Helge Toutenburg
Christian Heumann

Deskriptive Statistik

mit Beiträgen von

Michael Schomaker

Eine Einführung in Methoden und Anwendungen mit R und SPSS.

Siebente, aktualisierte und erweiterte Auflage

4. März 2009

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Inhaltsverzeichnis

i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-6	36	-1 440	2073600	8 640
2	-4	16	-740	547600	2 960
3	-1	1	-440	193600	440
4	4	16	560	313600	2 240
5	7	49	2 060	4243600	14 420
		$S_{xx} = 118$		$S_{yy} = 7372 \times 10^3$	$S_{xy} = 28\,700$

Damit ist

$$r = \frac{28\,700}{\sqrt{118 \cdot 7372 \times 10^3}} = \frac{287}{294.9} = 0.973\,.$$

Wir berechnen weiter $\tilde{\bar{x}} = 0.1\bar{x} = 1.6, \tilde{\bar{y}} = 0.625\bar{y} = 1\,525$ und

$$\begin{aligned} S_{\tilde{x}\tilde{x}} &= 0.1^2 S_{xx} = 1.18 \\ S_{\tilde{y}\tilde{y}} &= 0.625^2 S_{yy} = 287.97 \times 10^4 \\ S_{\tilde{x}\tilde{y}} &= 0.1 \cdot 0.625 S_{xy} = 1\,793.75 \end{aligned}$$

und erhalten damit (vgl. (4.42))

$$r = \frac{1\,793.75}{\sqrt{1.18 \cdot 287.97 \times 10^4}} = 0.973\,.$$

4.5 Explorative Grafiken für mehrere Variablen

Wollten wir in den vorangegangenen Kapiteln Daten und Variablen visualisieren, so haben wir uns stets auf den zwei- oder dreidimensionalen Fall beschränkt.
Stab-, Balken- oder Kreisdiagramme können uns einen Eindruck einer kategorialen Größe, Histogramme einer stetigen Variable geben. Wollen wir uns einen grafischen Überblick über zwei stetige Merkmale verschaffen, so betrachten wir Streudiagramme, sind die Merkmale diskret bzw. gemischt stetig-diskret so sind aufgesplittete Balkendiagramme oder Boxplots heranzuziehen. Sobald wir jedoch mehrere Variablen betrachten und diese zu visualisieren versuchen, benötigen wir mehr als die bisher beschriebenen Methoden.
In diesem Kapitel sollen einige einfache und schöne Konzepte vorgestellt werden, die es erlauben mehrere Variablen gleichzeitig grafisch darzustellen.

4.5.1 Coplots

Der Name Coplot entstand aus einer Abkürzung für die Bezeichnung 'conditioning scatter plots'. Dem Namen entsprechend werden dabei mehrere Streudiagramme für vorher definierte Bedingungen erstellt. Im einfachsten Fall bedeutet dies, dass neben zwei stetigen Variablen X und Y - für die

ein Streudiagramm gezeichnet werden soll - eine weitere Variable A vorliegt, die die Bedingungen vorgibt. Das heißt, dass für jede Ausprägung von A das bedingte Streudiagramm X-Y geplottet wird. Die Variable A sollte dabei natürlich nicht metrisch sein, sondern binär, kategorial oder klassiert. Tabelle 4.8 veranschaulicht noch einmal die hier vorliegende Datensituation.

Tabelle 4.8. Einfachste Datensituation für einen Coplot (links), sowie ein Beispiel mit möglichen Merkmalsausprägungen (rechts).

	X	Y	A
1	x_1	y_1	a_1
2	x_2	y_2	a_2
	.	.	.
	.	.	.
	.	.	.
n	x_n	y_n	a_n

	X	Y	A
1	4.3	2.4	0
2	3.8	2.6	1
3	4.1	2.1	0
4	3.1	2.4	0
	.	.	.
	.	.	.

