistungstests die Leistungen der Schüler in unterschiedlichen Fächern gemessen werden. Der Einfachheit halber zählen wir die Fehleranzahl in unterschiedlichen Klassenarbeiten aus. Stellen Sie sich also einen Datenbestand vor, der so aussehen könnte, wie es Abbildung 16.1 zeigt.

	deutsch	mathe	englisch	bio	franz	physik
1	2	4	3	3	4	2
2	5	3	4	3	5	2
3	3	5	5	4	4	2
4	3	6	6	5	5	4
5	0	4	2	4	3	3
6	1	3	3	3	2	3
7	7	2	8	3	7	4
8	3	6	2	4	3	4
9	4	4	3	5	0	3
10	5	5	2	7	2	5
11	2	2	2	3	3	3
12	0	5	2	5	1	5
13	0	7	1	5	2	7
14	6	3	4	2	3	3
15	9	0	7	1	6	2
16	8	2	10	3	8	4
17	2	3	3	3	4	4
18	4	6	5	6	3	4
19	5	4	4	5	4	5
20	8	3	7	4	6	3
21						

Abb. 16.1: Fehleranzahl in Klassenarbeiten unterschiedlicher Fächer

Stellen Sie sich einmal vor, der Datenbestand wäre noch umfangreicher, als er in Abbildung 16.1 vorgestellt wurde. Es wären beispielsweise n=200 Schüler und v=14 Fächer untersucht worden. Jeder Schüler könnte dann als ein Punkt in einem 14-dimensionalen Achsenkreuz dargestellt werden, wenn man sich überhaupt ein 14-dimensionales Achsenkreuz vorstellen kann.

Die Faktorenanalyse verfolgt nun zwei Aufgaben, die, wie sich gleich zeigen wird, eng miteinander verbunden sind:

Zunächst ist es ihr Anliegen, die eventuell recht hohe Dimensionalität des Datenbestandes zu reduzieren. So wie es ja schon Aufgabe der Methoden der deskriptiven Statistik war, *Datenreduktion* zu erreichen, so geht es auch hier darum, zu einem leichter überschaubaren Bild zu gelangen. Die Faktorenanalyse bestimmt zu diesem Zweck zusätzliche, gewissermaßen künstliche Variablen, Variablen also, die im Zuge des empirischen Datengewinnungsprozesses nicht erhoben worden waren, aus den gegebenen Daten aber „herausgerechnet“ werden können. Wie hat man das zu verstehen?

Man kann sich bei dem gewählten Beispiel vorstellen, dass es vielleicht zwei oder drei zusätzliche, nicht direkt erhobene Variablen gibt, die die konkret beobachteten Ausprägungen der gemessenen Variablen hinreichend gut erklären können. Vielleicht könnten diese zusätzlichen Variablen die folgenden sein:

1. Sprachbegabung
2. Naturwissenschaftliche Fähigkeiten
3. Häuslicher Fleiß

Wenn dem so ist, dann müssten beispielsweise die Variablen „Mathe“, „Bio“ und „Physik“ einerseits und „Deutsch“, „Englisch“ und „Franz“ andererseits relativ hoch miteinander korrelieren, denn die naturwissenschaftlichen Fähigkeiten müssten sich ja positiv auf alle naturwissenschaftlichen Fächer auswirken, und entsprechend müsste sich die Sprachbegabung positiv auf alle Fächer auswirken, die etwas mit Sprachen zu tun haben (zumindest auf die fremdsprachlichen Fächer). Anders formuliert: Wenn die Leistungen in den naturwissenschaftlichen Fächern einerseits und die Leistungen in den sprachlichen Fächern andererseits hoch miteinander korrelieren, dann können „dahinter stehend“ die Größen „naturwissenschaftliche Fähigkeiten“ und „Sprachbegabung“ vermutet werden.

Diese zusätzlichen Variablen werden als „*Faktoren*“ bzw. als „*Komponenten*“ bezeichnet, und man kann sich vorstellen, dass die genannten drei Faktoren, die man mit F1, F2 und F3 bezeichnen könnte, die konkret beobachteten Ausprägungen der erhobenen Variablen relativ gut statistisch erklären könnten.

Wenn dem so ist, dann könnten die Schüler auch als Punkte in einem dreidimensionalen F1-F2-F3-Achsenkreuz dargestellt werden, was Sie sich sicherlich schon sehr viel leichter als ein 14-dimensionales Achsenkreuz vorstellen können. Damit hätte man durch Reduzierung der Dimensionalität des Datenbestandes zugleich eine sehr wesentliche Datenreduktion erreicht – die Informationen werden leichter zugänglich und überschaubarer.

Zusätzlich zur Informations- und Dimensionsreduktion verfolgt die Faktorenanalyse aber auch noch eine zweite Aufgabe: Die Faktorenanalyse erkennt zusätzliche Variablen, die Faktoren, die sich quasi hinter den tatsächlich erhobenen Variablen „verstecken“. Insoweit zählt die Faktorenanalyse zu den sog. *hypothesengenerierenden Verfahren*, weil ihr Ergebnis beispielsweise lautet: Die Leistungen der Schüler werden zu soundsoviel Prozent durch Sprachbegabung, naturwissenschaftliche Fähigkeiten und durch häuslichen Fleiß bestimmt. Somit erhält man die Grundlage und den Anlass für weitere (hypothesengeleitete) empirische Untersuchungen.

16.2 Vorgehensweise

Die Vorgehensweise der Faktorenanalyse soll am Beispiel der Abbildung 16.1 demonstriert werden. Es handelt sich bei dieser Tabelle aus mathematischer Sicht um eine Matrix, die aus n=20 Zeilen besteht (Schüler i = 1, 2, 3, ..., 20) und spaltenweise v=6 Variablen X_j (j = 1, 2, ..., 6).

Nehmen Sie einmal an, die Faktorenanalyse würde die folgenden beiden Faktoren produzieren (wir greifen im Moment einmal dem Abschluss der Beschreibung der Vorgehensweise und den entsprechenden Berechnungen vor):

F_1 = Naturwissenschaftliche Fähigkeiten

F_2 = Sprachbegabung

Diese beiden Faktoren können als zusätzliche (nicht erhobene, sondern durch die Faktorenanalyse berechnete) Variablen angesehen werden, die ihrerseits jeweils n=20 Ausprägungen aufweisen: Für jeden Schüler gibt es einen Wert des ersten und einen des zweiten Faktors.

Wenn nun die Hypothese zutrifft, dass die beiden Faktoren die Leistungen der Schüler in den verschiedenen Fächern hinreichend gut erklären können, dann gelten die folgenden funktionalen Beziehungen:

1. Die Leistungen des ersten Schülers im ersten Fach (in Deutsch) hängen ab von seinen naturwissenschaftlichen Fähigkeiten und von seiner Sprachbegabung.
2. Die Leistungen des ersten Schülers im zweiten Fach (in Mathematik) hängen ab von seinen naturwissenschaftlichen Fähigkeiten und von seiner Sprachbegabung.
3. Die Leistungen des ersten Schülers im dritten Fach .... usw.
4. Die Leistungen des zweiten Schülers im ersten Fach ... usw.

Allgemein können diese Zusammenhänge so geschrieben werden:

$$x_{ij} = f(f_{i1}, f_{i2})$$

Dabei bedeuten:

x_{ij} = Wert der Variablen j für den Schüler i

f = funktionale Abhängigkeit

f_{i1} = Wert des Faktors 1 für den Schüler i

f_{i2} = Wert des Faktors 2 für den Schüler i

Die Faktorenanalyse unterstellt nun einen linearen Zusammenhang zwischen den Faktoren (F_1 und F_2) und der einzelnen Untersuchungsvariablen X_j nach dem Muster einer linearen Regression (hier im Drei-Variablen-Fall). Über diese lineare Regressionsfunktion (im Falle von zwei Faktoren ist dies eine ebene Fläche, bei mehr als zwei Faktoren spricht man von einer Hyperfläche) werden nun x_{ij}-Werte geschätzt; sie seien xt_{ij} genannt (theoretische X-Werte). Es wird also berechnet, welcher Wert einer Variablen X_j für einen Schüler i zu erwarten ist (xt_{ij}), wenn ein linearer Zusammenhang zwischen den Faktoren und der Variablen X_j unterstellt wird:

$$xt_{ij} = f(f_{i1}, f_{i2}) = a_{j1} * f_{i1} + a_{j2} * f_{i2}$$

Die Größen a_{j1} und a_{j2} entsprechen den aus der Regressionsstatistik bekannten Regressionskoeffizienten (a_{jk} bedeutet: Regressionskoeffizient für die j-te Variable und den k-ten Fak-

tor); sie werden im Zusammenhang mit der Faktorenanalyse mit dem Begriff *Ladungen* bezeichnet.

Eine Regressionskonstante (Ordinatenabschnitt) ist hier entbehrlich, wenn, wie in der Faktorenanalyse üblich, mit standardisierten x_{ij}-Werten gearbeitet wird, weil auch schon für den bivariaten Fall einer linearen Regressionsfunktion gilt, dass der Ordinatenabschnitt 0 ist, wenn die X-Y-Werte der bivariaten Punktwolke vor den entsprechenden Berechnungen standardisiert werden. Der Vollständigkeit halber sei in diesem Zusammenhang angemerkt, dass in diesem Fall die Steigung der Regressionsfunktion mit dem Korrelationskoeffizienten zwischen X und Y (standardisiert) identisch ist.

Es leuchtet nun unmittelbar ein, dass die beiden Faktoren umso besser die Leistungen in den sechs Fächern erklären, je näher die über die obige lineare Funktion geschätzten xt_{ij}-Werte bei den beobachteten x_{ij}-Werten liegen, d.h. je kleiner die „Reste" sind, also die Differenzen x_{ij} - xt_{ij}. Damit wird die weitere Vorgehensweise deutlich: Die *Faktorladungen* a_{jk} (k = 1,2) sind so zu bestimmen, dass die xt_{ij}-Werte möglichst gut mit den beobachteten x_{ij}-Werten übereinstimmen.

Um dies zu erreichen, schlägt die Faktorenanalyse einen „Umweg" ein, der folgendermaßen skizziert werden kann: Ausgangsbasis ist die Matrix der beobachteten Werte. Aus dieser Matrix können bivariate Pearson'sche Korrelationskoeffizienten berechnet werden. Die Faktorenanalyse erzeugt nun eine zweite 6*6-Korrelationskoeffizientenmatrix, und zwar ausgehend von den theoretischen xt_{ij}-Werten. Genau genommen müssten wir sagen, dass diese zweite Matrix erzeugt werden kann, wenn die Ladungen berechnet worden sind, und wenn es deshalb möglich geworden ist, die xt_{ij}-Werte zu berechnen.

Damit sind wir am Ende der Überlegungen angelangt:

Die Faktorenanalyse bestimmt die Ladungen a_{jk}, die Zahl der Faktoren und die *Faktorwerte* für die einzelnen Merkmalsträger f_{ik} so, dass die Korrelationskoeffizientenmatrix, die auf den über die Ladungen berechenbaren xt_{ij}-Werten aufbaut, möglichst gut der Korrelationskoeffizientenmatrix entspricht, die auf den empirischen Beobachtungen, also auf den x_{ij}-Werten, aufbaut.

Dabei wird die Zahl der, wie man sagt, zu extrahierenden Faktoren so festgelegt, dass ein Mindestprozentsatz der Varianz der beobachteten Variablen statistisch erklärt wird, z.B. 80%. Bringt ein weiterer, in die Berechnungen aufgenommener Faktor keinen nennenswerten Zuwachs an Varianzerklärung (z.B. weniger als 5%), wird der Berechnungsalgorithmus abgebrochen.

Auf diese Weise kann es also geschehen, dass der faktorenanalytische Ansatz aus dem Ausgangsdatenbestand zwei Faktoren extrahiert. Leider gibt er Ihnen keine Informationen darüber, wie diese Faktoren heißen (die Idee, dass die geeigneten Etiketten vielleicht „Sprachbegabung" und „naturwissenschaftliche Fähigkeiten" sein könnten, war ja nur ein Vorgriff).

Sie können nun aber im Nachhinein feststellen, mit welchen der sechs Ausgangsvariablen die einzelnen Faktoren hoch korrelieren und mit welchen nicht. Bei zwei Faktoren und sechs Ausgangsvariablen gibt es 2*6=12 Korrelationskoeffizienten zwischen Variablen und Faktoren. In ihrer Gesamtheit werden sie als das sogenannte *Ladungsmuster* bezeichnet. Wenn zum Beispiel der erste Faktor hoch und positiv korreliert mit den Variablen „Englisch" und „Franz" (Französisch) – man sagt, der Faktor lädt hoch in Englisch und Französisch, zugleich aber nur niedrig oder gar nicht mit Mathematik und Physik (geringe Ladungen) –, dann liegt die Idee nahe, diesen Faktor mit „Sprachbegabung" zu etikettieren.

