2 Elementare Behandlung der Daten

2.1 Beschreibung und Darstellung univariater Datensätze

Wie wir an den Beispielen in Kapitel 1 gesehen haben, werden im Rahmen der multivariaten Analyse an jedem von n Objekten p Merkmale erhoben. Die Werte dieser Merkmale werden in der *Datenmatrix* $\mathbf{X}$ zusammengefasst, wobei alle Werte numerisch kodiert werden:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}.$$

Diese Datenmatrix besteht aus n Zeilen und p Spalten. Dabei ist x_{ij} der Wert des j-ten Merkmals beim i-ten Objekt. In der i-ten Zeile der Datenmatrix $\mathbf{X}$ stehen also die Werte der p Merkmale beim i-ten Objekt. In der j-ten Spalte der Datenmatrix $\mathbf{X}$ stehen die Werte des j-ten Merkmals bei allen Objekten. Oft werden die Werte der einzelnen Merkmale beim i-ten Objekt benötigt. Man fasst diese in einem Vektor $\mathbf{x}_i$ zusammen:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}. \tag{2.1}$$

Beispiel 13. Im Beispiel 2 auf Seite 3 wurden 5 Merkmale bei 20 Studenten erhoben. Also ist $n = 20$ und $p = 5$. Wir müssen die Merkmale `Geschlecht`, `MatheLK` und `Abitur88` kodieren. Beim Merkmal `Geschlecht` weisen wir der Ausprägung `w` die 1 und der Ausprägung `m` die 0 zu. Bei den beiden anderen Merkmalen ordnen wir der Ausprägung `j` eine 1 und der Ausprägung `n` eine 0 zu. Die Datenmatrix sieht also folgendermaßen aus:

A. Handl, *Multivariate Analysemethoden*, Statistik und ihre Anwendungen, 2nd ed.,
DOI 10.1007/978-3-642-14987-0_2,

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 3 & 0 & 8 \\ 0 & 0 & 4 & 0 & 7 \\ 0 & 0 & 4 & 0 & 4 \\ 0 & 0 & 4 & 0 & 2 \\ 0 & 0 & 3 & 0 & 7 \\ 1 & 0 & 3 & 0 & 6 \\ 1 & 0 & 4 & 1 & 3 \\ 1 & 0 & 3 & 1 & 7 \\ 1 & 0 & 4 & 1 & 14 \\ 0 & 1 & 3 & 0 & 19 \\ 0 & 1 & 3 & 0 & 15 \\ 0 & 1 & 2 & 0 & 17 \\ 0 & 1 & 3 & 0 & 10 \\ 1 & 1 & 3 & 0 & 22 \\ 1 & 1 & 2 & 0 & 23 \\ 1 & 1 & 2 & 0 & 15 \\ 0 & 1 & 1 & 1 & 21 \\ 1 & 1 & 2 & 1 & 10 \\ 1 & 1 & 2 & 1 & 12 \\ 1 & 1 & 4 & 1 & 17 \end{pmatrix}. \tag{2.2}$$

□

Vor einer multivariaten Analyse wird man sich die Eigenschaften der Verteilungen der einzelnen Merkmale ansehen. Aus diesem Grunde beschäftigen wir uns zunächst mit der *univariaten Analyse*. Wir betrachten also die Werte in einer Spalte der Datenmatrix $\mathbf{X}$. Bei der Beschreibung und Darstellung der Merkmale werden wir in Abhängigkeit vom Merkmal unterschiedlich vorgehen. Man unterscheidet *qualitative* und *quantitative* Merkmale. Bei qualitativen Merkmalen sind die einzelnen Merkmalsausprägungen *Kategorien*, wobei jeder Merkmalsträger zu genau einer Kategorie gehört. Kann man die Ausprägungen eines qualitativen Merkmals nicht anordnen, so ist das Merkmal *nominalskaliert*. Kann man die Kategorien anordnen, so spricht man von einem *ordinalskalierten* Merkmal. Quantitative Merkmale zeichnen sich dadurch aus, dass die Merkmalsausprägungen Zahlen sind, mit denen man rechnen kann.

Beispiel 13. (fortgesetzt) Die Merkmale `Geschlecht`, `MatheLK`, `MatheNote` und `Abitur88` sind qualitative Merkmale, wobei die Merkmale `Geschlecht`, `MatheLK` und `Abitur88` nominalskaliert sind, während das Merkmal `MatheNote` ordinalskaliert ist. Das Merkmal `Punkte` ist quantitativ. □

2.1.1 Beschreibung und Darstellung qualitativer Merkmale

Wir gehen aus von den Ausprägungen $x_1, \ldots, x_n$ eines Merkmals bei n Objekten. Man spricht von der *Urliste*. Man nennt x_i auch die i-te Beobachtung.

Beispiel 13. (fortgesetzt) Wir wollen uns das Merkmal `MatheLK` näher ansehen. Hier sind die Werte der 20 Studenten:

```
0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1.
```

Konkret gilt $x_1 = 0$. □

Die Analyse eines qualitativen Merkmals mit den Merkmalsausprägungen $A_1, \ldots, A_k$ beginnt mit dem Zählen. Man bestimmt die *absolute Häufigkeit* n_i der i-ten Merkmalsausprägung A_i.

Beispiel 13. (fortgesetzt) Von den 20 Studenten haben 11 den Mathematik-Leistungskurs besucht, während 9 ihn nicht besucht haben. Die Merkmalsausprägung A_1 sei die 0 und die Merkmalsausprägung A_2 die 1. Es gilt also $n_1 = 9$ und $n_2 = 11$. □

Ob eine absolute Häufigkeit groß oder klein ist, hängt von der Anzahl n der untersuchten Objekte ab. Wir beziehen die absolute Häufigkeit n_i auf n und erhalten die *relative Häufigkeit* h_i mit

$$h_i = \frac{n_i}{n}.$$

Beispiel 13. (fortgesetzt) Es gilt $h_1 = 0.45$ und $h_2 = 0.55$. □

Absolute und relative Häufigkeiten stellt man in einer *Häufigkeitstabelle* zusammen. Tabelle 2.1 zeigt den allgemeinen Aufbau einer Häufigkeitstabelle.