Beispiel 4.5.1. Wir betrachten einen Datensatz zu seismologischer Aktivität im Gebiet der Fiji-Inseln und Tonga. Über einen Zeitraum von mehr als 30 Jahren wurde dabei bei einem Wert von mindestens '4' auf der Richter-Skala Ort und Stärke des Erdbebens festgehalten. Es liegen folgende Variablen vor:

lat	Längengrad des Ortes seismologischer Aktivität
long	Breitengrad des Ortes seismologischer Aktivität
depth	Tiefe des Bebens (in km)
mag	Stärke des Bebens (auf der Richter-Skala)
stations	Nummer der Kontrollstation

Um einen Überblick über die Schwerpunkte der Beben zu bekommen empfiehlt es sich ein Streudiagramm der Variablen der Längen- und Breitengrade zu zeichnen. Interessiert dabei nun aber auch noch ob sich die Schwerpunkte für eine unterschiedliche Tiefe der Beben unterscheiden, so könnte man einen Coplot zeichnen. Wir klassieren die Tiefe in eine binäre Variable, die den Wert '0' für eine Tiefe von '0 - 300 m' annimmt, und den Wert '1' bei einer Tiefe von mehr als 300 Metern. Insgesamt lagen 1000 Beobachtungen vor, 547 davon erhielten den Wert '0', 453 den Wert '1'.

In Abbildung 4.31 sind die Ergebnisse unseres Datensatzes zu betrachten. Es scheint so, als würde bei Beben in der Nähe der Erdoberfläche (also in einer Tiefe von 0-300 m) das Gebiet der seismologischen Aktivität weiter gestreut zu sein. Es liegen zwei größere örtliche Schwerpunkte vor, bei Beben in einer größeren Tiefe sind diese jedoch deutlich konzentrierter.

Anmerkung. Selbstverständlich muss die Variable 'depth' nicht binär kodiert sein. Wir hätten ebenfalls die Variable in beispielsweise 6 Kategorien ('0-100 m', '100-200 m',...) unterteilen können. Als Resultat hätten wir dann eben 6

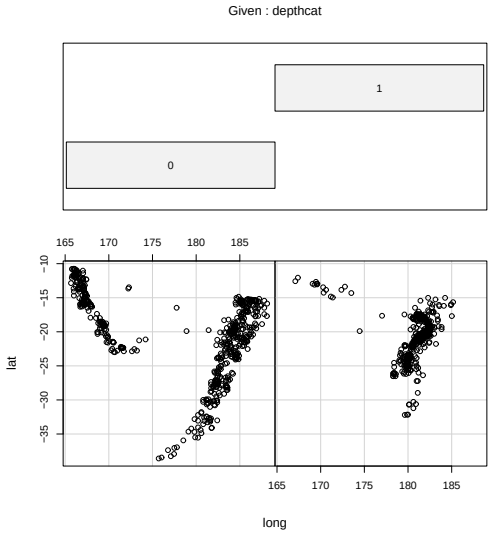

Abb. 4.31. Coplot für das Streudiagramm der Längen- und Breitengrade, aufgesplittet nach der binären Variable 'Tiefe'

verschiedene Streudiagramme erhalten. Zu beachten ist jedoch stets, dass bei einer Unterteilung für jeden Fall immer noch genug Beobachtungen vorliegen sollten.

Als Erweiterung zu der bisher beschriebenen Situation wollen wir nun den Fall näher untersuchen, bei dem zwei kategoriale Einflußgrößen vorliegen und somit ein zweidimensionaler Coplot gezeichnet werden könnte. Wir bezeichnen die neue Variable als 'B' und betrachten die Datensituation wie in Tabelle 4.9 veranschaulicht:

Tabelle 4.9. Datensituation für einen Coplot bei zwei metrischen und zwei kategorialen Variablen(links), sowie ein Beispiel mit möglichen Merkmalsausprägungen (rechts).

	X	Y	A	B
1	x_1	y_1	a_1	b_1
2	x_2	y_2	a_2	b_2
	.	.	.	.
	.	.	.	.
	.	.	.	.
n	x_n	y_n	a_n	b_n

	X	Y	A	B
1	4.3	2.4	0	6
2	3.8	2.6	1	6
3	4.1	2.1	0	7
4	3.1	2.4	0	7
	.	.	.	
	.	.	.	

Wollen wir nun Streudiagramme für die beiden Variablen X und Y zeichnen, so müssen wir dies für jede Kombination von Ausprägungen von A und B tun. Sind A und B binär erhalten wir also vier verschiedene Diagramme, bei höherer Anzahl entsprechend mehr.