SPSS gibt u.a. das Ladungsmuster aus, so dass der empirische Forscher hier mit seiner Interpretationsarbeit beginnen kann.

Hinweis:

Die Faktoren, die die Rechenprozedur, die sich hinter den gerade vorgetragenen Überlegungen versteckt, erzeugt, sind häufig recht schwierig zu interpretieren. Wie Abschnitt 16.3 zeigt, setzt hier die eigentliche Arbeit des empirischen Sozialforschers an. Man weiß ja von Vornherein nicht, dass die Faktoren die naturwissenschaftlichen Fähigkeiten und die Sprachbegabung sein könnten; wir hatten bisher aus didaktischen Gründen nur so getan, als sei dieses Ergebnis schon bekannt. Es handelt sich ja gewissermaßen um „künstliche Variablen". Diese lassen sich allerdings verzerrungsfrei so transformieren, dass sie in unterschiedlichen Koordinatensystemen dargestellt werden können. Es kann deshalb im Anwendungsfall sinnvoll sein, sie so zu transformieren, dass ihre Verbindung zu den beobachteten Variablen (Fachleistungen) deutlicher werden, womit sie dann leichter interpretierbar werden. Man nennt diesen Arbeitsschritt *Rotation*. Im folgenden Abschnitt kommen wir darauf noch im Detail zu sprechen.

16.3 Beispiel

Wir gehen im Folgenden von den Daten der Abbildung 16.1 aus (siehe oben). Um mit SPSS eine Faktorenanalyse durchzuführen, gehen Sie wie folgt vor:

1. Geben Sie die obigen Daten in eine neue SPSS-Tabelle ein.
2. Wählen Sie ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE...

Sie gelangen zum Fenster der Abbildung 16.2.

Abb. 16.2: Menü ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE...

3. Im Fenster der Abbildung 16.2 übertragen Sie alle sechs Ausgangsvariablen in den Bereich VARIABLEN:.
4. Klicken Sie auf die Schaltfläche WERTE...

Abb. 16.3: Menü ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE..., Schaltfläche WERTE...

5. Sorgen Sie für ein Häkchen beim Stichwort ALS VARIABLEN SPEICHERN.
6. Klicken Sie WEITER an. Sie gelangen zurück zum Fenster der Abbildung 16.2.
7. Klicken Sie OK an.

SPSS erzeugt jetzt die Ergebnisse, die im Folgenden besprochen werden. Schauen Sie sich zunächst Abbildung 16.4 an.

Kommunalitäten

	Anfänglich	Extraktion
deutsch	1,000	,800
mathe	1,000	,807
englisch	1,000	,913
bio	1,000	,825
franz	1,000	,836
physik	1,000	,715

Extraktionsmethode: Hauptkomponentenanalyse.

Erklärte Gesamtvarianz

	Anfängliche Eigenwerte			Summen von quadrierten Faktorladungen für Extraktion		
Komponente	Gesamt	% der Varianz	Kumulierte %	Gesamt	% der Varianz	Kumulierte %
1	3,617	60,276	60,276	3,617	60,276	60,276
2	1,280	21,332	81,608	1,280	21,332	81,608
3	,477	7,958	89,566			
4	,384	6,405	95,971			
5	,139	2,325	98,296			
6	,102	1,704	100,000			

Extraktionsmethode: Hauptkomponentenanalyse.

Abb. 16.4: Ergebnisse der Faktorenanalyse (Teil 1)

Unter der Überschrift KOMMUNALITÄTEN gibt Ihnen SPSS in Abbildung 16.4 sehr wichtige Informationen aus: Sie sehen, dass in der ersten Spalte alle Untersuchungsvariablen aufgelistet werden. Unter der Überschrift Extraktion finden sich die sog. Kommunalitäten. Dies ist einer der zentralen Begriffe der Faktorenanalyse.

Unter *Kommunalität* versteht man den Anteil der Varianz der einzelnen Untersuchungsvariablen, der durch die Faktoren, die SPSS bestimmt hat, statistisch erklärt wird. Wir kommen darauf gleich noch einmal zu sprechen. Hier aber möchten wir darauf aufmerksam ma-

chen, dass SPSS intern immer mit den standardisierten Werten der einzelnen Variablen arbeitet. Damit werden die unterschiedlichen Variablen auf vergleichbares Niveau gebracht. Offensichtlich ist nun aber die Standardabweichung einer standardisierten Variablen immer gleich 1 und somit auch die Varianz. Dies bedeutet zugleich, dass die Gesamtvarianz aller sechs Variablen den Wert 6 hat.

Der erste angegebene Wert (0,800) besagt, dass 80% der Varianz der ersten Variablen („deutsch“) durch die Gesamtheit der extrahierten Faktoren statistisch erklärt wird. Entsprechend sind die anderen Werte der Abbildung 16.4 zu interpretieren.

Besonders wichtig ist nun die zweite Tabelle unter der Überschrift ERKLÄRTE GESAMTVARIANZ. Hier werden zunächst die sog. *Eigenwerte* vorgestellt.

In diesem Teil der Ausgabe ist festzustellen, dass SPSS sechs Faktoren (Komponenten) extrahiert, die untereinander durchnummeriert ausgegeben werden. Zunächst muss deshalb darauf aufmerksam gemacht werden, dass es so aussieht, als hätte SPSS mit seiner Faktorenanalyse das erste Ziel dieser Methode, nämlich das der Informationsreduktion, verfehlt. Wenn nämlich die ursprünglichen sechs Untersuchungsvariablen durch sechs neue Variablen (die Faktoren) erklärt werden, ist eine Reduzierung der Dimensionalität des ursprünglichen Achsenkreuzes offenbar nicht gelungen.

Trotzdem sind die hier vorgestellten Informationen wichtig: Sie besagen nämlich, dass mit sechs Faktoren die Ausprägungen der sechs Untersuchungsvariablen völlig erklärt werden können. Sie erkennen das mit Blick auf den letzten Wert unter der Überschrift KUMULIERTE%, der bei 100% liegt.

Wie man nun gleichwohl auch das Ziel der Informationsreduktion erreichen kann, wird gleich besprochen. Zunächst aber noch ein Blick auf die übrigen Ergebnisse dieser zweiten Tabelle.

Unter der Überschrift ANFÄNGLICHE EIGENWERTE GESAMT werden für die sechs extrahierten Faktoren die Werte 3,617, 1,280 usw. ausgegeben. Auch das sind sehr wichtige Angaben. Der *Eigenwert* eines Faktors nämlich gibt an, welchen Anteil dieser eine Faktor an der Gesamtvarianz aller Untersuchungsvariablen aufklären kann.

Weiter oben wurde erwähnt, dass die Streuungen der Untersuchungsvariablen alle mit 1 zu bemessen sind, weil die Ausgangsvariablen standardisiert wurden. Daraus ergibt sich, dass die Gesamtstreuung von sechs Untersuchungsvariablen den Wert 6*1 = 6 aufweisen muss. Wenn nun unter dem Stichwort EIGENWERT von SPSS für den ersten Faktor der Wert 3,617 angegeben wird, so ist das ein Anteil von 60,276% vom Gesamtvarianzwert 6. Dieser Wert 60,276 findet sich in der Spalte mit der Überschrift % DER VARIANZin der vorangegangenen Abbildung. Der erste von SPSS extrahierte Faktor erklärt also 60,276% der Gesamtvarianz der sechs Untersuchungsvariablen (dies ist – nebenbei bemerkt – sehr viel). Der zweite Faktor erklärt weitere 21,332%.

In der nächsten Spalte sind diese Werte kumuliert (KUMULIERTE %), so dass Sie also sofort ablesen können, dass die beiden ersten Faktoren zusammen 81,608% der Gesamtstreuung der beobachteten Variablen statistisch erklären.

SPSS ordnet die Faktoren so an, dass der mit der höchsten Erklärungskraft ganz oben steht, dann folgt der mit der zweithöchsten Erklärungskraft usw.

Sie erkennen, dass ab dem dritten Faktor die Eigenwerte kleiner als 1 sind. Dies bedeutet, dass der dritte Faktor (und die folgenden erst recht) nicht in der Lage ist, mehr als die Varianz auch nur einer der beobachteten Untersuchungsvariablen statistisch zu erklären. Deshalb argumentiert man häufig wie folgt:

Faktoren, die einen Eigenwert aufweisen, der kleiner ist als 1, werden im Weiteren nicht berücksichtigt. Sie taugen noch nicht einmal dazu, eine Variable zu erklären, geschweige denn mehrere (Sie erinnern sich: Man sucht ja hinter den Variablen stehende Faktoren, die die Variablen, wenigstens einige davon, gemeinsam erklären können – „gemeinsam" bedeutet, mehr als eine Variable).

So kommt man nun auch, durch Ausscheiden der Faktoren mit Eigenwerten kleiner als 1, dem Ziel der Informationsreduktion näher. Deshalb werden im rechten Teil der Tabelle auch nur noch die ersten beiden Faktoren aufgelistet.

Des Weiteren wird die sog. *Komponentenmatrix* (*Faktormatrix*) ausgegeben, die an anderer Stelle als *Ladungsmuster* bezeichnet wurde (siehe Abbildung 16.5).

Komponentenmatrix[a]

	Komponente	
	1	2
deutsch	,804	,392
mathe	-,838	,324
englisch	,828	,477
bio	-,736	,532
franz	,835	,371
physik	-,586	,610

Extraktionsmethode: Hauptkomponentenanalyse.

a. 2 Komponenten extrahiert

Abb. 16.5: Komponentenmatrix (Ladungsmuster)

Hier werden die Koeffizienten angegeben (die Ladungen, wie man auch sagt), die in der Gleichung

$$xt_{ij} = f(f_{i1}, f_{i2}) = a_{j1} * f_{i1} + a_{j2} * f_{i2}$$

mit a_{j1} und a_{j2} bezeichnet wurden (für j von 1 bis 6 Variablen).

Diese Ladungen sind interpretierbar wie Regressionskoeffizienten (Steigungen einer Regressions-Hyperebene), und sie sind zugleich, weil SPSS von standardisierten Werten ausgeht, Korrelationskoeffizienten. Somit können Sie beispielsweise sagen, dass der Zusammenhang zwischen Faktor 1 (was immer das sein mag, wie auch immer er etikettiert werden mag) und der Variablen „Biologie" mit r = -0,736 bemessen werden kann. Entsprechend sind auch die anderen Koeffizienten zu interpretieren.

Am Ladungsmuster setzt die Interpretation der Faktoren an:

Sie sehen, dass Faktor 1 (Komponente 1) gleichgerichtet korreliert mit „Deutsch", „Englisch" und „Französisch", hingegen gegenläufig mit „Biologie", „Mathematik" und „Physik".

Nicht so eindeutig sieht das bei Faktor 2 aus: Er korreliert am höchsten mit „Biologie" und „Physik", weniger hoch dagegen mit den anderen Variablen.

Es liegt deshalb der interpretatorische Gedanke nahe, dass Faktor 1 etwas mit allen sprachlichen Fächern zu tun haben muss, Faktor 2 hingegen eher etwas mit den naturwissenschaftlichen Fächern – und deshalb kann jetzt dem Faktor 1 vielleicht das Etikett „Sprachbegabung" und dem Faktor 2 das Etikett „naturwissenschaftliche Fähigkeiten" zugewiesen werden.

Weiterhin bestimmt SPSS die Faktorwerte und speichert diese (siehe Abbildung 16.6).

	deutsch	mathe	englisch	bio	franz	physik	FAC1_1	FAC2_1
1	2	4	3	3	4	2	,08285	-1,19093
2	5	3	4	3	5	2	,67043	-,70321
3	3	5	5	4	4	2	,07640	-,32623
4	3	6	6	5	5	4	-,24510	1,16541
5	0	4	2	4	3	3	-,56126	-1,04503
6	1	3	3	3	2	3	-,22197	-1,36481
7	7	2	8	3	7	4	1,32603	1,03565
8	3	6	2	4	3	4	-,71525	-,04426
9	4	4	3	5	0	3	-,63438	-,58581
10	5	5	2	7	2	5	-1,09541	1,14493
11	2	2	2	3	3	3	,01286	-1,41124
12	0	5	2	5	1	5	-1,32556	-,14527
13	0	7	1	5	2	7	-1,82864	,88602
14	6	3	4	2	3	3	,53679	-,80329
15	9	0	7	1	6	2	2,08734	-,67552
16	8	2	10	3	8	4	1,71385	1,60405
17	2	3	3	3	4	4	-,03719	-,58887
18	4	6	5	6	3	4	-,63590	1,12474
19	5	4	4	5	4	5	-,24965	1,01110
20	8	3	7	4	6	3	1,04377	,91256

Abb. 16.6: Veränderte Ausgangstabelle

Unter den Spaltennamen FAC1_1 und FAC2_1 sind in der Ausgangstabelle der Abbildung 16.6 standardisierte Werte für die einzelnen Schüler angefügt worden. Dies sind die Werte der „künstlichen“ Variablen, also der Faktoren.