Tabelle 2.1. Allgemeiner Aufbau der Häufigkeitstabelle eines qualitativen Merkmals

Merkmals-ausprägung	absolute Häufigkeit	relative Häufigkeit
A_1	n_1	h_1
$\vdots$	$\vdots$	$\vdots$
A_k	n_k	h_k

Beispiel 13. (fortgesetzt) Für die 20 Studenten erhalten wir in Tabelle 2.2 die Häufigkeitstabelle. □

Tabelle 2.2. Häufigkeitstabelle des Merkmals MatheLK

Merkmals-ausprägung	absolute Häufigkeit	relative Häufigkeit
0	9	0.45
1	11	0.55

Die Informationen in einer Häufigkeitstabelle werden in einem *Stabdiagramm* graphisch dargestellt. Hierbei stehen in einem kartesischen Koordinatensystem auf der Abszisse die Merkmalsausprägungen und auf der Ordinate die relativen Häufigkeiten. Über jeder Merkmalsausprägung wird eine senkrechte Linie abgetragen, deren Länge der relativen Häufigkeit der Merkmalsausprägung entspricht.

Beispiel 13. (fortgesetzt) In Abbildung 2.1 ist das Stabdiagramm des Merkmals `MatheLK` zu finden. Um es leichter interpretieren zu können, haben wir bei der Achsenbeschriftung die Merkmalsausprägungen `n` und `j` gewählt. Wir erkennen an der Graphik auf einen Blick, dass die relativen Häufigkeiten der beiden Merkmalsausprägungen sich kaum unterscheiden. □

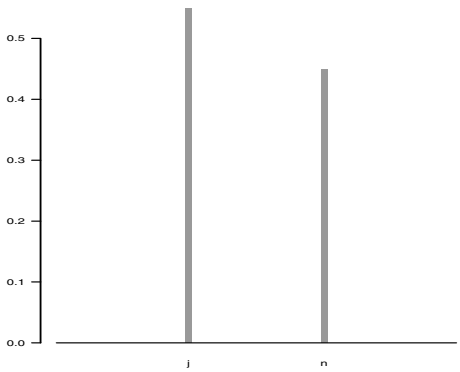

Abb. 2.1. Stabdiagramm des Merkmals MatheLK

2.1.2 Beschreibung und Darstellung quantitativer Merkmale

Beispiel 14. Im Beispiel 1 auf Seite 3 sind alle Merkmale quantitativ. Sehen wir uns das Merkmal `Mathematische Grundbildung` an. Die Urliste sieht folgendermaßen aus:

```
533 520 334 514 490 536 517 447 529 503 514 457 557
533 547 463 514 446 387 537 499 515 470 454 478 510
529 476 498 488 493 .
```

□

Die Urliste ist sehr unübersichtlich. Ordnen wir die Werte der Größe nach, so können wir bereits Struktur erkennen. Man bezeichnet die i-t kleinste Beobachtung mit $x_{(i)}$. Der *geordnete Datensatz* ist somit $x_{(1)}, \ldots, x_{(n)}$.

Beispiel 14. (fortgesetzt) Der geordnete Datensatz ist

```
334 387 446 447 454 457 463 470
476 478 488 490 493 498 499
503
510 514 514 514 515 517 520
529 529 533 533 536 537 547 557 .
```

□

Bei so vielen unterschiedlichen Werten ist es nicht sinnvoll, ein Stabdiagramm zu erstellen. Es werden Klassen gebildet. Eine Beobachtung gehört zu einer Klasse, wenn sie größer als die Untergrenze, aber kleiner oder gleich der Obergrenze dieser Klasse ist.

Beispiel 14. (fortgesetzt) Wir wählen 6 äquidistante Klassen so, dass die Untergrenze der ersten Klasse gleich 300 und die Obergrenze der letzten Klasse gleich 600 ist. Die Untergrenze der 4-ten Klasse ist 450 und die Obergrenze 500. Zur 4-ten Klasse gehören die Beobachtungen 454, 457, 463, 470, 476, 478, 488, 490, 493, 498, und 499. □

Die Häufigkeitsverteilung der Klassen wird in einem *Histogramm* dargestellt. Dabei trägt man die Klassen auf der Abszisse ab und zeichnet über jeder Klasse ein Rechteck, dessen Höhe gleich der relativen Häufigkeit der Klasse dividiert durch die Klassenbreite ist. Hierdurch ist die Fläche unter dem Histogramm gleich 1.

Beispiel 14. (fortgesetzt) Abbildung 2.2 zeigt das Histogramm des Merkmals `Mathematische Grundbildung`. Das Histogramm deutet auf eine *rechtssteile* Verteilung hin. Man bezeichnet diese auch als *linksschief*. □

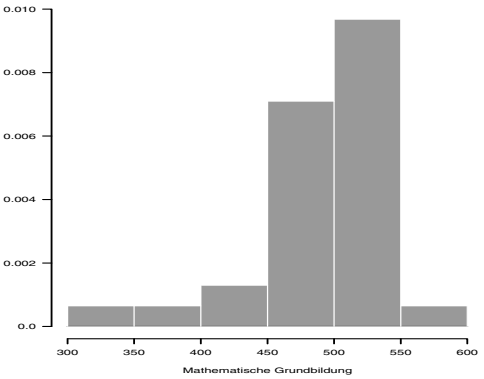

Abb. 2.2. Histogramm des Merkmals Mathematische Grundbildung

Wir wollen noch eine andere Art der Darstellung eines quantitativen univariaten Merkmals betrachten. Tukey (1977) hat vorgeschlagen, einen Datensatz durch folgende 5 Zahlen zusammenzufassen:

das *Minimum*	$x_{(1)}$,
das *untere Quartil*	$x_{0.25}$,
der *Median*	$x_{0.5}$,
das *obere Quartil*	$x_{0.75}$,
das *Maximum*	$x_{(n)}$.

Zunächst bestimmt man das Minimum $x_{(1)}$ und das Maximum $x_{(n)}$.