Beispiel 4.5.2. Wir betrachten erneut das Beispiel der seismologischen Aktivität im Gebiete der Fiji-Inseln und Tonga. Neben der Tiefe soll nun auch noch die Stärke des Bebens als relevante Variable erfasst werden. Ist der Wert auf der Richter-Skala geringer als 4.6, so erhält unsere Variable den Wert '0', ansonsten den Wert '1'. Von den 1000 Fällen wiesen 486 einen Wert von weniger als 4.6 auf, 514 dagegen waren größer. In Abbildung 4.32 ist der entsprechende Coplot abgebildet. Auch hier ist deutlich zu erkennen, dass bei größerer Tiefe (also Tiefe = 1) weniger große Schwerpunkte zu erkennen sind.

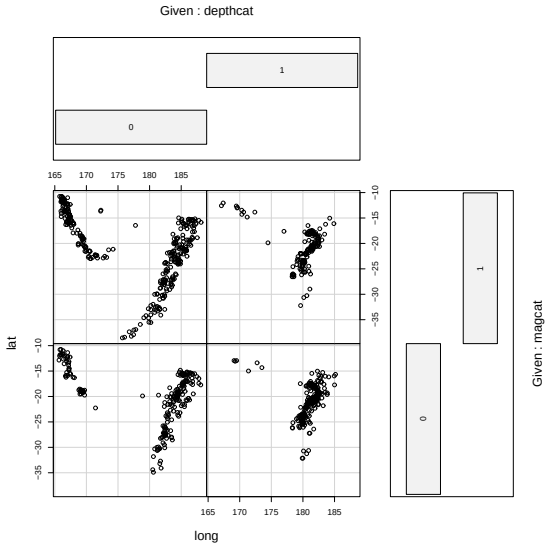

Abb. 4.32. Coplot für das Streudiagramm der Längen- und Breitengrade, aufgesplittet nach den binären Variablen 'Tiefe' und 'Stärke'

Verschiedene Programmpakete bieten auch die Möglichkeit Coplots für mehrere *stetige* Variablen zu konstruieren. Wir haben also eine Datensituation wie in Tabelle 4.10 veranschaulicht.

Das Programmpaket unterteilt dann, je nach Vorgabe, zwei der vier Variablen in sich überlappende Klassen und erstellt den Coplot wie gewohnt.

Tabelle 4.10. Datensituation für einen Coplot bei vier metrischen Variablen(links), sowie ein Beispiel mit möglichen Merkmalsausprägungen (rechts)

	X	Y	A	B
1	x_1	y_1	a_1	b_1
2	x_2	y_2	a_2	b_2
	.	.	.	.
	.	.	.	.
	.	.	.	.
n	x_n	y_n	a_n	b_n

	X	Y	A	B
1	4.3	2.4	200	66.6
2	3.8	2.6	212	62.9
3	4.1	2.1	198	71.5
4	3.1	2.4	234	70.3
	.	.	.	
	.	.	.	

Beispiel 4.5.3. Betrachten wir erneut das Beispiel der seismologischen Aktivität. Für die beiden metrischen Variablen 'Längengrad' und 'Breitengrad', erhalten wir unter der Bedingung der beiden anderen metrischen Variablen 'Tiefe' und 'Stärke' einen Coplot wie in Abbildung 4.33. Eine detailliertere Betrachtung bringt hier jedoch keine neueren Erkenntnisse.

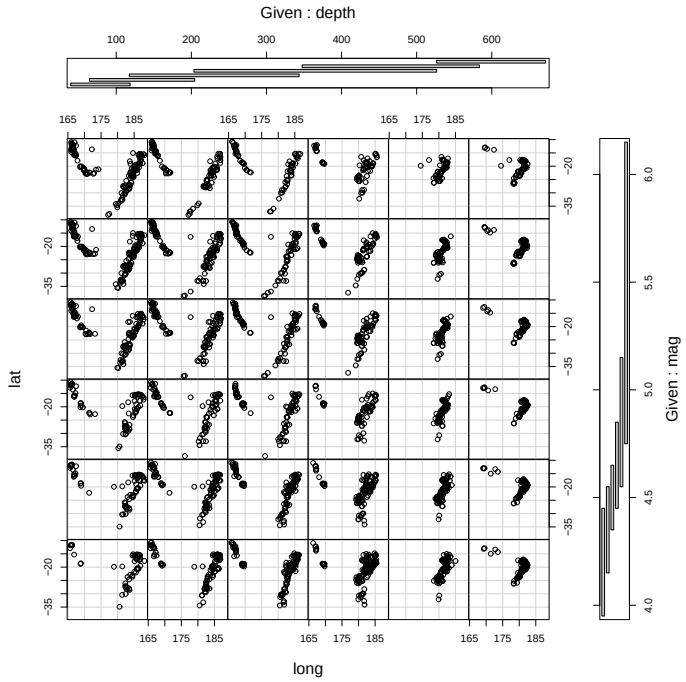

Abb. 4.33. Coplot für das Streudiagramm der Längen- und Breitengrade, aufgesplittet nach den stetigen aber klassierten Variablen 'Tiefe' und 'Stärke'