Wenn man diese beiden neuen Variablen miteinander korreliert, ergibt sich der Wert 0, d.h. die von SPSS extrahierten Faktoren korrelieren nicht miteinander. Auch dies ist einer der wichtigsten Aspekte der Faktorenanalyse: Die angestrebte Informationsreduktion durch Verminderung der Dimensionalität des Datenbestandes kommt in der Weise zustande, dass künstliche Variablen erzeugt werden, die Faktoren, die die Eigenschaft haben, nicht miteinander zu korrelieren – sie sind unabhängig voneinander.

Dies erinnert an das Problem der *Multikollinearität*, auf das wir schon früher aufmerksam gemacht hatten. Es wurde schon festgestellt, dass im Mehr-Variablen-Fall, also bei der Erklärung einer interessierenden Untersuchungsvariablen durch mehrere erklärende Variablen, solche zweckmäßigerweise ausgeschieden werden, die mit anderen hoch korrelieren, weil diese anderen ja die auszuscheidende Variable in einem solchen Fall hinreichend repräsentieren. Wünschenswert wäre es, wenn die beeinflussenden Variablen ihrerseits alle voneinander unabhängig wären, was in der praktischen Anwendung aber so gut wie nie der Fall sein dürfte. Nun haben Sie aber mit der Faktorenanalyse ein Instrument in der Hand, welches unter anderem dazu taugt, voneinander unabhängige Variablen, die Faktoren, zu produzieren. Geht man mit diesen neuen Variablen in ein multiples Regressionsmodell, kann das Problem der Multikollinearität nicht mehr auftauchen.

16.4 Schlussbemerkungen

Zum Abschluss dieses Kapitels sind einige Anmerkungen angebracht, die sich auf die Durchführung einer Faktorenanalyse und auf die Interpretation der Ergebnisse beziehen:

1. Extraktion

Zunächst muss festgehalten werden, dass die Faktorenanalyse genau genommen eine ganze Gruppe von Verfahren umfasst, weil beispielsweise in Abhängigkeit von eventuell verwendeten Rotationsalgorithmen und unterschiedlichen Extraktionsverfahren die Ergebnisse ganz verschieden ausfallen können. Zu Details verweisen wir auf die zuständige Spezialliteratur (siehe z.B. *Revenstorf*: Lehrbuch der Faktorenanalyse, Stuttgart 1976; *Überla*: Faktorenanalyse, Berlin/Heidelberg/New York, 1971).

Die verschiedenen angebotenen Extraktionsmethoden erkennen Sie über ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE..., Schaltfläche EXTRAKTION..., Stichwort METHODE... (siehe Abbildung 16.7), wenn Sie dort auf den Listenpfeil klicken.

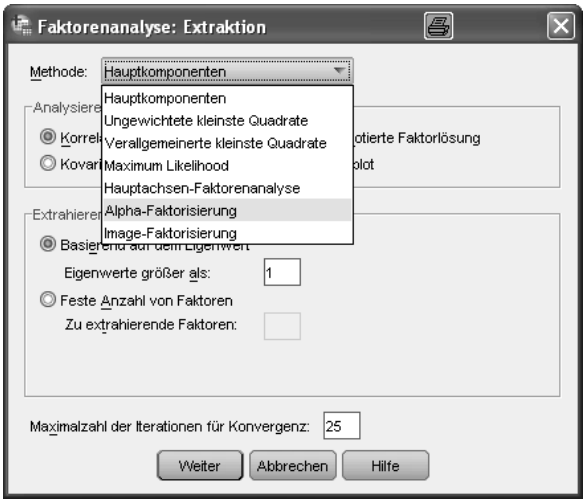


Abb. 16.7: Menü ANALYSIEREN/DIMENSIONSREDZEIRUNG/FAKTORENANALYSE..., Schaltfläche EXTRAKTION..., Stichwort METHODE:

Die einzelnen Extraktionsmethoden sollen hier nicht besprochen werden.

2. Rotation

Bei dem Beispiel des vorangegangenen Abschnitts wurde auf eine *Rotation* verzichtet, weil die Interpretation der Faktoren auch so recht gut gelungen erscheint. In anderen, vor allem umfangreicheren Datenbeständen, mag das anders sein.

Eine Rotation wird durchgeführt, wenn Sie über ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE... im Fenster der Abbildung 16.2 (siehe oben) die Schaltfläche ROTATION... anklicken. Sie gelangen dann ins Fenster der Abbildung 16.8, wo Ihnen verschiedene Rotationsmethoden angeboten werden.

Abb. 16.8: Menü ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE..., Schaltfläche ROTATION...

Es würde zu weit führen, diese Rotationsmethoden im Detail zu besprechen (wir verweisen auf die oben genannte Spezialliteratur). Es schadet aber nichts, wenn Sie die verschiedenen Methoden einmal anhand des gegebenen Beispiels erproben und die unterschiedlichen Ergebnisse, insbesondere die Ladungsmuster, miteinander vergleichen.

3. Werte

Bei der Schaltfläche WERTE: im Fenster ANALYSIEREN/DIMENSIONSREDUZIERUNG/FAKTORENANALYSE... zeigt sich, dass auch andere Schätzmethoden zur Bestimmung der xt_{ij}-Werte verwendet werden können, als das lineare Modell, das im Beispiel unterstellt wurde. Dies zeigt Abbildung 16.3 weiter oben.

Auch auf diese Verfahren soll hier nicht näher eingegangen werden. Wir verweisen wieder auf die oben angeführte Spezialliteratur.

In Abhängigkeit von den verwendeten Algorithmen können die Ergebnisse der Faktorenanalyse sehr unterschiedlich ausfallen, was einen der wichtigsten Kritikpunkte an diesem Verfahren begründet: Die Interpretation der extrahierten Faktoren wird in weitem Maße vom Verfahren abhängig, wobei ganz unterschiedliche Ladungsmuster entstehen können.

Dies bedeutet, dass die Faktorenanalyse nicht zu eindeutigen Ergebnissen führt – zweifelsohne ein recht unbefriedigender Umstand. Zudem liegt es weitgehend im Belieben des jeweiligen Forschers, wie er die Faktoren interpretiert. Beispielsweise wurden die im obigen Beispiel extrahierten Faktoren mit „Sprachbegabung“ und „naturwissenschaftliche Fähigkeiten“ etikettiert. Es könnte aber auch durchaus sein, dass der erste Faktor mit Auslandsaufenthalten der befragten Schüler, der zweite mit ihrem Geschlecht zu tun hat (vielleicht sind Knaben in den mathematisch-naturwissenschaftlichen Fächern ja prinzipiell besser als Mädchen). Sie sehen, genau genommen bieten die Ergebnisse der Faktorenanalyse Anlass und sogar die Notwendigkeit zu weiterführenden empirischen Untersuchungen. Nicht zuletzt deshalb werden sie als „hypothesengenerierende Verfahren“ bezeichnet.

17 Clusteranalyse

17.1 Aufgabenstellung

Auch die Clusteranalyse gehört, wie die im vorangegangenen Kapitel besprochene Faktorenanalyse, zu den multivariaten statistischen Verfahren. Zentrales Ziel der clusteranalytischen Verfahren ist die Beantwortung der Frage, ob sich die Merkmalsträger eines Datenbestandes in systematischer Weise klassifizieren lassen.

Wenn man beispielsweise anhand der Ausprägungen der verschiedenen Untersuchungsvariablen feststellen kann, dass sich die Merkmalsträger in zwei oder drei Gruppen unterteilen lassen, so kann ein solches Gruppierungsergebnis zum Anlass genommen werden, auf interpretatorischem Weg die Frage zu beantworten, warum sich die beobachtete Gruppenbildung ergeben hat. Welches sind die dafür verantwortlichen Größen?

Sie erkennen an dieser Frage, dass es auch hier darum geht, hinter den erhobenen Daten sich versteckende Größen aufzuspüren, wie bei der Faktorenanalyse auch, jetzt aber nicht von gemeinsamen Korrelationen zwischen den Untersuchungsvariablen ausgehend, sondern von dem Umstand der Gruppenbildung.

Ein einfaches gedankliches Beispiel mag diese Idee illustrieren: Stellen Sie sich vor, Sie hätten die beiden Untersuchungsvariablen X = „Körpergröße“ und Y = „Körpergewicht“ zufällig ausgewählter Erwachsener erfasst. Zusätzlich haben wir die Variable „Geschlecht“ erfasst (0=„männlich“, 1=„weiblich“). Ein entsprechender Datenbestand wird in Abbildung 17.1 präsentiert. Wir haben diese fiktiven Daten der Abbildung 17.1 so angeordnet, dass die ersten 10 Fälle die befragten Frauen sind, die übrigen 10 Fälle sind Männer.

	cm	kg	sex
1	165	55	1
2	161	63	1
3	167	61	1
4	165	65	1
5	170	66	1
6	158	62	1
7	171	65	1
8	166	61	1
9	168	65	1
10	163	65	1
11	163	66	0
12	182	85	0
13	179	72	0
14	191	89	0
15	179	85	0
16	175	78	0
17	183	81	0
18	185	77	0
19	182	79	0
20	188	85	0

Abb. 17.1: Körpergröße, Körpergewicht und Geschlecht

Stellt man die Variablen „Körpergröße“ und „Körpergewicht“ in einem Streudiagramm dar (siehe Abschnitt 7.3), ergibt sich das Diagramm der Abbildung 17.2.

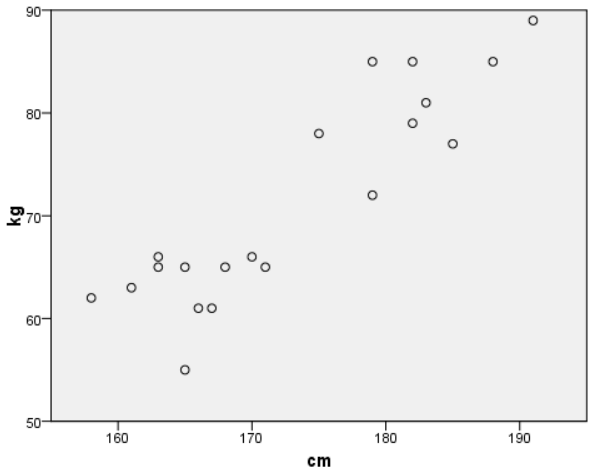


Abb. 17.2: Streudiagramm

Die Punktwolke in Abbildung 17.2 steigt tendenziell von links unten nach rechts oben im X-Y-Achsenkreuz. Allerdings zeigt sich, dass sie eigentlich aus zwei Teilpunktwolken besteht, eine liegt links unten, die andere weiter rechts und weiter oben im Achsenkreuz. Es treten Klumpungseffekte auf (Klumpen = cluster).

Warum ist das so? Sofort fällt Ihnen ein, dass es möglicherweise die Variable „Geschlecht“ sein könnte (selbst wenn diese Variable nicht erhoben worden wäre), die für diesen Klumpungseffekt verantwortlich ist. Frauen dürften im Schnitt kleiner und leichter sein als Männer. Die linkeTeilpunktwolke repräsentiert Frauen, die rechte Männer. Natürlich ist es aber nicht ausgeschlossen, dass sich die beiden Teilpunktwolken in bestimmten Bereichen überschneiden und durchmischen (kleine, leichte Männer und große, schwere Frauen).

Hinter den beiden dargestellten Untersuchungsvariablen versteckt sich also die dritte Variable „Geschlecht“. Ziel der Clusteranalyse ist es nun, einer solchen dritten (oder weiteren) Variablen auf die Spur zu kommen, selbst wenn sie nicht, wie in diesem kleinen Demonstrationsbeispiel, erhoben wurde.

Der Grundgedanke bei der Clusteranalyse ist also folgender: Bei n Merkmalsträgern und v Untersuchungsvariablen lässt sich der gegebene Datenbestand gedanklich als Punktwolke in einem v-dimensionalen Achsenkreuz vorstellen. Weiterhin ist vorstellbar, dass sich die n Punkte in bestimmten Teilbereichen klumpen. Gibt es mehrere dieser Zusammenballungen, so taucht die Frage auf, welche Größe(n) für die relative Homogenität innerhalb der Klumpen, bzw. für die relative Heterogenität zwischen den Klumpen, maßgeblich ist (sind).

Um die eventuellen Klumpen zu identifizieren, geht man von den Abständen zwischen den einzelnen Punkten im v-dimensionalen Achsenkreuz aus und fasst diejenigen Punkte jeweils zu einem Cluster zusammen, die nahe beieinander liegen.