Beispiel 14. (fortgesetzt) Es gilt $x_{(1)} = 334$ und $x_{(n)} = 557$. □

Durch Minimum und Maximum kennen wir den Bereich, in dem die Werte liegen. Außerdem können wir mit Hilfe dieser beiden Zahlen eine einfache Maßzahl für die *Streuung* bestimmen. Die Differenz aus Maximum und Minimum nennt man die *Spannweite* R. Es gilt also

$$R = x_{(n)} - x_{(1)}. \tag{2.3}$$

Beispiel 14. (fortgesetzt) Die Spannweite beträgt 223. □

Eine Maßzahl für die *Lage* des Datensatzes ist der Median $x_{(0.5)}$. Dieser ist die Zahl, die den geordneten Datensatz in zwei gleiche Teile teilt. Ist der Stichprobenumfang ungerade, dann ist der Median die Beobachtung in der Mitte des geordneten Datensatzes. Ist der Stichprobenumfang gerade, so ist der Median der Mittelwert der beiden mittleren Beobachtungen im geordneten Datensatz. Formal kann man den Median folgendermaßen definieren:

$$x_{0.5} = \begin{cases} x_{(0.5(n+1))} & \text{falls } n \text{ ungerade ist} \\ 0.5(x_{(0.5n)} + x_{(1+0.5n)}) & \text{falls } n \text{ gerade ist.} \end{cases} \tag{2.4}$$

Beispiel 14. (fortgesetzt) Der Stichprobenumfang ist gleich 31. Der Median ist somit die Beobachtung an der 16-ten Stelle des geordneten Datensatzes. Der Wert des Medians beträgt somit 503. □

Neben dem Minimum, Maximum und Median betrachtet Tukey (1977) noch das untere Quartil $x_{0.25}$ und das obere Quartil $x_{0.75}$. 25 Prozent der Beobachtungen sind kleiner oder gleich dem unteren Quartil $x_{0.25}$ und 75 Prozent der Beobachtungen sind kleiner oder gleich dem oberen Quartil $x_{0.75}$. Das untere Quartil teilt die untere Hälfte des geordneten Datensatzes in zwei gleich große Hälften, während das obere Quartil die obere Hälfte des geordneten Datensatzes in zwei gleich große Hälften teilt. Somit ist das untere Quartil der Median der unteren Hälfte des geordneten Datensatzes, während das obere Quartil der Median der oberen Hälfte des geordneten Datensatzes ist. Ist der Stichprobenumfang gerade, so ist die untere und obere Hälfte des geordneten Datensatzes eindeutig definiert. Bei einem ungeraden Stichprobenumfang gehört der Median sowohl zur oberen als auch zur unteren Hälfte des geordneten Datensatzes.

Beispiel 14. (fortgesetzt) Das untere Quartil ist der Mittelwert aus $x_{(8)} = 470$ und $x_{(9)} = 476$ und beträgt somit 473, während das obere Quartil der Mittelwert aus $x_{(23)} = 520$ und $x_{(24)} = 529$ ist. Es beträgt 524.5. Die 5 Zahlen sind somit

$$\begin{aligned} x_{(1)} &= 334, \\ x_{0.25} &= 473, \\ x_{0.5} &= 503, \\ x_{0.75} &= 524.5, \\ x_{(n)} &= 557. \end{aligned}$$

□

Tukey (1977) hat vorgeschlagen, die 5 Zahlen in einem sogenannten *Boxplot* graphisch darzustellen. Beim Boxplot wird ein Kasten vom unteren Quartil bis zum oberen Quartil gezeichnet. Außerdem wird der Median als Linie in den Kasten eingezeichnet. Von den Rändern des Kastens bis zu den Extremen werden Linien gezeichnet, die an sogenannten *Zäunen* enden. Um Ausreißer zu markieren, wird der letzte Schritt modifiziert: Sind Punkte mehr als das 1.5-fache der Kastenbreite von den Quartilen entfernt, so wird die Linie nur bis zum 1.5-fachen der Kastenbreite gezeichnet. Alle Punkte, die außerhalb liegen, werden markiert.

Beispiel 14. (fortgesetzt) Abbildung 2.3 zeigt den Boxplot des Merkmals `Mathematische Grundbildung`. Der Boxplot deutet auch auf eine linksschiefe Verteilung hin. Außerdem ist ein Ausreißer gut zu erkennen.

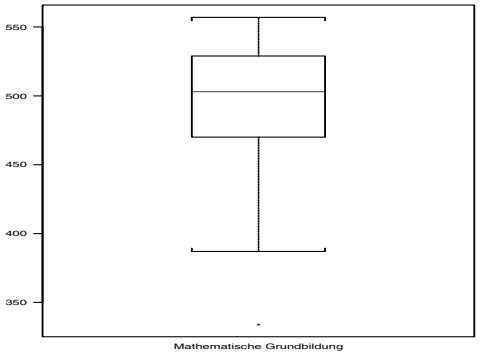

Abb. 2.3. Boxplot des Merkmals Mathematische Grundbildung im Rahmen der PISA-Studie

□

Ein wichtiger Aspekt der Verteilung eines quantitativen Merkmals ist die Lage. Wir haben bisher den Median als eine Maßzahl zur Beschreibung der Lage kennengelernt. Neben dem Median ist der *Mittelwert* $\bar{x}$ die wichtigste Maßzahl zur Beschreibung der Lage. Dieser ist folgendermaßen definiert:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Beispiel 14. (fortgesetzt) Es gilt $\bar{x} = 493.16$. Der Mittelwert ist kleiner als der Median. Dies ist bei einer linksschiefen Verteilung der Fall. □

Transformieren wir alle Beobachtungen x_i linear zu $y_i = b + a\,x_i$, so gilt

$$\bar{y} = b + a\,\bar{x}. \tag{2.5}$$

Dies sieht man folgendermaßen:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} (b + a\,x_i) = \frac{1}{n} \sum_{i=1}^{n} b + \frac{1}{n} \sum_{i=1}^{n} a\,x_i$$

$$= \frac{1}{n}\, n\, b + a\, \frac{1}{n} \sum_{i=1}^{n} x_i = b + a\,\bar{x}.$$

Bei einer symmetrischen Verteilung sind Mittelwert und Median identisch. Der Mittelwert ist ausreißerempfindlich. Eine Beobachtung, die stark von den anderen Beobachtungen abweicht, hat einen großen Einfluss auf den Mittelwert. Man sagt auch, dass der Mittelwert nicht *robust* ist. Da Ausreißer einen

starken Einfluss auf den Mittelwert haben, liegt es nahe, einen Anteil α auf beiden Seiten der geordneten Stichprobe zu entfernen und den Mittelwert der restlichen Beobachtungen zu bestimmen. Man spricht in diesem Fall von einem *getrimmten Mittelwert* $\bar{x}_\alpha$. Formal kann man diesen so beschreiben:

$$\bar{x}_\alpha = \frac{1}{n - 2\lfloor n\,\alpha \rfloor} \sum_{i=1+\lfloor n\alpha \rfloor}^{n-\lfloor n\alpha \rfloor} x_{(i)}. \tag{2.6}$$

Dabei ist $\lfloor c \rfloor$ der ganzzahlige Teil der positiven reellen Zahl c. Typische Werte für α sind 0.05 und 0.10.