Anmerkung. Unter SPSS können Coplots sehr einfach für binäre, kategoriale oder klassierte Bedingungsvariablen erstellt werden. Unter 'Grafiken' → 'Streudiagramm' → 'einfaches Streudiagramm' → 'Felder anordnen nach' kann optional eine Unterteilungs- und damit auch Bedingungsvariable eingestellt werden.
Andere Programmpakete wie R oder S-Plus lassen unter ihrer Funktion *coplot()* auch eine metrische Bedingungsvariable zu.

4.5.2 Chernoff Faces

Die grundlegende Idee der 'Chernoff faces' ist, dass jeder Teil eines Gesichtes eine einzelne Variable repräsentiert. So kann für jede Beobachtungseinheit ein Gesicht gezeichnet werden, das die individuellen Eigenschaften dieser Einheit für mehrere (in der Regel bis zu 15) Variablen widerspiegelt. In Tabelle 4.11 ist die detaillierte Auflistung zu sehen, wie jede Variable in einem 'Chernoff face' wiederzufinden ist.

Tabelle 4.11. Chernoff faces. Repräsentation der Variablen durch Gesichtszüge

Var 1	Fläche des Gesichts	Var 9	Blickwinkel der Augen
Var 2	Form des Gesichts	Var 10	Form der Augen
Var 3	Länge der Nase	Var 11	Breite der Augen
Var 4	Ort des Mundes	Var 12	Ort der Pupille
Var 5	Krümmung des Lachens	Var 13	Ort der Augenbraue
Var 6	Breite des Mundes	Var 14	Winkel der Augenbraue
Var 7	Ort der Augen	Var 15	Breite der Augenbraue
Var 8	Distanz der Augenbrauen		

Beispiel 4.5.4. Als Beispiel betrachten wir einen Datensatz mit den Eigenschaften verschiedener Biere. So wurden für 32 verschiedene Biere die Merkmale 'Alkoholgehalt (in %)', 'Stammwürze', 'Kilokalorien pro 0.33l', 'Brauereigründung' und 'Bittereinheiten' erhoben. Tabelle 4.12 zeigt einen Auszug aus den Daten. Wir können uns nun die dazugehörigen Chernoff faces plotten lassen (siehe Abbildung 4.34).

Gemäß Tabelle 4.11 sollte also nun ein höherer Alkoholgehalt des Bieres zu einem größeren Gesicht führen, ein höherer Bittergehalt (unpassenderweise) zu einem größeren Lachen. Tatsächlich scheint auch beispielsweise Bier Nummer 10 (Clausthaler Alkoholfrei) ein sehr kleines Gesicht zu haben, Bier Nummer 3 (Erdinger) hat mit einem sehr geringen Bittergehalt von 9 ein sehr trauriges Gesicht, Bier Nummer 11 (Jever) dagegen ein sehr fröhliches.

Als eine freie Variante und Interpretation der Chernoff faces ist für das Programmpaket R eine Funktion erhältlich, die ähnliche Gesichter plottet.

Tabelle 4.12. Auszug aus den Bierdaten

Nr.	Biersorte	Alkoholgehalt	Stammwürze	kcal (pro 0.33l)	Gründungsjahr	Bittereinheiten
.	.	.	.	.	.	.
.	.	.	.	.	.	.
14	Paulaner	6.0	13.7	165	1634	24
15	Holsten	4.8	11.2	136	1879	29
16	Astra	4.9	11.2	136	1897	28
17	Maisels	5.4	12.3	142	1887	12
.	.	.	.	.	.	.
.	.	.	.	.	.	.