17.2 Hierarchische Clusteranalyse

Für kleinere Datenbestände bietet SPSS die Methode der hierarchischen Clusterung an. Bei diesem Verfahren wird zunächst jeder einzelne Fall als ein Cluster betrachtet. Die beiden ähnlichsten Fälle, also die beiden, die die geringste Distanz voneinander aufweisen, werden dann zu einem ersten neuen Cluster zusammengefasst, was natürlich voraussetzt, dass SPSS zunächst alle Distanzen berechnet (bei 20 Fällen gibt es 190 Distanzen). Damit verringert sich die Anzahl der Cluster um 1. Von den nun noch vorhandenen Fällen werden wieder die beiden ähnlichsten zu einem neuen Cluster zusammengefasst. Dieser Schritt wird so lange wiederholt, bis es schließlich nur noch ein einziges Gesamtcluster gibt. Dieses letzte Ergebnis ist natürlich nicht besonders aufregend. Sie können diesen Prozess der Clusterung aber steuern, indem Sie dem Programm beispielsweise aufgrund theoretisch-inhaltlicher Überlegungen eine bestimmte kleinere Zahl als letzte Clusteranzahl vorgeben. Zusätzlich können Sie sich für jede Stufe der Clusterbildung mitteilen lassen, wie groß der Abstand zwischen den zusammengefassten Clustern ist. Wenn z.B. dieser Abstand ab einer bestimmten Stufe der Clusterbildung sprunghaft ansteigt, kann dies Hinweis für die sinnvolle Clusteranzahl sein.

Dieses Verfahren heißt hierarchisch, weil die Zuordnung der Fälle zu einem Cluster immer in eine Richtung erfolgt. Fälle, die einem Cluster zugeordnet wurden, bleiben in diesem Cluster, wenn weitere Fälle hinzukommen, selbst wenn sich auf späteren Stufen der Clusterbildung durch eine andere Zuordnung eine geringere Distanz ergeben könnte. Bei einem nicht-hierarchischen Verfahren hingegen können Fälle, die einem Cluster schon zugeordnet wurden, aus diesem auch wieder herausgenommen und einem anderen Cluster zugeordnet werden, wenn dies der Steigerung der Homogenität innerhalb der Cluster (Verringerung des Gesamtabstandes) dienlich ist.

Wichtig bei diesem Verfahren ist die Festlegung der Abstände zwischen den Punkten der ursprünglichen v-dimensionalen Punktwolke. Üblicherweise werden die quadrierten *Euklidischen Distanzen* verwendet (ihre Berechnung beruht auf dem Lehrsatz des Phytagoras über die Länge der Hypotenuse rechtwinkliger Dreiecke). Es gibt allerdings auch andere *Distanzmaße* (siehe Abschnitt 17.5).

Um nun vom Datenbestand der Abbildung 17.1 ausgehend eine hierarchische Clusterung mit SPSS durchzuführen, sind die folgenden Arbeitsschritte erforderlich:

1. Gebe Sie die obigen Daten in eine neue SPSS-Tabelle ein.
2. Wählen Sie ANALYSIEREN/KLASSIFIZIEREN/HIERARCHISCHE CLUSTER...

Sie gelangen ins Fenster der Abbildung 17.3.

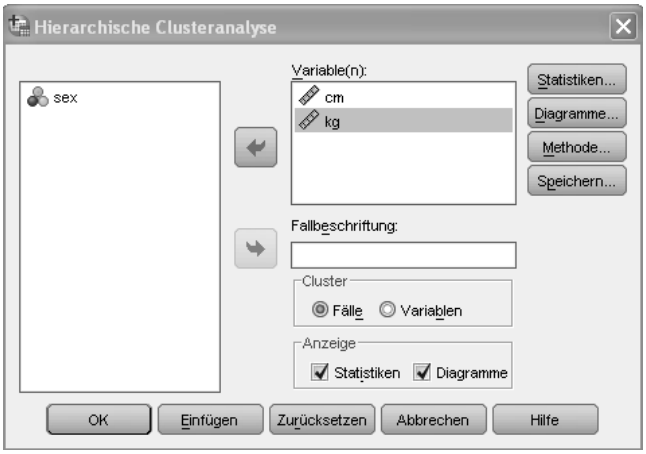


Abb. 17.3: Menü ANALYSIEREN/KLASSIFIZIEREN/HIERARCHISCHE CLUSTER...

3. Im Fenster der Abbildung 17.3 übertragen Sie die beiden Variablen „cm“ und „kg“ in den Bereich VARIABLE(N):.
4. Im Bereich CLUSTER muss der Optionsschalter bei FÄLLE angeklickt sein.
5. Sorgen Sie im Bereich ANZEIGEN für Häkchen bei STATISTIK und bei DIAGRAMME.
6. Klicken Sie die Schaltfläche STATISTIK...

Sie gelangen zum Fenster der Abbildung 17.4.

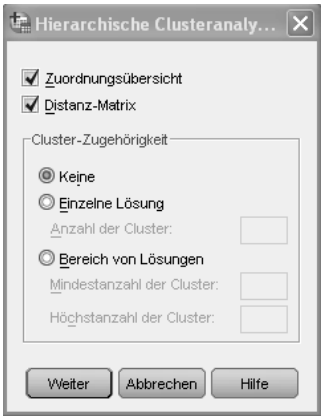


Abb. 17.4: Menü ANALYSIEREN/KLASSIFIZIEREN/HIERARCHISCHE CLUSTER..., Schaltfläche STATISTIK...

7. Im Fenster der Abbildung 17.4 sorgen Sie für Häkchen bei ZUORDNUNGSÜBERSICHT und bei DISTANZ-MATRIX.
8. Klicken Sie auf WEITER. Sie gelangen zum Fenster der Abbildung 17.3 zurück.
9. Klicken Sie auf die Schaltfläche DIAGRAMME...

Sie gelangen zum Fenster der Abbildung 17.5.

Abb. 17.5: Menü ANALYSIEREN/KLASSIFIZIEREN/HIERARCHISCHE CLUSTER..., Schaltfläche DIAGRAMME...

10. Im Fenster der Abbildung 17.5 sorgen Sie für ein Häkchen bei DENDROGRAMM.
11. Bei EISZAPFEN klicken Sie auf KEINE.
12. Klicken Sie auf WEITER. Sie gelangen zum Fenster der Abbildung 17.3 zurück und dann auf OK.
13. Klicken Sie im Fenster der Abbildung 17.3 auf OK.

Hinweis:

Wenn sich die Ausgangsdaten auf sehr unterschiedlichem Niveau bewegen, kann über die Schaltfläche METHODEeine Standardisierung vorgenommen werden, wenn Sie den Listenpfeil anklicken und dann z.B. Z-WERTE anklicken. Es handelt sich hierbei um die sog. *Z-Transformation*, die Sie in Abschnitt 6.4 kennen gelernt haben. Wir haben in diesem Beispiel auf die Standardisierung der Ausgangswerte verzichtet. Es lässt sich zeigen, dass die Clusterbildung dadurch nicht wesentlich beeinflusst wird. Die Abstände zwischen je zwei Merkmalsträgern ändern sich zwar durch die Standardisierung, da sie sich aber alle ändern, bleibt die Relation zwischen den Abständen, zumindest bei der Verwendung Euklidischer Distanzen, praktisch unverändert – und diese Relationen sind es, die schließlich über die Clusterzuordnung entscheiden.

SPSS erzeugt jetzt die Ergebnisse, die im folgenden Abschnitt besprochen werden.

17.3 Ergebnisse

SPSS gibt eine Reihe von Informationen aus, von denen hier nur das sog. Dendrogramm (Baumdiagramm) interessiert. Es stellt sich folgendermaßen dar:

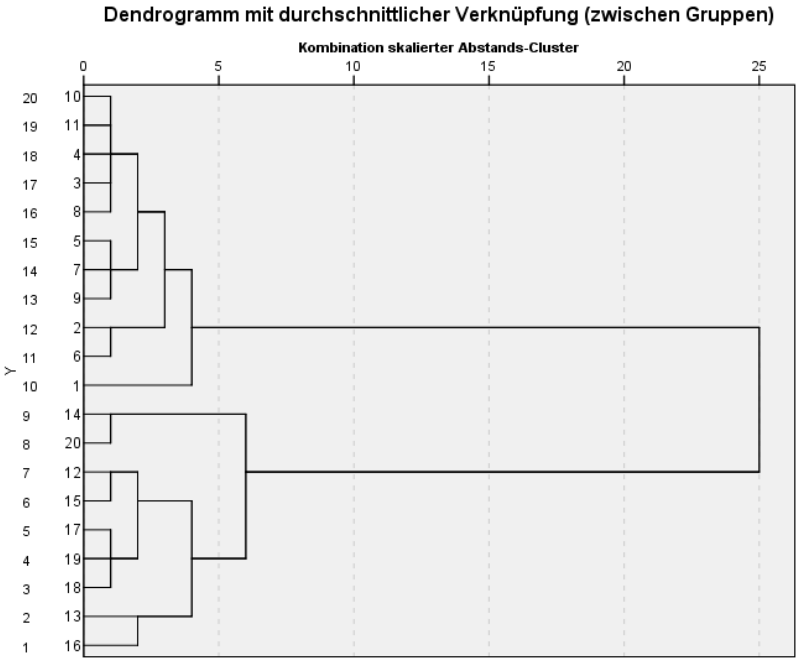


Abb. 17.6: Dendrogramm

In Abbildung 17.6 erkennen Sie, dass SPSS zwei Hauptcluster bildet. Dem ersten gehören die Fälle 12 bis 20 an, dem zweiten die übrigen der 20 Fälle. Mit Blick auf die Ausgangsdaten können Sie nun feststellen, dass im ersten Cluster nur Männer sind, im zweiten Cluster Frauen, aber auch ein Mann (Fall 11).

Die SPSS-Clusteranalyse hat also die beiden Hauptcluster, so wie es auch erwartet werden konnte, nach Maßgabe der dritten Variablen „Geschlecht“ gebildet.

Besonders wichtig ist aber jetzt der folgende Umstand: Selbst wenn Sie die dritte Variable („Geschlecht“) nicht erhoben hätten (sie ging ja auch in die entsprechenden Berechnungen nicht mit ein), hätte sich selbstverständlich diese Clusterbildung ergeben, und Sie würden jetzt vor der Notwendigkeit der Interpretation dieses Befundes stehen. Warum hat SPSS in dieser Weise geclustert?

Um diese Frage zu beantworten, ist es erforderlich, sich den Ausgangsdatenbestand anzuschauen, um zu überprüfen, was – ausgehend von den erhobenen Untersuchungsvariablen und ihren Ausprägungen pro Fall – die Fälle des ersten von denen des zweiten Clusters unterscheidet. Der Blick auf die Ausgangstabelle zeigt: Fälle im ersten Cluster sind eher große und schwere Personen, Fälle im zweiten Cluster sind im Vergleich dazu eher kleiner und leichter. Daran schließt sich die zweite Frage an: Welche (nicht erhobenen) Variablen könnten dafür verantwortlich sein, dass sich die Fälle so elegant in zwei Gruppen einteilen lassen, in kleine, leichte Personen einerseits und in große, schwere Personen andererseits? Antwort: Es könnte das Geschlecht sein, das hier für diese Einteilung sorgt.

Diese Interpretation bietet sich an, selbst wenn zu konstatieren ist, dass der Clusteralgorithmus nicht jeden Fall zweifelsfrei zuordnen kann. Sie wissen ja (aber nur bei diesem kleinen Demonstrationsbeispiel), dass sich im „Frauen-Cluster“ auch ein Mann versteckt. SPSS ordnet diesen (irrtümlich) der Gruppe der Frauen zu. Mit Blick auf die Ausgangsdaten ist dies verständlich: Beim Fall Nr. 11 handelt es sich um einen ausgesprochen

kleinen und leichten Mann (er unterscheidet sich ja kaum vom Fall 10, und das war definitiv eine Frau). Kein Wunder also, dass SPSS diesen Hänfling zu den Frauen sortiert.

Sie erkennen übrigens des Weiteren in Abbildung 17.6, dass SPSS auch „Unter-Cluster" bildet. Zum Beispiel wird das Hauptcluster 1 (Männer) in zwei Untercluster unterteilt, wobei in dem ersten sich die Fälle 14 und 20 befinden, im zweiten die Fälle 12, 13 und 15 bis 20. Mit Blick auf die Daten erkennen Sie, dass die Fälle 14 und 20 besonders große und besonders schwere Männer sind.

Es leuchtet unmittelbar ein, dass man auch für andere Untercluster auf entsprechende Weise zu Erklärungen dafür gelangen kann, warum SPSS so geclustert hat, wie es das Dendrogramm der Abbildung 17.6 zeigt.