Beispiel 14. (fortgesetzt) Für $n = 31$ gilt $\lfloor n\,0.05 \rfloor = 1$ und $\lfloor n\,0.1 \rfloor = 3$. Für das Merkmal `Mathematische Grundbildung` erhalten wir $\bar{x}_{0.05} = 496.45$ und $\bar{x}_{0.10} = 499.2$. □

Den Median kann man als getrimmten Mittelwert mit $\alpha = 0.5$ auffassen. Je höher der Wert von α ist, umso mehr Beobachtungen können vom Rest der Beobachtungen abweichen, ohne dass dies den getrimmten Mittelwert beeinflusst.

Neben der Lage ist die Streuung von größtem Interesse. Wir haben bereits die Spannweite R als Maß für die Streuung kennengelernt. Ein anderes Maß für die Streuung ist die *Stichprobenvarianz*. Diese ist definiert durch

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{2.7}$$

Beispiel 14. (fortgesetzt) Es gilt $s_x^2 = 2192.873$. □

Die Stichprobenvarianz besitzt nicht die gleiche Maßeinheit wie die Beobachtungen. Zieht man aus der Stichprobenvarianz die Quadratwurzel, so erhält man eine Maßzahl, die die gleiche Dimension wie die Beobachtungen besitzt. Diese heißt *Standardabweichung* s_x.

Beispiel 14. (fortgesetzt) Es gilt $s_x = 46.83$. □

Transformieren wir alle Beobachtungen x_i linear zu $y_i = b + a\,x_i$, so gilt

$$s_y^2 = a^2 s_x^2. \tag{2.8}$$

Dies sieht man mit Gleichung (2.5) folgendermaßen:

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (b + a\,x_i - b - a\,\bar{x})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (a\,(x_i - \bar{x}))^2 = a^2\,\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = a^2 s_x^2.$$

2.2 Beschreibung und Darstellung multivariater Datensätze

Bisher haben wir nur ein einzelnes Merkmal analysiert. Nun wollen wir mehrere Merkmale gemeinsam betrachten, um zum Beispiel Abhängigkeitsstrukturen zwischen den Merkmalen aufzudecken. Wir gehen davon aus, dass an jedem von n Objekten p Merkmale erhoben wurden. Wir wollen zeigen, wie man Informationen in Datenmatrizen einfach darstellen kann. Dabei wollen wir wieder zwischen qualitativen und quantitativen Merkmalen unterscheiden.

2.2.1 Beschreibung und Darstellung von Datenmatrizen quantitativer Merkmale

Beispiel 15. Wir betrachten den Datensatz im Beispiel 1 auf Seite 3 und stellen die Daten in einer Datenmatrix zusammen. In der ersten Spalte stehen die Werte des Merkmals `Lesekompetenz`, in der zweiten Spalte die Werte des Merkmals `Mathematische Grundbildung` und in der letzten Spalte die Werte des Merkmals `Naturwissenschaftliche Grundbildung`:

$$\mathbf{X} = \begin{pmatrix} 528 & 533 & 528 \\ 507 & 520 & 496 \\ 396 & 334 & 375 \\ 497 & 514 & 481 \\ 484 & 490 & 487 \\ \vdots & \vdots & \vdots \\ 492 & 498 & 511 \\ 480 & 488 & 496 \\ 504 & 493 & 499 \end{pmatrix}.$$

In der fünften Zeile der Matrix $\mathbf{X}$ stehen die Merkmalsausprägungen von Deutschland:

$$\mathbf{x}_5 = \begin{pmatrix} 484 \\ 490 \\ 487 \end{pmatrix}.$$

□

Wir wollen nun das Konzept des Mittelwerts auf mehrere Merkmale übertragen. Dies ist ganz einfach. Wir bestimmen den Mittelwert jedes Merkmals und fassen diese Mittelwerte zum Vektor der Mittelwerte zusammen. Wir bezeichnen den Mittelwert des j-ten Merkmals mit $\bar{x}_j$. Es gilt also

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}\,.$$

Für den Vektor $\bar{\mathbf{x}}$ der Mittelwerte gilt also

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}.$$

Beispiel 15. (fortgesetzt) Es gilt

$$\bar{\mathbf{x}} = \begin{pmatrix} 493.45 \\ 493.16 \\ 492.61 \end{pmatrix}. \tag{2.9}$$

Wir sehen, dass im Bereich `Lesekompetenz` im Durchschnitt am meisten Punkte erreicht wurden, während die Leistungen im Bereich `Naturwissenschaftliche Grundbildung` im Mittel am schlechtesten waren. □

Mit den Beobachtungsvektoren $\mathbf{x}_i$, $i = 1, \ldots, n$, aus Gleichung (2.1) können wir den Vektor $\bar{\mathbf{x}}$ der Mittelwerte auch bestimmen durch

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \,.$$

Dies sieht man folgendermaßen:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ip} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \,.$$

Manche der multivariaten Verfahren, die wir betrachten werden, gehen davon aus, dass die Merkmale *zentriert* sind. Wir zentrieren die Werte des i-ten Merkmals, indem wir von jedem Wert x_{ij} den Mittelwert $\bar{x}_j$ subtrahieren:

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j \,.$$

Der Mittelwert eines zentrierten Merkmals ist gleich 0. Dies sieht man folgendermaßen:

$$\bar{\tilde{x}}_j = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) = \frac{1}{n} \sum_{i=1}^{n} x_{ij} - \frac{1}{n} \sum_{i=1}^{n} \bar{x}_j = \bar{x}_j - \frac{1}{n}\, n\, \bar{x}_j = 0 \,.$$