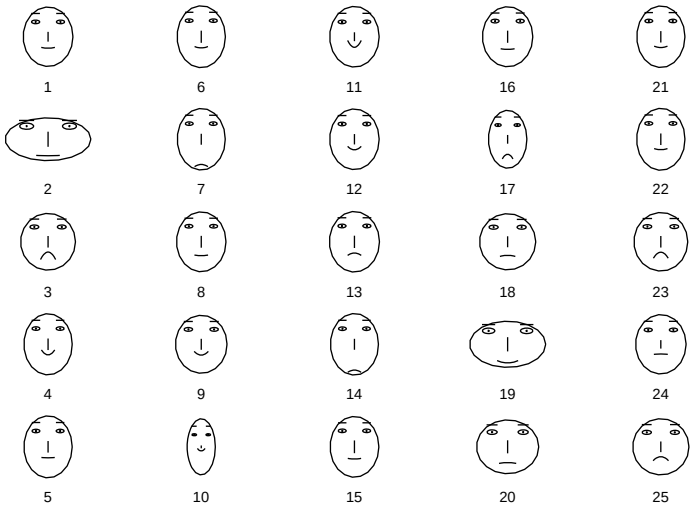

Abb. 4.34. Chernoff faces für 25 der 32 verschiedenen Biersorten

Vorteil daran ist ein detaillierteres äußeres Erscheinungsbild der Gesichter sowie eine Kennzeichnung der Gesichter mit Namen (siehe auch Abbildung 4.35).
Bei dieser Funktion würde beispielsweise ein 'kurzes' Gesicht auf einen geringen Alkoholgehalt hinweisen, ein breiter Mund dagegen auf hohe Bittereinheiten.

Anmerkung. Auch wenn die originelle Idee von Chernoff teilweise einen schnellen Überblick über viele Variablen verschaffen kann, so sind doch noch folgende wichtige Bemerkungen zu machen:

- Das Aussehen der einzelnen Gesichter hängt von der Reihenfolge der Variablen ab. Bei einer Umordnung ändert sich aus das Aussehen der 'Chernoff

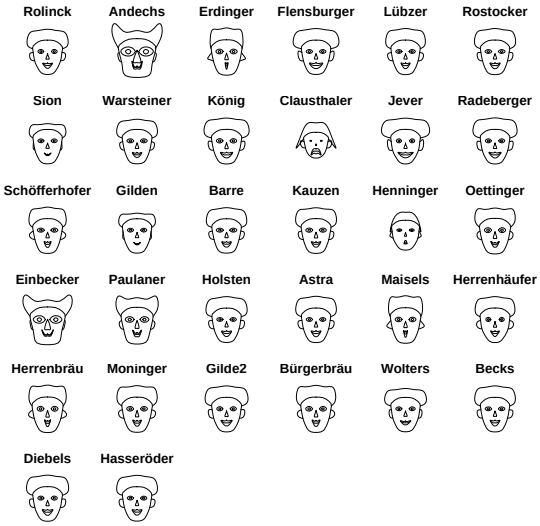

Abb. 4.35. Chernoff faces für die 32 verschiedenen Biersorten

faces'.

- Teilweise ist es schwierig die Gesichter schnell und genau zu interpretieren.

- In SPSS sind die 'Chernoff' faces leider nicht implementiert. Für andere Programmpakete (z.B. R und S-Plus), gibt es jedoch Funktionen die die Gesichter plotten. Innerhalb der verschiedenen Software gibt es jedoch teilweise kleinere Unterschiede bei der Konstruktion der Gesichter.

4.6 Sachgemäße Gestaltung von Grafiken

Bei den ersten Erfahrungen im Umgang mit statistischen Softwarepaketen oder der Interpretation von Grafiken ergibt sich für den Anwender oft das Problem, dass zu schnell falsche Rückschlüsse innerhalb eines Sachverhaltes gezogen werden. Outputs weisen eine irreführende Skalierung auf, Hypothesen werden unsachgemäß formuliert und überladene Grafiken laden zu falschen Interpretationen ein.
In diesem Kapitel soll ein erster Einblick gegeben werden, welche Fehler gemacht werden können und wie diese am besten vermieden werden.

8. Fehlende Daten

Zu Beginn wollen wir einen kleinen Rückblick auf das erste Kapitel (Grundlagen) machen. Wir erinnern uns, dass dort ein Kernsatz lautete: „Je höher die Qualität der erhobenen Daten ist, desto besser sind die Chancen für eine aussagekräftige statistische Analyse“. Dabei hatten wir sehr viel Wert gelegt auf die **Planung vor einer Datenerhebung** in einer Studie oder in einem Experiment (Auswahl der geeigneten Untersuchungseinheiten, Festlegung der zu erhebenden Merkmale). In der Praxis taucht nun häufig das Problem auf, dass trotz aller Bemühungen bei der Erhebung der Daten die Ausprägungen eines oder mehrerer Merkmale an einigen, oft auch an vielen Untersuchungseinheiten, nicht erhoben werden konnten. Wir sind also **nach Erhebung der Daten** in der Situation, die wahren Merksmalsausprägungen nicht immer beobachtet zu haben.