Derartige Interpretationsversuche werden umso anstrengender, je tiefer man in die Cluster-Hierarchie einsteigt und je kleiner die Fallzahlen in den jeweiligen Unterclustern werden. Deshalb nutzt man das Dendrogramm, um zu entscheiden, wie viele Cluster überhaupt betrachtet werden sollen. Man kann generell sagen, dass die Interpretation von Teilclustern umso uninteressanter und unergiebiger wird, je weiter links man sich im Dendrogramm befindet. Es interessieren die Hauptäste des Baums, nicht die vielen kleinen Zweiglein.

17.4 Clusterzentrenanalyse

Bei großen Datenbeständen wird die hierarchische Clusteranalyse zu rechenaufwendig. Selbst Computer sind kaum in der Lage, mit den 499.500 Distanzen, die sich bei z.B. n=1000 Fällen ergeben, in angemessener Zeit umzugehen („per Hand" sind solche Berechnungen sowieso nur theoretisch möglich, aber praktisch wegen des erforderlichen Zeitaufwandes nicht durchführbar). SPSS bietet für diesen Fall die Clusterzentrenanalyse an. Hier werden keine paarweisen Vergleiche durchgeführt, sondern die Vorgehensweise ist folgende:

Zunächst wird vom Benutzer festgelegt, wie viele Cluster gebildet werden sollen. Dies bedeutet, dass vor dem Start des Programms Überlegungen hinsichtlich der unbekannten, die Clusterbildung beeinflussenden Variablen und der Zahl ihrer Ausprägungen angestellt werden müssen. Zweckmäßigerweise führt man deshalb zunächst mit einer kleinen Zufallsstichprobe aus dem größeren Datenbestand eine hierarchische Clusteranalyse durch, um anhand des Dendrogramms eine Vorstellung darüber zu erlangen, wie viele Hauptcluster vermutlich erwartet werden können.

Eine weitere Beschleunigung des Verfahrens wird dadurch erzielt, dass die Zentren der Cluster vorgegeben werden.

> **Hinweis:**
>
> Ein Clusterzentrum wird im obigen Beispiel gegeben durch die durchschnittliche Größe und das durchschnittliche Gewicht der Personen im jeweiligen Cluster.

Sie können diese Zentren selbst angeben oder durch SPSS aus dem Ausgangsdatenbestand ausrechnen lassen. Es handelt sich in beiden Fällen um vorläufige Zentren. SPSS nimmt dann eine Einteilung der Fälle gemäß der so vorgegebenen Clusterzentren vor, wobei diese sich allerdings verändern.

Dann werden die Fälle erneut, nun anhand der veränderten Zentren, zugeordnet. Dies wird dann in der Regel zu leichten Veränderungen der Zuordnungen führen. Dadurch verändern sich die Zentren erneut, und der Zuordnungsprozess wird erneut wiederholt. Dies geschieht dann so lange, bis eine vom Anwender vorzugebende Höchstzahl von Zuordnungsvor-

gängen (Anzahl der Iterationen) erreicht ist, oder bis die Unterschiede der sich verändernden Clusterzentren zwischen den einzelnen Zuordnungsdurchgängen eine vorzugebende Untergrenze unterschreiten (Konvergenzkriterium).

Obwohl das Demonstrationsbeispiel, das in den vorangegangenen Abschnitten verwendet wurde, keinen großen Datenbestand darstellt, soll es auch hier verwendet werden, um diesen Algorithmus zu erproben, zudem auf diese Weise Unterschiede in den Ergebnissen der beiden Verfahren deutlich werden.

Um die Clusterzentrenanalyse mit SPSS durchzuführen, ist folgendermaßen vorzugehen:

1. Arbeiten Sie mit den obigen Ausgangsdaten.
2. Wählen Sie ANALYSIEREN/KLASSIFIZIEREN/CLUSTERZENTRENANALYSE...

Sie gelangen zum Fenster der Abbildung 17.7.

Abb. 17.7: Menü ANALYSIEREN/KLASSIFIZIEREN/CLUSTERZENTRENANALYSE...

3. Im Fenster der Abbildung 17.7 übertragen Sie die Variablen „cm“ und „kg“ in den Bereich VARIABLEN:.
4. Geben Sie bei ANZAHL CLUSTER den Wert 2 ein, wenn er dort nicht schon erscheint.
5. Klicken Sie auf die Schaltfläche OPTIONEN... Sie gelangen zum Fenster der Abbildung 17.8.
6. Im Fenster der Abbildung 17.8 sorgen Sie für Häkchen bei ANFÄNGLICHE CLUSTERZENTREN und bei CLUSTER-INFORMATIONEN FÜR JEDEN FALL.
7. Klicken Sie WEITER an. Sie gelangen zurück zum Fenster der Abbildung 17.7.
8. Klicken Sie OK an.

Abb. 17.8: Menü ANALYSIEREN/KLASSIFIZIEREN/CLUSTERZENTRENANALYSE..., Schaltfläche OPTIONEN

SPSS erzeugt jetzt die Ergebnisse der Clusterzentrenanalyse, von denen die wichtigsten vorgestellt werden sollen.

Cluster-Zugehörigkeit

Fallnummer	Cluster	Distanz
1	1	8,093
2	1	4,183
3	1	2,771
4	1	1,918
5	1	5,628
6	1	7,264
7	1	6,123
8	1	2,245
9	1	3,404
10	1	2,899
11	1	3,636
12	2	3,836
13	2	9,924
14	2	11,399
15	2	5,265
16	2	8,316
17	2	,401
18	2	4,824
19	2	2,320
20	2	6,536

Clusterzentren der endgültigen Lösung

	Cluster	
	1	2
cm	165	183
kg	63	81

Abb. 17.9: Ergebnisse der Clusterzentrenanalyse

Zunächst wird unter der Überschrift CLUSTER-ZUGEHÖRIGKEIT eine Liste ausgegeben, der Sie entnehmen können, welcher Fall zu welchem Cluster zugeordnet wurde.

Sie erkennen, dass die Fälle 1 bis 11, wie schon bei der hierarchischen Clusterung in Abschnitt 17.3, dem ersten Cluster, die übrigen Fälle dem zweiten Cluster zugeordnet wurden.

Danach zeigt SPSS die Koordinaten der endgültigen Clustermittelpunkte unter der Überschrift CLUSTERZENTREN DER ENDGÜLTIGEN LÖSUNG.

Wir erkennen, dass im ersten Cluster kleine, leichte, im zweiten Cluster große, schwere Personen versammelt sind. Zudem wird eine Übersicht über die Anzahl der Fälle pro Cluster ausgegeben:

Anzahl der Fälle in jedem Cluster

Cluster	1	11,000
	2	9,000
Gültig		20,000
Fehlend		,000

Abb. 17.10: Anzahl der Fälle in den Clustern

Nun zur Interpretation: Welche Größe(n) könnte dafür verantwortlich sein, dass sich die Punkte im Achsenkreuz (hier die beobachteten Wertepaare „Größe“ und „Gewicht“) so anordnen, dass SPSS zu der vorgestellten Clusterung gelangt?

Wie schon im Abschnitt zuvor ist nun der Blick auf die Ausgangsdaten erforderlich, um festzustellen, was die Fälle auszeichnet, die zum Cluster 1 gehören, und was im Unterschied dazu die Fälle auszeichnet, die zum Cluster 2 gehören. Wieder haben Sie erkannt, dass im Cluster 1 kleine, leichte Menschen, im Cluster 2 große, schwere Menschen sind, weshalb Sie jetzt wieder auf die Idee kommen dürften, dass das Geschlecht die für die vorgenommene Clusterzuordnung maßgebliche Variable sein könnte.

Um bei einem Datenbestand mit mehr als nur zwei oder drei Variablen zu einer brauchbaren Interpretation zu gelangen, empfiehlt es sich, die Häufigkeitsverteilungen für alle nicht in die Clusterung eingegangenen Variablen auszugeben – im Beispiel ist dies nur die Variable „sex“ – und zwar für die einzelnen Cluster getrennt. Dies ist dann sehr leicht möglich, wenn Sie sich die Clusternummern mit speichern lassen.

Dazu müssen Sie im Fenster der Abbildung 17.7 die Schaltfläche SPEICHERN… anklicken und dort bei CLUSTER-ZUGEHÖRIGKEIT; danach WEITER und OK.

Im Datenbestand wird dann eine Variable (Spalte) mit dem Namen „QCL_1“ angefügt. Wählen Sie dann Menü DATEN/FÄLLE AUSWÄHLEN… Sie gelangen damit in das Fenster der Abbildung 17.11:

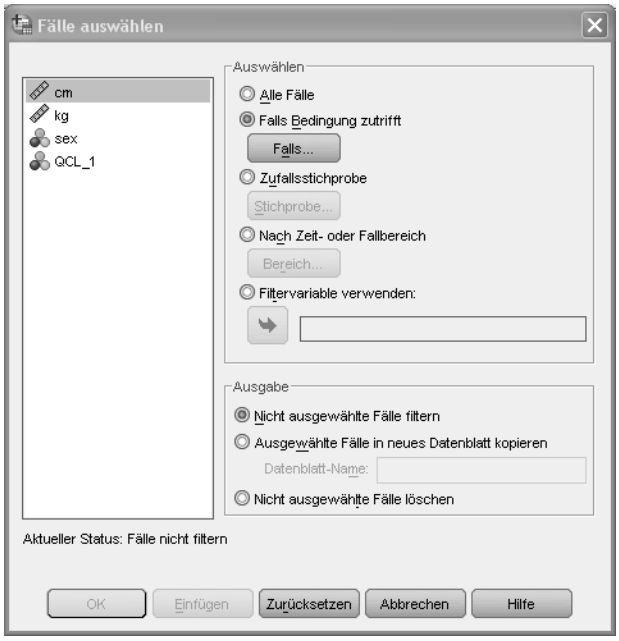


Abb. 17.11: Menü DATEN/FÄLLE AUSWÄHLEN…

Klicken Sie hier an bei FALLS BEDINGUNG ZUTRIFFT und dann auf die Schaltfläche FALLS…

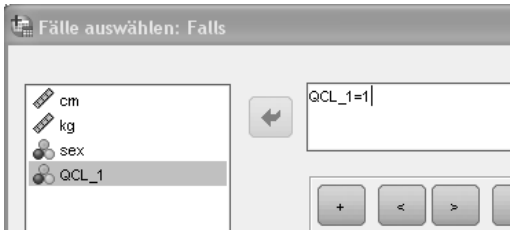


Abb. 17.12: Formulierung einer Auswahlbedingung (Ausschnitt)

Hier übertragen Sie die Variable QCL_1 nach rechts gefolgt von „=1“; dann WEITER und OK. Damit werden die Fälle ausgewählt, die dem Cluster 1 angehören. Folgende Auswertungsprozeduren beziehen sich dann ausschließlich auf diese Fälle.

Erzeugen Sie jetzt Häufigkeitsverteilungen für die anderen Variablen (hier nur für die Variable „sex“), ergibt sich Abbildung 17.13:

sex

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	1	9,1	9,1	9,1
	1	10	90,9	90,9	100,0
	Gesamt	11	100,0	100,0	

Abb. 17.13: Geschlechtsverteilung im Cluster 1

Wiederholen Sie diese Schritte, wählen nun aber als Auswahlbedingung QCL_1=2, erhalten Sie die Geschlechtsverteilung im Cluster 2:

sex

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	9	100,0	100,0	100,0

Abb. 17.14: Geschlechtsverteilung im Cluster 2

Sie erkennen, was schon bekannt ist: Im Cluster 1 befinden sich Frauen (und ein Mann), im Cluster 2 befinden sich nur Männer.

Hinweis:

Wenn Sie mit der Prozedur DATEN/FÄLLE AUSWÄHLEN gearbeitet haben und dann die Arbeit mit SPSS mit Prozeduren fortsetzen wollen, die sich wieder auf alle Fälle beziehen sollen, müssen Sie zunächst im Fenster der Abbildung 17.11 bei ALLE FÄLLE anklicken.

17.5 Ergänzungen

Zum Verfahren der Clusteranalyse sollen in diesem Abschnitt ergänzende Anmerkungen zu den Voraussetzungen präsentiert werden.