Die *zentrierte Datenmatrix* ist

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & \ldots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \ldots & x_{np} - \bar{x}_p \end{pmatrix}. \tag{2.10}$$

Beispiel 15. (fortgesetzt) Es gilt

$$\tilde{\mathbf{X}} = \begin{pmatrix} 34.55 & 39.84 & 35.39 \\ 13.55 & 26.84 & 3.39 \\ -97.45 & -159.16 & -117.61 \\ 3.55 & 20.84 & -11.61 \\ -9.45 & -3.16 & -5.61 \\ \vdots & \vdots & \vdots \\ -1.45 & 4.84 & 18.39 \\ -13.45 & -5.16 & 3.39 \\ 10.55 & -0.16 & 6.39 \end{pmatrix}.$$

An der zentrierten Datenmatrix kann man sofort erkennen, wie sich jedes Land vom Mittelwert unterscheidet. Wir sehen, dass Deutschland als fünftes Land in der Matrix in allen Bereichen unter dem Durchschnitt liegt, während Australien als erstes Land in der Matrix in allen Bereichen über dem Durchschnitt liegt. □

Wir wollen uns nun noch anschauen, wie man die zentrierte Datenmatrix durch eine einfache Multiplikation mit einer anderen Matrix gewinnen kann. Diese Matrix werden wir im Folgenden öfter verwenden. Sei

$$\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}'. \tag{2.11}$$

Dabei ist $\mathbf{I}_n$ die Einheitsmatrix und $\mathbf{1}$ der Einservektor. Es gilt

$$\tilde{\mathbf{X}} = \mathbf{M}\mathbf{X}. \tag{2.12}$$

Um Gleichung (2.12) zu zeigen, formen wir sie um:

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X} = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}.$$

Wir betrachten zunächst $\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}$. Da $\mathbf{1}$ der summierende Vektor ist, gilt

$$\mathbf{1}'\mathbf{X} = (\sum_{i=1}^{n} x_{i1}, \ldots, \sum_{i=1}^{n} x_{ip}).$$

Da $\frac{1}{n}$ ein Skalar ist, gilt

$$\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{1}\frac{1}{n}\mathbf{1}'\mathbf{X}.$$

Es gilt

$$\frac{1}{n}\mathbf{1}'\mathbf{X} = (\bar{x}_1, \ldots, \bar{x}_p).$$

Somit folgt

$$\frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}'\mathbf{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (\bar{x}_1, \ldots, \bar{x}_p) = \begin{pmatrix} \bar{x}_1 & \ldots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \ldots & \bar{x}_p \end{pmatrix}.$$

Also gilt

$$\mathbf{M}\mathbf{X} = \mathbf{X} - \frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 & \ldots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \ldots & \bar{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} x_{11} - \bar{x}_1 & \ldots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \ldots & x_{np} - \bar{x}_p \end{pmatrix} = \tilde{\mathbf{X}}\,.$$

Man nennt $\mathbf{M}$ auch die *Zentrierungsmatrix*. Sie ist symmetrisch. Es gilt nämlich

$$\mathbf{M}' = (\mathbf{I}_n - \frac{1}{n}\,\mathbf{1}\mathbf{1}')' = \mathbf{I}_n' - (\frac{1}{n}\,\mathbf{1}\mathbf{1}')' = \mathbf{I}_n - \frac{1}{n}\,\mathbf{1}\mathbf{1}' = \mathbf{M}\,.$$

Multipliziert man die Datenmatrix also von rechts mit der Matrix

$$\mathbf{M} = \mathbf{I}_p - \frac{1}{p}\,\mathbf{1}\mathbf{1}',$$

so werden die Zeilen zentriert.

Bei der univariaten Datenanalyse haben wir robuste Schätzer wie den Median und den getrimmten Mittelwert betrachtet. Wir wollen nun aufzeigen, wie man diese Konzepte auf den multivariaten Fall übertragen kann. Hierbei werden wir uns aber auf den zweidimensionalen Fall beschränken, da nur in diesem Fall Funktionen in `S-PLUS` existieren. Beginnen wir mit dem Trimmen. Hierzu stellen wir die Werte der beiden Merkmale in einem *Streudiagramm* dar. Die beiden Merkmale bilden die Achsen in einem kartesischen Koordinatensystem. Die Werte jedes Objekts werden als Punkt in dieses Koordinatensystem eingetragen.

Beispiel 15. (fortgesetzt) Abbildung 2.4 zeigt das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung`. □

Bei nur einem Merkmal ist das Trimmen eindeutig. Man ordnet die Werte der Größe nach und entfernt jeweils einen Anteil α der extremen Werte auf beiden Seiten der geordneten Stichprobe. Bei zwei Merkmalen gibt es keine

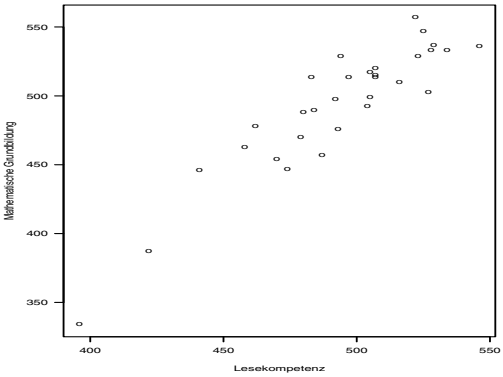

Abb. 2.4. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

natürliche Ordnung. Natürlich kann man jedes der beiden Merkmale getrennt trimmen. Hierbei berücksichtigt man aber nicht, dass beide Merkmale an demselben Objekt erhoben wurden. Es gibt nun eine Reihe von Vorschlägen, wie man im zweidimensionalen Raum trimmen kann. Wir wollen uns einen von diesen anschauen. Man bestimmt hierzu zunächst die *konvexe Hülle* der Menge der Beobachtungen. Die konvexe Hülle ist das kleinste Polygon, in dem entweder jede Beobachtung auf dem Rand oder innerhalb des Polygons liegt. Büning (1991), S.202 veranschaulicht die Konstruktion der konvexen Hülle folgendermaßen:

> Wir können uns die Punkte $\mathbf{x}_1, \ldots, \mathbf{x}_n$ als Nägel auf einem Brett vorstellen, um die ein (großes) elastisches Band gespannt und dann losgelassen wird; das Band kommt in Form eines Polygons zur Ruhe.