Allerdings kommt es nicht immer ungewollt zu fehlenden Daten. Es kann auch der Fall auftreten, dass Daten per Design, also geplant fehlen. Beispielsweise sind in Fragebögen oftmals Verzweigungen eingebaut, die dazu führen, dass bestimmte Fragen nur dann beantwortet werden sollen, wenn eine andere Frage zuvor mit einer bestimmten Merkmalsausprägung beantwortet wurde. Als triviales Beispiel diene die Frage nach Anzahl und Alter der Kinder, die nur dann sinnvoll beantwortet werden kann, wenn die Frage „Haben Sie Kinder?“ zuvor mit „ja“ beantwortet wurde. Wir werden geplantes oder systematisches Fehlen allerdings nicht näher behandeln. In den fogenden Abschnitten 8.1 und 8.2 werden wir uns mit dem ungeplanten Fehlen von Daten beschäftigen.

8.1 Betrachtung eines einzelnen Merkmals

Im Folgenden wollen wir das Problem fehlender Daten zunächst durch univariate Betrachtungen anhand eines einführenden Beispiels erläutern.

Beispiel 8.1.1. Ein Unternehmen, welches davon überzeugt ist, dass motivierte und zufriedene Mitarbeiter wichtig für den Erfolg des Unternehmens sind, führt eine schriftliche, anonyme Befragung seiner Mitarbeiter hinsichtlich der Zufriedenheit am Arbeitsplatz durch. Eines der erhobenen Merkmale lautet:

9. Einführung in SPSS

SPSS ist ein statistisches Softwarepaket und in seiner ursprünglichen Version ("Statistical Package for the Social Sciences") als anwendungsorientiertes Analyseinstrument für die Sozialwissenschaften konzipiert. Heutzutage steht das Kürzel SPSS für "Statistical Product and Service Solution" und zielt damit auf die Integration zwischen Statistik und Service ab.

Im Vergleich zu anderen statistischen Softwarepaketen wie S-Plus, R, SAS, MINITAB, etc. ist SPSS noch immer im Wesentlichen auf den Anwender fokussiert und erlaubt dadurch statistische Instrumente einfach und interaktiv einzusetzen. Dies bringt viele Vorteile, jedoch auch einige Nachteile, mit sich. Prinzipiell ist SPSS intuitiv und einfach bedienbar, es existieren eine gute Online Hilfe sowie gute Handbücher, SPSS ist Windows-konform und erstellt automatisch Programmcodes (Syntax).
Leider birgt die einfache Bedienung auch Gefahren, so werden schnell falsche Methoden angewandt und interpretiert. Auch ist die automatische Manipulation von Grafiken nur beschränkt möglich. Neben typischen Programmierwerkzeugen wie beispielsweise Schleifen fehlen auch statistische Verfahren, die in anderen Programmpaketen implementiert sind. Einzelne Prozeduren weisen Inkonsistenzen auf. Wer mit dem Textsatzprogramm Latex arbeitet wird schnell bemerken, dass ein Einbinden der Grafiken oft sehr mühselig ist.

9.1 Grundaufbau des Programms

SPSS besteht im Wesentlichen aus drei verschiedenen Fenstern bzw. Dateien:

1. Datendatei.*sav* → Hier werden die Daten entweder eingelesen oder eingegeben. Variablen können modifiziert werden, Berechnungen sind möglich und Anweisungen werden hier erteilt.
2. Ausgabedatei.*spo* → Hier werden Grafiken und Berechnungen ausgegeben. Per Mausklick können die Ausgabegrafiken und Tabellen noch verändert werden.
3. Syntaxdatei.*sps* → Hier kann der Programmcode (also die Syntax) eingesehen, gespeichert und modifiziert werden. Um Speicherplatz zu sparen wird meist die Syntax anstelle der Outputs gespeichert.

9.1.1 Das Datenfenster

Das Datenfenster spaltet sich in zwei Teile auf, die "Variablenansicht" und die "Datenansicht".