In der Standardliteratur zur Clusteranalyse finden Sie häufig den Hinweis, dass die Variablen, die in einer Clusteranalyse Verwendung finden sollen, unkorreliert sein sollten. Dies dürfte in der Praxis aber eher die Ausnahme sein. Bei dem oben verwendeten Beispiel ging es um die Variablen „Körpergröße“ und „Körpergewicht“, und natürlich korrelieren diese Variablen miteinander. Somit wird gegen die genannte Voraussetzung der Unabhängigkeit der Variablen voneinander verstoßen. Diese Voraussetzung erklärt sich aus dem Umstand, dass sich in den Distanzen zwischen den Wertepaaren korrelierender Variablen Verzerrungen niederschlagen, die um so gravierender sind, mithin auch die Zuordnungen der Clusterung beeinflussen, je enger die Zusammenhänge zwischen einzelnen Variablen sind. Hier bietet sich nun aber ein überzeugender Ausweg an:

Stellen Sie sich vor, eine große Zahl v von Untersuchungsvariablen sollen in einer Clusteranalyse verarbeitet werden. Paarweise werden diese Untersuchungsvariablen mehr oder weniger stark miteinander korrelieren, wobei nicht ausgeschlossen werden kann, dass auch einige voneinander unabhängig sein können. Führt man nun für diese v Variablen zunächst eine Faktorenanalyse durch (siehe Kapitel 16), so erzeugt man einen Set von k Faktoren (Komponenten), deren Werte für die einzelnen Merkmalsträger gespeichert werden können Diese Faktoren sind nichts anderes als „künstliche“ Variablen, und diese sind, wie schon angemerkt wurde, voneinander unabhängig, d.h. sie korrelieren nicht miteinander. Die Faktoren eignen sich mithin in hervorragender Weise dazu, der Voraussetzung für die Clusteranalyse zu entsprechen und somit als Ausgangsvariablen für die Clusteranalyse zu dienen.

Weiterhin wird vorausgesetzt, dass die Variablen ungefähr gleiche Wertebereiche aufweisen sollten, denn auch hier gilt, dass Variablen, die sich in hohen Wertebereichen bewegen, einen zu großen Einfluss gewinnen gegenüber solchen, die sich in niedrigen

Wertebereichen bewegen. Um dieses Problem zu lösen, könnte man statt mit den ursprünglichen Variablen mit standardisierten Variablen arbeiten.

Schließlich ist anzumerken, dass die Clusteranalyse nur für metrische Daten sinnvoll eingesetzt werden kann. Wenn Sie nichtmetrische Variablen haben, können Sie diese durch Umkodierungen in 0/1-kodierte dichotome Variablen umwandeln – und diese können, das wurde an anderer Stelle ja schon erläutert, wie metrische Variablen behandelt werden.

18 Logit-Analyse

Bei den Ausführungen dieses Kapitels stützen wir uns mit freundlicher Genehmigung des Verfassers weitgehend auf den Übersichtsartikel von *M. Tiede*: Statistische Logit-Analyse, eine Orientierungshilfe für die Verwendung des binären Logit-Modells, Diskussionspapiere aus der Fakultät für Sozialwissenschaft der Ruhr-Universität Bochum, Heft 95-3, 1995.

18.1 Logit-Modell

Die Logit-Analyse zählt in amerikanischen empirisch orientierten sozialwissenschaftlichen Zeitschriftenbeiträgen derzeit zu den häufig verwendeten Analyseformen. Aber auch im deutschsprachigen Raum gewinnt dieses Verfahren zunehmend an Bedeutung. Da das Verfahren aber trotzdem noch nicht allzu sehr verbreitet ist, zunächst einige Anmerkungen zur Methodik:

Ein Logit-Modell ist ein Regressionsmodell, allerdings nicht für metrische Variablen (siehe dazu Kapitel 9), sondern für eine zu erklärende Variable, die Nominalskalenqualität aufweist, und die nur zwei Ausprägungen aufweist (dichotome Variable). Allerdings ist auch eine Ausweitung auf polytome Variablen möglich (mehr als zwei Ausprägungen), was zur sog. Multinomialen logistischen Regression führt, auf die hier aber nicht eingegangen werden soll.

Bei der logistischen Regression wird die Beziehung zwischen der abhängigen, zu erklärenden Variablen und den beeinflussenden Variablen als nichtlineare Beziehung formuliert.

Wir hatten schon an anderer Stelle darauf aufmerksam gemacht, dass bei einer dichotomen Variablen, deren beide Ausprägungen mit 0 und 1 kodiert werden, das Instrumentarium der Regressions- und Korrelationsrechnung, das ja eigentlich auf das Vorliegen metrischer Variablen abstellt, verwendet werden kann (siehe z.B. bezüglich der Korrelationsrechnung Abschnitt 10.6). Zur Verdeutlichung dieses Sachverhaltes sei das Beispiel vorgestellt, das auch *Tiede* in dem o.a. Beitrag verwendet:

Die abhängige, zu erklärende, dichotome Variable Y möge sich auf die Frage beziehen, ob zufällig ausgewählte Versuchspersonen sich für eine 14-tägige Urlaubsreise nach Bali entschließen, oder nicht. Diese Variable Y hat die beiden Ausprägungen 0 (es wird nicht gereist) und 1 (es wird gereist).

In einem sehr einfachen gedanklichen Modell soll als erklärende Variable X das monatliche Nettoeinkommen (in 1000 Euro) verwendet werden. In einer Zufallsstichprobe ergibt sich der Datenbestand der Abbildung 18.1.

Berechnet man aus diesen Daten mit der Methode der kleinsten Quadrate eine lineare Regressionsfunktion, erhält man die folgenden Parameter:

Ordinatenabschnitt a	= -0,724
Steigung b	= 0,226
Determinationskoeffizient r^2	= 0,34

Als Diagramm ergibt sich das Bild in Abbildung 18.2:

	eink1000	reise
1	3,00	0
2	2,50	0
3	5,00	1
4	4,50	0
5	3,50	0
6	7,50	1
7	6,00	1
8	6,10	0
9	5,50	0
10	4,60	0
11	3,70	0
12	3,00	0
13	3,30	0
14	4,20	0
15	3,90	0
16	5,20	1
17	5,00	1
18	4,80	1
19	4,60	0
20	4,70	0

Abb. 18.1: Ausgangsdaten

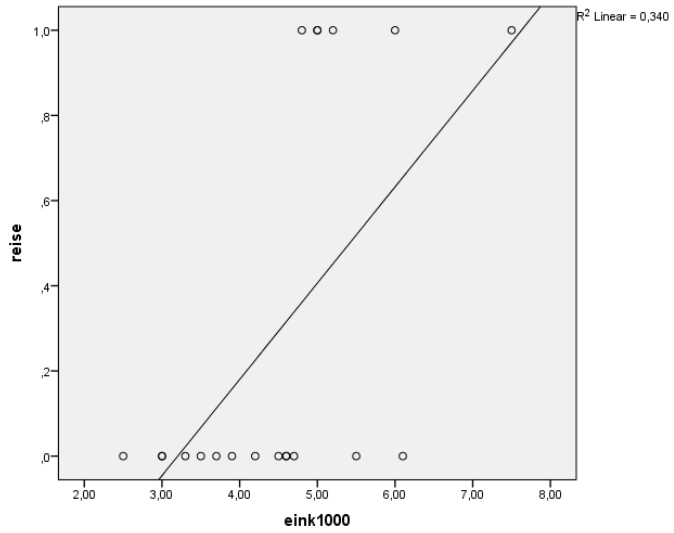


Abb. 18.2: Lineare Regressionsrechnung

Der Determinationskoeffizient besagt, dass 34% der Streuung der abhängigen Variablen Y durch die Variation der Variablen X (Einkommen) statistisch erklärt wird (siehe auch Abschnitt 10.3). Dies ist kein sehr befriedigendes Ergebnis, weil 34% der Variablen Y, also der Wert 0,34 keinen Wert der dichotomen Variablen repräsentiert. Zudem erkennen Sie in der Abbildung 18.2, und auch das ist recht unbefriedigend, dass bei einem Einkommen, das größer ist als ca. 7500 Euro, die Wahrscheinlichkeit, dass eine entsprechende, zufällig ausgewählte Person die fragliche Reise bucht, größer als 1 wird. Wahrscheinlichkeiten größer als 1 sind aber unsinnig, genauso, wie Wahrscheinlichkeiten kleiner als 0 unsinnig

sind. Eine Person, die weniger als ca. 3000 Euro bezieht, bucht mit einer (sinnlosen) negativen Wahrscheinlichkeit eine Reise nach Bali.

Deshalb sollte das lineare Modell modifiziert werden, so dass es

1. nur solche Werte für Y prognostiziert, die zwischen 0 und 1 liegen,
2. die Wahrscheinlichkeitsveränderungen nicht linear abbildet.

Zu diesem zweiten Punkt eine plausible Erläuterung, ausgehend von dem Reisebeispiel: Es darf unterstellt werden, dass bei einer Erhöhung des Einkommens von z.B. 9000 Euro auf 10000 Euro die Neigung, eine Bali-Reise zu buchen, weniger zunimmt (da sie sowieso schon recht hoch sein dürfte), als bei einem Einkommensanstieg von z.B. 5000 Euro auf 6000 Euro. Der im linearen Modell stets konstante Zuwachs für die Wahrscheinlichkeit, die Bali-Reise zu buchen, ist nicht sehr plausibel (gemäß des Regressionskoeffizienten b = 0,226 steigt die Wahrscheinlichkeit immer um diesen konstanten Betrag, wenn sich das Einkommen um 1000 Euro erhöht).

Auf diese beiden Einwände reagiert das Logit-Modell, indem die Zielgröße π_i wie folgt transformiert wird:

$$\Gamma(p_i) = \ln \frac{p_i}{1 - p_i}$$

Die Funktion Γ wird in diesem Zusammenhang als *Link-Funktion* bezeichnet (link = Verbindung); ln ist der natürliche Logarithmus. Wenn beispielsweise gilt, dass $e^y = b$, so ist $y = \ln b$ (e ist die Euler'sche Zahl, die sog. Wachstumskonstante, deren numerischer Wert bei e=2,718... liegt).

Diese Transformation wird als *Logit* bezeichnet. Wird p_i eingesetzt, so wird $\Gamma(p_i)$ als *Logitwert* für p_i bezeichnet.

Die bivariate lineare Regressionsfunktion, die jetzt nicht die zu erklärende Ausgangsvariable Y, sondern $\Gamma(p_i)$ verwendet, sieht so aus:

$$\Gamma(p_i) = \ln \frac{p_i}{1 - p_i} = a + b*x_i$$

Anmerkung: Der Quotient, der hinter dem natürlichen Logarithmus ln steht, wird als *odds* bezeichnet.

Im Englischen bedeutet „odds" beispielsweise das Verhältnis zwischen Einsatz und Gewinn bei Wetten beim Pferderennen. Im Zusammenhang mit den hier angestellten Überlegungen könnte mit diesem Begriff eine gewisse Dominanz zum Ausdruck gebracht werden: Falls der Odds den Wert z hat, beträgt die Wahrscheinlichkeit für die Kategorie 1 der Variablen Y das z-fache der Wahrscheinlichkeit für die Kategorie 0.

Entspricht nun ein solches Modell den oben formulierten beiden Einwänden gegen das einfache lineare Modell?

Zum ersten Problem des Wertebereichs ist folgendes anzumerken: Wie der mathematisch vorgebildete Leser weiß, variiert $\Gamma(p_i)$ zwischen -unendlich und +unendlich. Dies entspricht einem Wertebereich für p_i, der von nahe Null bis nahe Eins reicht. Damit ist dem ersten Einwand Rechnung getragen.

Dem zweiten Einwand wird dadurch Rechnung getragen, dass eine (logistische) Regressionsfunktion entsteht, wie in Abschnitt 18.5 gezeigt wird, die zunächst langsam, dann steiler und dann wieder schwächer ansteigt.

18.2 Logit-Koeffizienten

Ausgehend von der Beziehung

$$\Gamma(p_i) = \ln \frac{p_i}{1-p_i} = a + b*x_i$$

sollen nun die Koeffizienten a und b näher betrachtet werden.

Der Regressionskoeffizient a gibt an, welcher Logitwert zu erwarten ist, falls die erklärende Variable den Wert X=0 aufweist.

Der Regressionskoeffizient b gibt an, welche Änderung für den Logitwert zu erwarten ist, wenn sich X um eine Einheit ändert (Wirkungsstärke von X). Beachten Sie aber: Die Änderung von p_i, die auf einer Änderung von X beruht, hängt nicht linear mit b zusammen, sondern zusätzlich von der Höhe des jeweiligen X-Wertes.

Hat man mehrere beeinflussende Variablen (multivariates Modell), so ist ein Vergleich der Wirkungsstärken der unterschiedlichen Variablen X_j nur möglich auf der Basis normierter (standardisierter) Logitkoeffizienten. Die Standardisierung

$$b^* = b/\sqrt{VAR(X)} = b*\text{Standardabweichung}$$

führt zu einem dimensionslosen Koeffizienten. Diese Größe b^* (normierter Logit-Koeffizient) gibt die Änderung des Logitwertes an, falls der Wert der beeinflussenden Variablen X um eine Standardabweichung erhöht wird.

18.3 Maximum-Likelihood-Schätzung

Ausgehend von einem Stichprobendatenbestand können die Koeffizienten des Logit-Modells geschätzt werden. Es stehen dafür Kleinst-Quadrate-Schätzungen oder *Maximum-Likelihood-Schätzungen* zur Verfügung, die auch von den zuständigen SPSS-Prozeduren verwendet werden. Bei dieser Methode werden die Regressionskoeffizienten des Logit-Modells in der Weise bestimmt, dass die in der Stichprobe beobachteten Logitwerte die größte Chance hatten, realisiert zu werden.