Beispiel 15. (fortgesetzt) Abbildung 2.5 zeigt die konvexe Hülle der Beobachtungen der Merkmale `Lesekompetenz` und `Mathematische Grundbildung`. Auf der konvexen Hülle liegen die Länder `IRL`, `BR`, `L`, `FL`, `CH`, `J` und `FIN`. □

Einen auf der konvexen Hülle basierenden getrimmten Mittelwert erhält man dadurch, dass man alle Beobachtungen auf der konvexen Hülle aus dem Datensatz entfernt und den Mittelwert der restlichen Beobachtungen bestimmt.

Beispiel 15. (fortgesetzt) Es sind 7 Punkte auf der konvexen Hülle. Somit beträgt der Trimmanteil $7/31 = 0.23$. Der getrimmte Mittelwert beträgt

$$\bar{\mathbf{x}}_{0.23} = \begin{pmatrix} 495.33 \\ 494.54 \end{pmatrix}.$$

Ein Vergleich mit den ersten beiden Komponenten von $\bar{\mathbf{x}}$ in (2.9) zeigt, dass sich dieser nicht stark vom Mittelwert unterscheidet. □

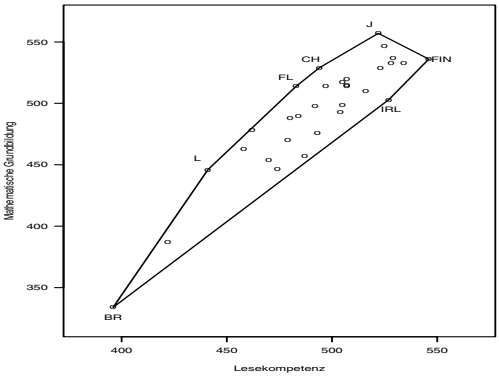

Abb. 2.5. Konvexe Hülle der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

Im Englischen nennt nennt diese Vorgehensweise *Peeling*. Heiler und Michels (1994) verwenden den Begriff *Schälen*. Man kann nun eine konvexe Hülle nach der anderen entfernen, bis nur noch eine übrig bleibt. Liegt innerhalb dieser Hülle noch ein Punkt, so ist dieser der *multivariate Median*. Liegt innerhalb dieser Hülle kein Punkt, so wählt man den Mittelwert der Beobachtungen auf der innersten Hülle als multivariaten Median. Heiler und Michels (1994), S.237 nennen ihn auch *Konvexe-Hüllen-Median*.

Beispiel 15. (fortgesetzt) Abbildung 2.6 zeigt die innerste Hülle des Datensatzes. Wir sehen, dass innerhalb dieser konvexen Hülle ein Punkt liegt. Es

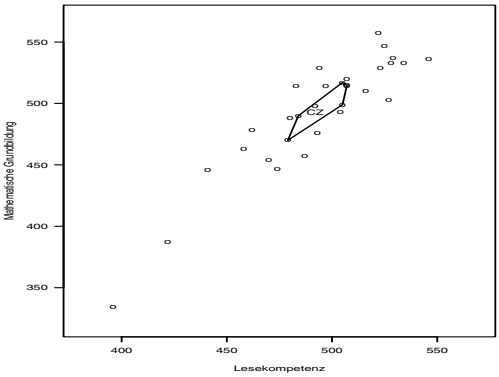

Abb. 2.6. Innerste konvexe Hülle der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

handelt sich um CZ. Also bilden die Werte von Tschechien den multivariaten Median. Dieser ist gegeben durch

$$\begin{pmatrix} 492 \\ 498 \end{pmatrix}.$$

□

Weitere Ansätze zur Bestimmung eines multivariaten Medians sind bei Büning (1991), Heiler und Michels (1994) und Small (1990) zu finden.

Ein Maß für die Streuung eines univariaten Merkmals ist die Stichprobenvarianz. In Analogie zu (2.7) ist die Stichprobenvarianz des j-ten Merkmals definiert durch

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \,. \tag{2.13}$$

Die Standardabweichung des j-ten Merkmals ist $s_j = \sqrt{s_j^2}$.

Beispiel 15. (fortgesetzt) Die Stichprobenvarianzen der einzelnen Merkmale sind $s_1^2 = 1109.4$, $s_2^2 = 2192.9$ und $s_3^2 = 1419.0$. Wir sehen, dass die Punkte am stärksten im Bereich Mathematische Grundbildung und am wenigsten im Bereich Lesekompetenz streuen. Die Standardabweichungen der Merkmale sind $s_1 = 33.3$, $s_2 = 46.8$ und $s_3 = 37.7$. □

Wir haben in Gleichung (2.10) die Merkmale zentriert. Dividiert man die Werte eines zentrierten Merkmals noch durch die Standardabweichung dieses Merkmals, so erhält man *standardisierte* Merkmale

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \,.$$

Der Mittelwert eines standardisierten Merkmals ist gleich 0. Dies sieht man folgendermaßen:

$$\overline{x_j^*} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^* = \frac{1}{n} \sum_{i=1}^{n} \frac{x_{ij} - \bar{x}_j}{s_j} = \frac{1}{n\, s_j} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) = 0 \,.$$

Die Stichprobenvarianz der standardisierten Merkmale ist gleich 1. Dies sieht man folgendermaßen:

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2 &= \frac{1}{s_j^2} \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \\ &= \frac{1}{s_j^2}\, s_j^2 = 1. \end{aligned}$$

Die *Matrix der standardisierten Merkmale* ist:

$$\mathbf{X}^* = \begin{pmatrix} \frac{x_{11} - \bar{x}_1}{s_1} & \cdots & \frac{x_{1p} - \bar{x}_p}{s_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_1}{s_1} & \cdots & \frac{x_{np} - \bar{x}_p}{s_p} \end{pmatrix}. \tag{2.14}$$

Beispiel 15. (fortgesetzt) Es gilt

$$\mathbf{X}^* = \begin{pmatrix} 1.037 & 0.851 & 0.939 \\ 0.407 & 0.573 & 0.090 \\ -2.926 & -3.399 & -3.122 \\ 0.107 & 0.445 & -0.308 \\ -0.284 & -0.068 & -0.149 \\ \vdots & \vdots & \vdots \\ -0.044 & 0.103 & 0.488 \\ -0.404 & -0.110 & 0.090 \\ 0.317 & -0.003 & 0.170 \end{pmatrix}.$$

□

Es ist nicht üblich, in Analogie zum Vektor der Mittelwerte einen Vektor der Stichprobenvarianzen zu bilden. Die Stichprobenvarianzen sind Bestandteil der *empirischen Varianz-Kovarianz-Matrix*. Um diese zu erhalten, benötigen wir die *empirische Kovarianz*, die wir nun herleiten wollen. Bisher haben wir uns die Charakteristika jedes einzelnen Merkmals angeschaut. In der multivariaten Analyse sind aber Zusammenhänge zwischen Merkmalen von Interesse.