Datenansicht. Im Datenfenster mit Datenansicht können im Wesentlichen Daten eingelesen und ausgewertet werden. Typisch ist die Datenbankform der Daten:

Die Spalten beschreiben dabei Variablen bzw. Merkmale, die Zeilen stehen für die Untersuchungseinheiten.

Werden die Daten nicht eingetippt sondern liegen schon als Datei vor, so öffnet man sie über

Datei → Öffnen → Daten.

Liegen die Daten bereits als *.sav*-Datei vor, so erscheinen sofort alle Werte, ansonsten müssen noch interaktiv Fragen zu dem Datenfile beantwortet werden (z.B.: Wie sind die einzelnen Werte voneinander getrennt?).

Für die Datenanalyse wird die obere Schaltleiste benützt. Durch Mausklick können folgende Menüs aufgerufen werden:

Datei	→	Hier können mit allen *.sav, .sps, .spo* Dateien administrative Dinge wie Speichern, Laden und Umbenennen erledigt werden.
Bearbeiten	→	Ermöglicht im Wesentlichen Kopier- und Einfügearbeiten.
Ansicht	→	Regelt die visuelle Ausrichtung der Datenansicht.
Daten	→	Ermöglicht die Strukturierung eines Datenfiles. Verschiedene Datensätze können verschmolzen und (Fall-)Bedingungen ausgewählt werden.
Transformieren	→	Erlaubt die Transformation oder Umkodierung von Variablen.
Analysieren	→	Das Herzstück von SPSS. Alle statistischen Prozeduren werden hier ausgewählt.
Grafiken	→	Grafiken, speziell im Bereich deskriptiver Analysen, können hier ausgewählt werden.
Extras	→	Einige zusätzliche Optionen.
Fenster	→	Ermöglicht verschiedene Ansichten der Fenster.
Hilfe	→	Hilfe zu Themen und Syntax.

Variablenansicht. In der Variablenansicht werden die Eigenschaften der Merkmale angegeben. Auch hier können interaktiv alle erforderlichen Dinge angegeben werden:

Name	→	Der Name der Variable.
Typ	→	Ist meine Variable numerisch oder ein Wort (also ein 'String')? Liegen die Ausprägungen als Zahl, als Datum oder gar als Währung vor?
Spaltenformat	→	Hier kann die Anzahl der angezeigten Zahlen pro Feld ausgewählt werden.
Dezimalstellen	→	Wieviele Dezimalstellen sind für meine Variable relevant?
Variablenlabel	→	Wird hier ein zusätzlicher Name eingetragen, so erscheint dieser bei den ausgegebenen Grafiken und Analysen.
Wertelabels	→	Sehr wichtig und hilfreich. Speziell für binäre oder kategoriale Variablen können die Kodierungen in Worte übersetzt werden. So kann SPSS beispielsweise mitgeteilt werden, dass die Zahl '0' für männlich steht, die Zahl '1' dagegen für weiblich. Bei Outputs wird dies berücksichtigt.
Fehlende Werte	→	Zum Auswählen von Bereichen oder Werten, die fehlende Daten kodieren.
Spalten	→	Hier kann die Breite eines Feldes reguliert werden.
Ausrichtung	→	es kann ausgewählt werden ob der Text (bzw. die Werte) mittig, links oder rechts stehen soll.
Messniveau	→	es kann zwischen 'nominal', 'ordinal' und 'metrisch' gewählt werden.

9.1.2 Das Grafikfenster

Das Grafikfenster besteht im Wesentlichen aus drei Teilen: Die obere Schaltleiste ermöglicht die Analyse eines Datensatzes und unterscheidet sich nur unwesentlich von der Leiste des Datenfensters. Speziell für kurze Analysen ist so ein ständiges Wechseln zwischen den einzelnen Fenstern nicht unbedingt notwendig.
Der Großteil des Fensters besteht natürlich aus den Ausgaben selbst, also Tabellen und Grafiken. Per doppeltem Mausklick können vor allem die Grafiken editiert werden. Sehr schnell lassen sich so Farbe, Achsenskalierung und Beschriftungen ändern. Sollen Grafiken separat abgespeichert werden, so können unter *Rechte Maustaste → Exportieren* viele gängige Grafikformate ausgewählt werden.

Auf der linken Leiste des Fensters sind übersichtlich alle bisherigen Outputs angeordnet. Diese können per Mausklick schnell aufgerufen, editiert bzw. gelöscht werden.