Für die Regressionskoeffizienten des Logit-Modells kann die Likelihood-Funktion L wie folgt ermittelt werden:

Bekannt sind aus der Stichprobe die relativen Häufigkeiten p_i bzw. die absoluten Häufigkeiten $n*p_i$. Diese absoluten Häufigkeiten sind Werte für die mit 1 kodierte Kategorie der abhängigen Untersuchungsvariablen Y, die binomisch verteilt ist mit den Parametern n und π_i. Somit kann man die Likelihood-Funktion herleiten, die angibt, wie wahrscheinlich ein beobachteter Stichprobenbefund in Abhängigkeit von den zu schätzenden π_i-Werten ist. Wird mit Hilfe der Differentialrechnung das Maximum dieser Funktion bestimmt, erhält man Bestimmungsgleichungen für die Regressionskoeffizienten des Logit-Modells. Auf Einzelheiten dieser Berechnungen, die SPSS automatisch durchführt, soll hier nicht eingegangen werden.

18.4 Modellgüte

Wenn die Regressionskoeffizienten des Logit-Modells berechnet worden sind, interessiert eine Aussage zur Güte des Modells. Wie gut konnte das Logit-Modell an die empirischen Daten angepasst werden? Um diese Frage zu beantworten, müssen die Unterschiede zwischen den in der Stichprobe beobachteten Logitwerten (bzw. den Anteilswerten) und den entsprechenden Werten, die das geschätzte Modell ergibt, untersucht werden. Es bietet sich in diesem Zusammenhang ein Anpassungstest an, der die beobachteten den geschätzten Anteilswerten gegenüberstellt.

Eine wichtige Maßzahl in diesem Zusammenhang, die in der zuständigen Fachliteratur häufig auftaucht, ist das sog. Pseudo-R^2. Es kann damit die Hypothese getestet werden, die Zielvariable Y („Buchung einer Bali-Reise“, um an das Ausgangsbeispiel zu erinnern) sei unabhängig von der Variablen X („Einkommen“), oder unabhängig von den Variablen X_j im multiplen Modell.

Als Prüfvariable wird $G = -2*\ln(L_0/L_1) = -2(\ln L_0 - \ln L_1)$

verwendet. Dabei bedeuten:

L_0 = Maximum der Likelihood-Funktion ohne beeinflussende Variable

L_1 = Maximum der Likelihood-Funktion mit beeinflussender Variablen (bzw. mit allen beeinflussenden Variablen).

Die Prüfvariable G folgt für große Stichprobenumfänge approximativ einer Chi-Quadrat-Verteilung (Freiheitsgrade = Parameterzahl des vollständigen Modells minus Parameterzahl des um die Zahl der beeinflussenden Variablen reduzierten Modells). Als Pseudo-R^2 wird nun die folgende Größe bezeichnet:

$$\text{Pseudo-}R^2 = 1 - \ln L_1/\ln L_0$$

Diese Größe ist im Wertebereich zwischen 0 und +1 definiert. Der Wert 0 ergibt sich, wenn $L_0 = L_1$. In diesem Fall stellen die Modelle, mit und ohne beeinflussende Variablen, die Stichprobenwerte gleich gut (also schlecht) dar; die beeinflussende(n) Variable(n) tragen nichts zur Erklärung bei. Der Wert 1 ergibt sich, wenn $L_1 = 0$, falls also sicher ist, dass das geschätzte Modell die Stichprobenwerte fehlerfrei darstellen kann. Relativ kleine Werte dieser Maßzahl sprechen also gegen eine gute Modelldarstellung der empirischen Daten.

Ob das Maß signifikant von Null abweicht, kann mit einem Liklihood-Quotienten-Test geprüft werden. Dabei gilt allerdings, dass eine signifikante Abweichung von Null nur besagt, dass das verwendete Logit-Modell besser ist als das (theoretische) Modell ohne beeinflussende Variable(n). Nähere Details sind der oben erwähnten Arbeit von *Tiede* zu entnehmen.

18.5 Beispiel

Nun zu einem Anwendungsbeispiel, bei dem wir einen etwas umfangreicheren Datenbestand verwenden.

Es soll untersucht werden, in welcher Weise die Buchung einer Reise (Variable Y = „reise“, mit den Werten 0 für „Nein“ und 1 für „Ja“) durch das monatliche Nettoeinkommen (Variable X = „eink1000“ = Einkommen in 1000 Euro) zufällig ausgewählter Personen beeinflusst wird. Die Ausgangsdaten liegen gruppiert vor und stellen sich so dar, wie in Abbildung 18.3 gezeigt wird.

	gruppe	eink	reise	n
1	1	2	0	10
2	2	3	2	20
3	3	4	5	20
4	4	6	10	30
5	5	7	10	25
6	6	9	11	22
7	7	10	10	18
8	8	13	12	18
9	9	18	6	10
10	10	30	5	5

Abb. 18.3: Ausgangsdaten

In dieser Abbildung 18.3 erkennen Sie, dass wir Gruppen von Befragten gebildet haben:

In der ersten Gruppe liegt das Durchschnittseinkommen bei 2000 Euro, und von $n_1 = 10$ befragten Personen haben sich $y_2 = 0$ für eine Reise entschieden (Anteil $p_1 = 0{,}00$). In der zweiten Gruppe liegt entsprechend das Durchschnittseinkommen bei 3000 Euro, und von $n_2 = 20$ Befragten haben sich $y_2 = 2$ für die Reise entschieden ($p_2 = 0{,}10$), usw.

Zur Durchführung der Logit-Analyse gehen Sie folgendermaßen vor:

1. Geben Sie die obigen Daten in eine neue SPSS-Tabelle ein.
2. Wählen Sie ANALYSIEREN/REGRESSION/PROBIT...

Sie gelangen zum Fenster der Abbildung 18.4.

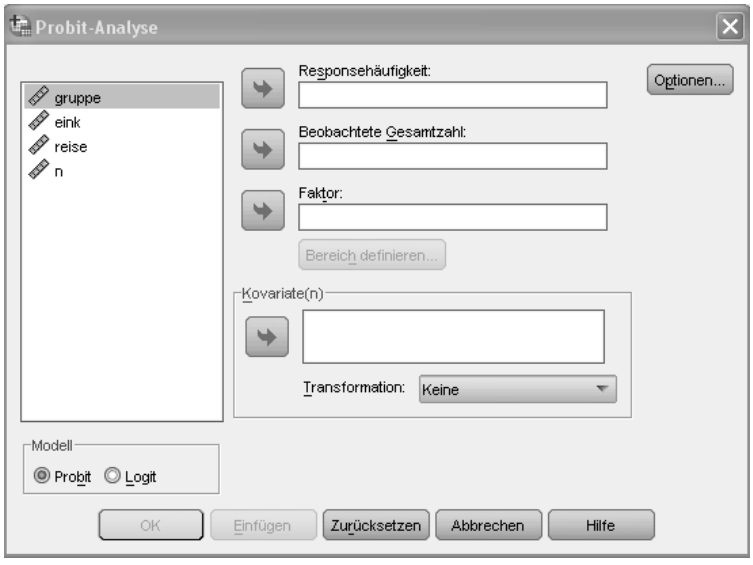


Abb. 18.4: Menü ANALYSIEREN/REGRESSION/PROBIT...

3. Im Fenster der Abbildung 18.4 übertragen Sie die Variable „reise“ in den Bereich RESPONSEHÄUFIGKEIT:.

4. Übertragen Sie die Variable „n“ in den Bereich BEOBACHTETE GESAMZAHL:.
5. Übertragen Sie die Variable „eink“ in den Bereich KOVARIATE(N):.
6. Klicken Sie im Bereich MODELL den Optionsschalter bei LOGIT an.
7. Klicken Sie auf OK.

Die wichtigste Ausgabe von SPSS ist die der Koeffizienten in Abbildung 18.5:

Parameterschätzer

	Parameter	Schätzer	Standardfehler	Z-Wert	Sig.	95%-Konfidenzintervall	
						Untergrenze	Obergrenze
LOGIT[a]	eink	,199	,043	4,581	,000	,114	,284
	Konstante	-2,025	,381	-5,318	,000	-2,405	-1,644

a. LOGIT-Modell: LOG(p/(1-p)) = Konstante + BX

Abb. 18.5: Koeffizienten der logistischen Regression

In dieser Abbildung 18.5 sehen Sie, dass der Regressionskoeffizient (es gibt nur einen einzigen, weil nur eine beeinflussende Variable unterstellt wurde) mit b = 0,199 und der Ordinatenabschnitt mit -2,025 angegeben wird.

Weiterhin erkennen Sie in Abbildung 18.6 eine Überschreitungswahrscheinlichkeit von p = 0,444, was besagt, dass die vom verwendeten Modell geschätzten Anteilswerte (zu erwartende Anteile derjenigen, die die Reise buchen, in den einzelnen Einkommensgruppen) hinreichend gut mit den empirischen Anteilswerten übereinstimmen.

Chi-Quadrat-Tests

		Chi-Quadrat	Freiheitsgrade[b]	Sig.
LOGIT	Anpassungstest nach Pearson	7,897	8	,444[a]

a. Da das Signifikanzniveau größer als ,150 ist, wird bei der Berechnung der Konfidenzgrenzen kein Heterogenitätsfaktor verwendet.

b. Statistiken auf der Grundlage einzelner Fälle unterscheiden sich von Statistiken auf der Grundlage aggregierter Fälle.

Abb. 18.6: Test der Anpassungsgüte

Zur Kontrolle der Anpassungsgüte dienen in erster Linie die Angaben der Abbildung 18.7.

In dieser Abbildung 18.7 werden die beobachteten Werte (Anzahl derjenigen, die die Reise buchen in den einzelnen Einkommensgruppen) unter der Überschrift BEOBACHTETE RESPONSES den erwarteten Werten unter der Überschrift ERWARTETE RESPONSES einander gegenübergestellt. In der letzten Spalte finden Sie unter dem Stichwort WAHRSCHEINLICHKEIT die geschätzten Anteilswerte.

Zellenhäufigkeiten und Residuen

	Anzahl	eink	Anzahl Personen	Beobachtete Responses	Erwartete Responses	Residuum	Wahrscheinlichkeit
LOGIT	1	2,000	10	0	1,643	-1,643	,164
	2	3,000	20	2	3,869	-1,869	,193
	3	4,000	20	5	4,528	,472	,226
	4	6,000	30	10	9,103	,897	,303
	5	7,000	25	10	8,676	1,324	,347
	6	9,000	22	11	9,717	1,283	,442
	7	10,000	18	10	8,841	1,159	,491
	8	13,000	18	12	11,462	,538	,637
	9	18,000	10	6	8,258	-2,258	,826
	10	30,000	5	5	4,905	,095	,981

Abb. 18.7: Beobachtete und erwartete Häufigkeiten

Ergänzend stellen wir in der folgenden Tabelle für Euro-Beträge von 0 Euro bis 30000 Euro (in Tausenderschritten) die zu erwartenden Logitwerte und die zu erwartenden Anteilswerte dar:

Euro in 1000	logit	p
0	-2,025	0,117
1	-1,826	0,139
2	-1,627	0,164
3	-1,428	0,193
4	-1,229	0,226
5	-1,03	0,263
6	-0,831	0,303
7	-0,632	0,347
8	-0,433	0,393
9	-0,234	0,442
10	-0,035	0,491
11	0,164	0,541
12	0,363	0,590
13	0,562	0,637
14	0,761	0,682
15	0,96	0,723
16	1,159	0,761
17	1,358	0,795
18	1,557	0,826
19	1,756	0,853
20	1,955	0,876
21	2,154	0,896
22	2,353	0,913
23	2,552	0,928
24	2,751	0,940
25	2,95	0,950
26	3,149	0,959
27	3,348	0,966
28	3,547	0,972
29	3,746	0,977
30	3,945	0,981

Abb. 18.8: Logit-Werte und zu erwartende Wahrscheinlichkeiten (p)

Stellt man die Wahrscheinlichkeiten unter p grafisch dar, ergibt sich Abbildung 18.9:

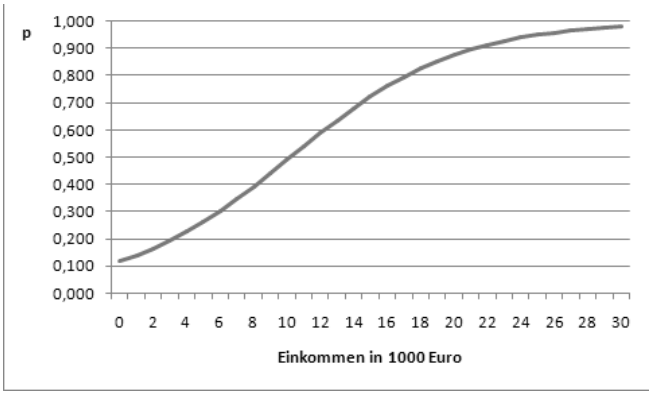


Abb. 18.9: Zu erwartende Reisewahrscheinlichkeiten

Sie erkennen, dass ab einem Einkommen von etwas über 10000 Euro die Wahrscheinlichkeit für Y=1 (Reise wird gebucht) den 50%-Wert übersteigt. Sie erkennen weiterhin, dass der Wertebereich zwischen 0 und 1 eingehalten wird, da sich die Kurve links und rechts asymptotisch diesen Grenzen annähert. Dies entspricht der weiter oben vorgetragenen Überlegung, dass in einem mittleren Einkommensbereich die Wahrscheinlichkeitszuwächse relativ konstant sind, während z.B. bei höheren Einkommen die Wahrscheinlichkeit, eine Bali-Reise zu buchen, zwar weiter ansteigt, aber mit abnehmenden Zuwächsen.