Beispiel 15. (fortgesetzt) Schauen wir uns unter diesem Aspekt noch einmal das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung` in Abbildung 2.4 auf Seite 26 an. Wir sehen, dass Länder, die eine hohe Punktezahl im Bereich `Lesekompetenz` aufweisen, auch im Bereich `Mathematische Grundbildung` eine hohe Punktezahl erreichen. Länder mit einer niedrigen Punktezahl im Bereich `Lesekompetenz` weisen in der Regel auch einen niedrigen Wert im Bereich `Mathematische Grundbildung` auf. Ist ein Land also über dem Durchschnitt in einem Bereich, so ist es in der Regel auch über dem Durchschnitt im anderen Bereich. Dies wird auch am Streudiagramm deutlich, wenn wir die Mittelwerte der beiden Merkmale in diesem berücksichtigen. Hierzu zeichnen wir eine Gerade parallel zur Ordinate in Höhe des Mittelwerts der Punktezahl im Bereich `Lesekompetenz` und eine Gerade parallel zur Abszisse in Höhe des Mittelwerts der Punktezahl im Bereich `Mathematische Grundbildung`. Abbildung 2.7 veranschaulicht dies. Hierdurch erhalten wir 4 Quadranten, die in der Graphik durchnummeriert sind. Im ersten Quadranten sind die Länder, deren Punktezahl in den Bereichen `Lesekompetenz` und `Mathematische Grundbildung` über

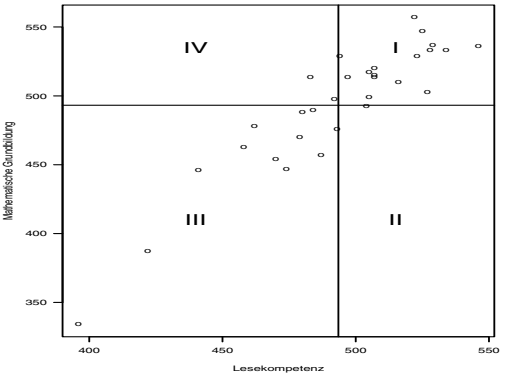

Abb. 2.7. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie, aufgeteilt in 4 Quadranten

dem Durchschnitt liegen, während sich im dritten Quadranten die Länder befinden, deren Punktezahl in den Bereichen `Lesekompetenz` und `Mathematische Grundbildung` unter dem Durchschnitt liegen. Im zweiten Quadranten sind die Länder, deren Punktezahl im Bereich `Lesekompetenz` über dem Durchschnitt, im Bereich `Mathematische Grundbildung` hingegen unter dem Durchschnitt liegen, während im vierten Quadranten die Länder liegen, deren Punktezahl im Bereich `Lesekompetenz` unter dem Durchschnitt, im Bereich `Mathematische Grundbildung` hingegen über dem Durchschnitt liegen. Besteht ein positiver Zusammenhang zwischen den beiden Merkmalen, so werden wir die meisten Beobachtungen in den Quadranten I und III erwarten, während wir bei einem negativen Zusammenhang die meisten in den Quadranten II und IV erwarten. Verteilen sich die Punkte gleichmäßig über die Quadranten, so liegt kein Zusammenhang zwischen den Merkmalen vor.

□

Um den im Beispiel veranschaulichten Sachverhalt in eine geeignete Maßzahl für den Zusammenhang zwischen den beiden Merkmalen umzusetzen, gehen wir davon aus, dass das i-te Merkmal auf der Abszisse und das j-te Merkmal auf der Ordinate stehe. Sei x_{ki} die Ausprägung des i-ten Merkmals beim k-ten Objekt und x_{kj} die Ausprägung des j-ten Merkmals beim k-ten Objekt. Dann gilt in den einzelnen Quadranten:

Quadrant I: $x_{ki} > \bar{x}_i \ , \ x_{kj} > \bar{x}_j,$

Quadrant II: $x_{ki} > \bar{x}_i \ , \ x_{kj} < \bar{x}_j,$

Quadrant III: $x_{ki} < \bar{x}_i \ , \ x_{kj} < \bar{x}_j,$

Quadrant IV: $x_{ki} < \bar{x}_i \ , \ x_{kj} > \bar{x}_j.$

Also gilt

$$\begin{aligned}
&\text{Quadrant I:} && x_{ki} - \bar{x}_i > 0\ ,\ x_{kj} - \bar{x}_j > 0,\\
&\text{Quadrant II:} && x_{ki} - \bar{x}_i > 0\ ,\ x_{kj} - \bar{x}_j < 0,\\
&\text{Quadrant III:} && x_{ki} - \bar{x}_i < 0\ ,\ x_{kj} - \bar{x}_j < 0,\\
&\text{Quadrant IV:} && x_{ki} - \bar{x}_i < 0\ ,\ x_{kj} - \bar{x}_j > 0.
\end{aligned}$$

Also ist das Produkt $(x_{ki} - \bar{x}_i) \cdot (x_{kj} - \bar{x}_j)$ im ersten und dritten Quadranten positiv, während es im zweiten und vierten Quadranten negativ ist. Dies legt nahe, folgende Maßzahl zu betrachten:

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)\,. \tag{2.15}$$

s_{ij} heißt empirische Kovarianz zwischen dem i-ten und j-ten Merkmal. Es gilt

$$s_{jj} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)(x_{kj} - \bar{x}_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2 = s_j^2.$$

Bei p Merkmalen $x_1, \ldots, x_p$ bestimmt man zwischen allen Paaren von Merkmalen die Kovarianz und stellt diese Kovarianzen in der empirischen Varianz-Kovarianz-Matrix zusammen:

$$\mathbf{S} = \begin{pmatrix} s_1^2 & \ldots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \ldots & s_p^2 \end{pmatrix}.$$

Wegen $s_{ij} = s_{ji}$ ist die empirische Varianz-Kovarianz-Matrix symmetrisch.