9.1.3 Das Syntaxfenster

Um sich einen schnellen Überblick zu verschaffen, genügt es oft über eine Bestätigung per Mausklick ('*OK*') Grafiken und Tabellen ins Outputfenster zu schicken. Damit Ergebnisse nicht verloren gehen, kann selbstverständlich der gesamte Output abgespeichert werden (*.spo*). Leider ist dies sehr speicherplatzintensiv und bietet auch nicht die Möglichkeit im Nachhinein nachzuvollziehen wie Grafiken und Tabellen produziert wurden.
Als Lösung bietet es sich an, anstelle der Ausgabe die Syntax zu speichern. Hierfür wird einfach anstelle der einfachen Bestätigung (also 'OK') der Button 'Einfügen' angeklickt. Es öffnet sich ein neues Fenster (das Syntaxfenster) mit dem Programmcode. Durch Klicken auf den ▷ - Button wird der *markierte* Teil der Syntax ausgeführt.
Durch *Kopieren* und *Einfügen* können so auch sehr schnell ähnliche Operationen durchgeführt werden.

9.2 Ein praktisches Beispiel

Wir betrachten den Datensatz 'autodatensatz.sav' (entspricht dem SPSS-Datensatz 'cars.sav'). Dieser ist auch unter

'*www.stat.uni-muenchen.de/institut/ag/toutenb/daten*'

zu finden. An verschiedenen Autos wurden die Merkmale 'mpg' (Gefahrene Meilen pro Gallone), 'Hubraum', 'PS', 'Gewicht', 'Beschleunigung', 'Baujahr', 'Herkunftsland' und 'Zylinder' erhoben. Tabelle 9.1 zeigt einen Auszug aus dem Datensatz.

Tabelle 9.1. Auszug aus dem Datensatz

ID	mpg	Hubraum	PS	Gewicht	Beschl.	Baujahr	Land	Zylinder
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
3	18	307	130	3504	12	70	1	8
4	15	350	165	3693	12	70	1	8
5	18	318	150	3436	11	70	1	8
6	16	304	150	3433	12	70	1	8
7	17	302	140	3449	11	70	1	8
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

10. Einführung in R

R (R Development Core Team, 2007) ist ein statistisches Softwarepaket, das über das Internet zur Verfügung gestellt wird. Es handelt sich um ein sogenanntes *open source* Projekt, bei dem der komplette Quelltext der Software eingesehen werden kann und das unter der GNU General Public License steht. Dadurch kann es auf unterschiedlichen Betriebssystemen verwendet werden, u.a. Apple Mac OS X, Linux, Sun Solaris und Microsoft Windows. Während bei dem in Kapitel 9 eingeführten statistischen Softwarepaket SPSS die einfache Handhabung fest implementierter Prozeduren über eine grafische Benutzeroberfläche im Vordergrund steht, zeichnet sich R durch eine praktisch unbegrenzt mögliche Erweiterung durch neue Funktionen und Verfahren aus. Neben dem gut ausgetesteten Basispaket, welches bereits eine hohe Funktionalität hinsichtlich statistischer Verfahren und grafischer Darstellungsmöglichkeiten für Daten besitzt, gibt es eine große Anzahl an zusätzlichen R-Paketen mit modernsten statistischen Verfahren für die unterschiedlichsten Datensituationen und Einsatzzwecke. Dazu zählen auch Methoden, die weit über die in diesem Buch besprochenen Verfahren hinausgehen, wie z.B. modernste Verfahren zur statistischen Modellierung, zur Zeitreihenanalyse und zu datengesteuerten multivariaten Analysen mit sogenannten Data Mining Methoden. Mit Hilfe solcher Zusatzpakete kann jeder, der es wünscht, der gesamten R Benutzergemeinde neue Funktionalität zugänglich machen. Für einen Abriß der Geschichte von R, sowie Hintergründe zur Programmiersprache S, welche durch R im wesentlichen implementiert wurde, verweisen wir auf die Bücher von Ligges (2007), Dalgaard (2002) und Venables und Ripley (2002).

10.1 Installation und Grundaufbau des Programmpakets R

Der Einfachheit halber beschränken wir uns hier auf die Beschreibung der Version für das Betriebssystem Microsoft Windows. Wie bereits erwähnt, wird die Software über das Internet zur Verfügung gestellt. Einstiegspunkt ist die Webseite `http://www.r-project.org/`. Von dort bewegt man sich weiter zum sogenannten *Comprehensive R Archive Network (CRAN)*. Dort