19 Ergänzungen

Das Programm SPSS bietet derartig vielfältige Möglichkeiten, dass nur ein Teil davon hier besprochen werden konnte, um den vorgegebenen Rahmen nicht zu sprengen. Wir haben uns dabei auf die für die Praxis wichtigsten Verfahren beschränkt, wollen aber in diesem Kapitel noch einige Stichworte zu anderen Aspekten des Programms nachtragen, die ab und zu auch eine Rolle spielen dürften.

19.1 Ausgabe

Gedruckt wird über die Menüposition DATEI/DRUCKEN... oder, indem Sie die Schaltfläche mit dem Druckersymbol anklicken. Es öffnet sich das Fenster der Abbildung 19.1:

Abb. 19.1: Menü DATEI/DRUCKEN

In Abbildung 19.1 erkennen Sie, dass Sie auch eine Auswahl Ihrer Informationen drucken können. Diese muss zunächst durch entsprechende Markierung festgelegt werden.

Einstellungen für die Ausgabe am Bildschirm nehmen Sie über das Menü BEARBEITEN/OPTIONEN... vor Was in diesem Fenster der Abbildung 19.2 im Einzelnen in den verschiedenen Registern möglich ist, erkennen Sie an den angegebenen Stichworten, so dass Detailerläuterungen entbehrlich sein dürften.

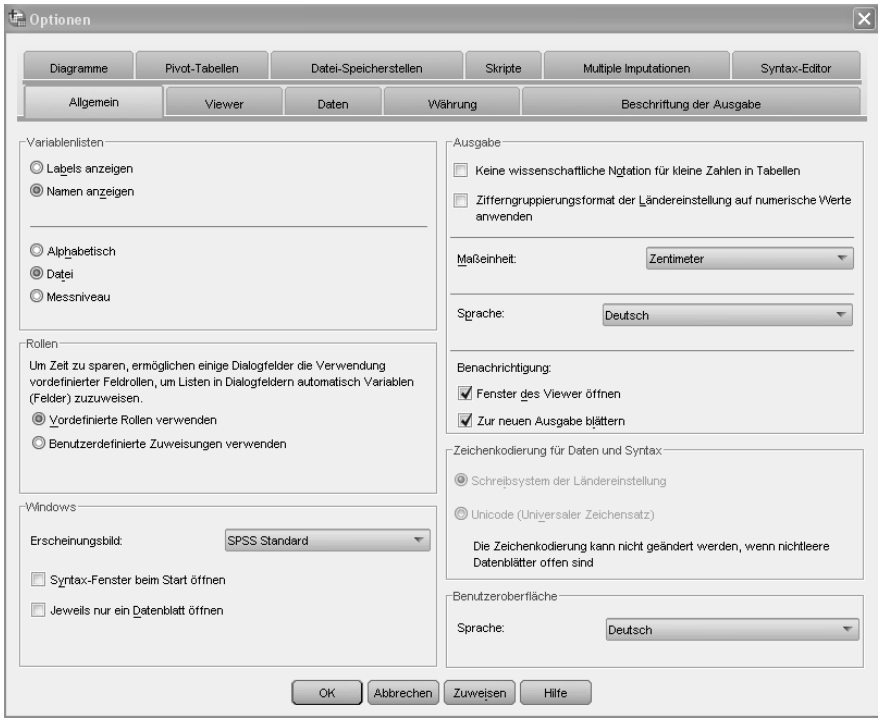


Abb. 19.2: Menü BEARBEITEN/OPTIONEN...

19.2 Datentransformationen

Über das Transformieren von Daten wurde schon im Zusammenhang mit Umkodierungen gesprochen (Menü TRANSFORMIEREN/UMKODIEREN). Weitere interessante Möglichkeiten, die dieses Menü bietet, sind die folgenden:

TRANSFORMIEREN/VARIABLE BERECHNEN...

Dieser Befehl führt zum Dialogfenster der Abbildung 19.3.

Sie sehen in der Abbildung 19.3, dass Sie eine neue Variable aus den vorhandenen Variablen über einen mathematischen Algorithmus erzeugen können.

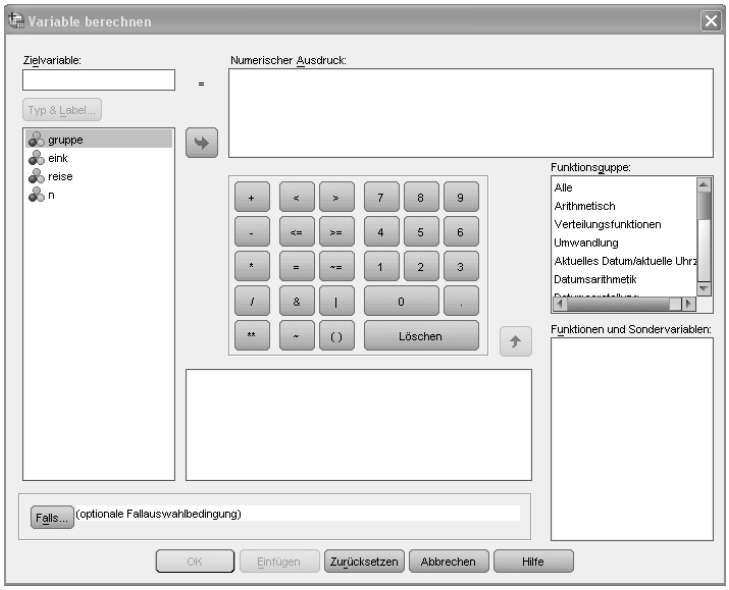


Abb. 19.3: Menü TRANSFORMIEREN/VARIABLE BERECHNEN…

TRANSFORMIEREN/WERTE IN FÄLLEN ZÄHLEN...

Dieser Befehl führt zum Dialogfenster der Abbildung 19.4.

Abb. 19.4: Menü TRANSFORMIEREN/WERTE IN FÄLLEN ZÄHLEN…

Hier wird Ihnen die Möglichkeit geboten, das Auftreten einzelner Werte bei auszuwählenden Variablen auszuzählen. Darüber wurde schon an anderer Stelle gesprochen.

19.3 Mehrfachantworten

Mit dem Menü ANALYSIEREN/MEHRFACHANTWORTEN wird ein Tatbestand angesprochen, der in der praktischen statistischen Arbeit nicht ganz unwichtig ist. Deshalb soll hierzu ein kleines Illustrationsbeispiel vorgeführt werden.

Stellen Sie sich vor, in einer Befragung wäre u. a. danach gefragt worden, welche Zeitschriften regelmäßig gelesen werden. In dem Fragebogen tauchen verschiedene Zeitschriften auf, und der Befragte ist gebeten, anzukreuzen:

Der Spiegel	O
Emma	O
FAZ	O
Focus	O
Die Zeit	O

usw.

Solche Angaben werden in der Regel so kodiert, dass für jede Zeitschrift eine Variable definiert wird, die jeweils nur den Wert 0 (kein Kreuz) oder 1 (Kreuz) aufweist. Der Datenbestand sieht dann so aus, wie es Abbildung 19.6 zeigt.

	spiegel	emma	faz	focus	zeit
1	0	1	1	1	0
2	1	0	0	0	1
3	1	0	0	0	0
4	0	0	1	1	0
5	0	1	0	1	0
6	0	0	0	0	1
7	1	0	0	0	0
8	0	1	0	1	0
9	1	1	1	1	0
10	1	0	0	0	0
11					

Abb. 19.5: Daten mit Mehrfachantworten

Bei der univariaten Datenauswertung kann dann festgestellt werden, wie viele befragte Personen den Spiegel lesen, wie viele Emma lesen usw. Zusätzlich kann man aber auch eine Mehrfachantworten-Analyse vornehmen. Dazu gehen Sie wie folgt vor:

1. Geben Sie die obigen Daten in eine neue SPSS-Tabelle ein.
2. Wählen Sie ANALYSIEREN/MEHRFACHANTWORTEN/SETS DEFINIEREN...
3. Im Fenster der Abbildung 19.7 übertragen Sie alle Variablen in den Bereich VARIABLEN IM SET:.
4. Im Bereich VARIABLENKODIERT ALS klicken Sie auf den Optionsschalter bei DICHOTOMIEN, wenn dies noch erforderlich ist.
5. Bei GEZÄHLTER WERT geben Sie die 1 ein.

6. Im Bereich NAME: geben Sie z.B. Anzahl ein (SPSS ändert diesen Namen in $Anzahl im Bereich MEHRFACHANTWORTEN-SETS☺ .
7. Klicken Sie auf die Schaltfläche HINZUFÜGEN.
8. Klicken Sie auf die Schaltfläche SCHLIEßEN.

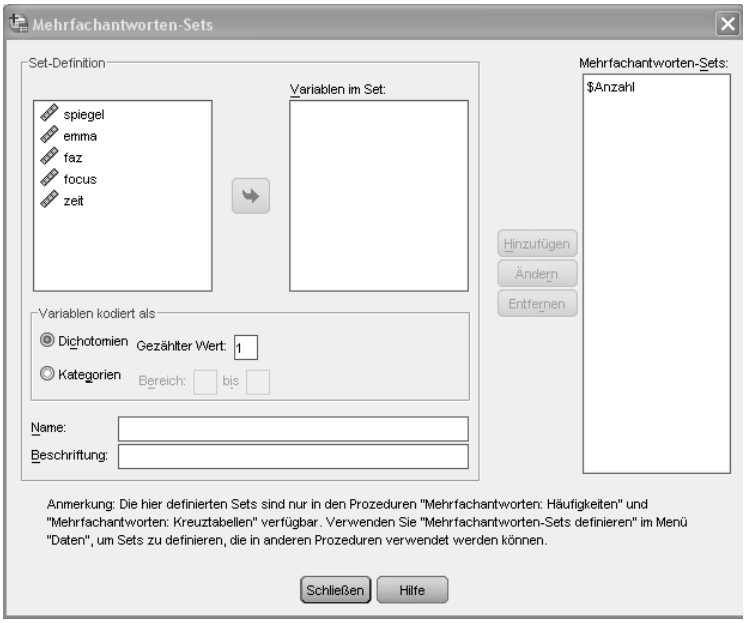


Abb. 19.6: Menü ANALYSIEREN/MEHRFACHANTWORTEN/VARIABLENSETS DEFINIEREN...

Wählen Sie dann erneut Menü ANALYSIEREN/MEHRFACHANTWORTEN und dort jetzt HÄUFIGKEITEN...

Abb. 19.7: Menü ANALYSIEREN/MEHRFACHANTWORTEN/HÄUFIGKEITEN...

9. Im Fenster der Abbildung 19.7 übertragen Sie die Variable „$Anzahl“ in den Bereich TABELLE(N) FÜR:.
10. Klicken Sie OK an.

Häufigkeiten von $Anzahl

		Antworten		Prozent der Fälle
		N	Prozent	
$Anzahl[a]	spiegel	5	26,3%	50,0%
	emma	4	21,1%	40,0%
	faz	3	15,8%	30,0%
	focus	5	26,3%	50,0%
	zeit	2	10,5%	20,0%
Gesamt		19	100,0%	190,0%

a. Dichotomie-Gruppe tabellarisch dargestellt bei Wert 1.

Abb. 19.8: Auswertung der Mehrfachantworten

Unter N wird die Anzahl der Nennungen ausgewiesen.

Unter PROZENT wird gezeigt, wie viel Prozent aller ausgewerteten Antworten auf die einzelnen Zeitschriften entfallen (es ergibt sich die Summe 100).

Unter PROZENT DER FÄLLE finden Sie schließlich Angaben darüber, wie viel Prozent der befragten Personen jeweils die einzelnen Zeitschriften angekreuzt haben (wegen der Möglichkeit der Mehrfachantwort ist die Summe dieser Anteile größer als 100%, nämlich hier 190%).