Beispiel 15. (fortgesetzt) Es gilt

$$\mathbf{S} = \begin{pmatrix} 1109.4 & 1428.3 & 1195.6 \\ 1428.3 & 2192.9 & 1644.0 \\ 1195.6 & 1644.0 & 1419.0 \end{pmatrix}.$$

Wir sehen, dass alle empirischen Kovarianzen positiv sind. Die empirische Kovarianz zwischen den Merkmalen `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` ist am größten. □

Man kann die empirische Varianz-Kovarianz-Matrix auch folgendermaßen bestimmen:

$$\mathbf{S} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})\,(\mathbf{x}_k - \bar{\mathbf{x}})'. \tag{2.16}$$

Mit

$$\mathbf{x}_k = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kp} \end{pmatrix}$$

und

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

gilt

$$(\mathbf{x}_k - \bar{\mathbf{x}})\,(\mathbf{x}_k - \bar{\mathbf{x}})' = \begin{pmatrix} x_{k1} - \bar{x}_1 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} (x_{k1} - \bar{x}_1 \ \ldots \ x_{kp} - \bar{x}_p\ )$$

$$= \begin{pmatrix} (x_{k1} - \bar{x}_1)\,(x_{k1} - \bar{x}_1) & \ldots & (x_{k1} - \bar{x}_1)\,(x_{kp} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{kp} - \bar{x}_p)\,(x_{k1} - \bar{x}_1) & \ldots & (x_{kp} - \bar{x}_p)\,(x_{kp} - \bar{x}_p) \end{pmatrix}.$$

Summieren wir diese Matrizen von $k = 1$ bis n und dividieren die Summe durch $n - 1$, so erhalten wir die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$. Es gibt noch eine weitere Darstellung der empirischen Varianz-Kovarianz-Matrix, auf die wir noch häufiger zurückkommen werden. Sei $\tilde{\mathbf{X}}$ die zentrierte Datenmatrix aus Gleichung (2.10) auf Seite 23. Dann gilt

$$\mathbf{S} = \frac{1}{n-1}\,\tilde{\mathbf{X}}'\tilde{\mathbf{X}}. \tag{2.17}$$

Das Element in der i-ten Zeile und j-ten Spalte von $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ erhält man dadurch, dass man das innere Produkt aus den Vektoren bildet, die in der i-ten und der j-ten Spalte von $\tilde{\mathbf{X}}$ stehen. Dieses ist

$$\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)\,(x_{kj} - \bar{x}_j).$$

Dividiert man diesen Ausdruck durch $n - 1$, so erhält man die empirische Kovarianz zwischen dem i-ten und j-ten Merkmal, wie man durch einen Vergleich mit Gleichung (2.15) erkennt.

Die empirische Kovarianz ist nicht skaleninvariant. Multipliziert man alle Werte des einen Merkmals mit einer Konstanten b und die Werte des anderen Merkmals mit einer Konstanten c, so wird die empirische Kovarianz bc-mal so groß.

Mit (2.5) gilt nämlich

$$\begin{aligned}\frac{1}{n-1}\sum_{k=1}^{n}(b\,x_{ki}-\overline{b\,x_i})(c\,x_{kj}-\overline{c\,x_j}) &= \frac{1}{n-1}\sum_{k=1}^{n}(b\,x_{ki}-b\,\overline{x}_i)(c\,x_{kj}-c\,\overline{x}_j)\\ &= b\,c\,\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)\\ &= b\,c\,s_{ij}.\end{aligned}$$

Man kann die empirische Kovarianz normieren, indem man sie durch das Produkt der Standardabweichungen der beiden Merkmale dividiert. Man erhält dann den *empirischen Korrelationskoeffizienten*

$$r_{ij}=\frac{s_{ij}}{s_i\,s_j}. \tag{2.18}$$

Für den empirischen Korrelationskoeffizienten r_{ij} gilt:

1. $-1 \le r_{ij} \le 1$,
2. $r_{ij} = 1$ genau dann, wenn zwischen den beiden Merkmalen ein exakter linearer Zusammenhang mit positiver Steigung besteht,
3. $r_{ij} = -1$ genau dann, wenn zwischen den beiden Merkmalen ein exakter linearer Zusammenhang mit negativer Steigung besteht.

Wir wollen diese Eigenschaften hier nicht beweisen. Wir beweisen sie in Kapitel 3 für den Korrelationskoeffizienten. Hier wollen wir diese Eigenschaften aber interpretieren. Die erste Eigenschaft besagt, dass der empirische Korrelationskoeffizient Werte zwischen -1 und 1 annimmt, während die beiden anderen Eigenschaften erklären, wie wir die Werte des empirischen Korrelationskoeffizienten zu interpretieren haben. Liegt der Wert des empirischen Korrelationskoeffizienten in der Nähe von 1, so liegt ein positiver linearer Zusammenhang zwischen den beiden Merkmalen vor, während ein Wert in der Nähe von -1 auf einen negativen linearen Zusammenhang hindeutet. Ein Wert in der Nähe von 0 spricht dafür, dass kein linearer Zusammenhang zwischen den beiden Merkmalen vorliegt. Dies bedeutet aber nicht notwendigerweise, dass gar kein Zusammenhang zwischen den beiden Merkmalen besteht, wie das Beispiel in Tabelle 2.3 zeigt. Der Wert des Korrelationskoeffizienten zwischen den beiden Merkmalen beträgt 0. Schaut man sich die Werte in der Tabelle genauer an, so stellt man fest, dass $x_{k2} = x_{k1}^2$ gilt. Zwischen den beiden Merkmalen besteht also ein funktionaler Zusammenhang.

Tabelle 2.3. Werte der Merkmale x_1 und x_2

k	x_{k1}	x_{k2}
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

Wir stellen die Korrelationen in der *empirischen Korrelationsmatrix* $\mathbf{R}$ zusammen:

$$\mathbf{R} = \begin{pmatrix} r_{11} & \dots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \dots & r_{pp} \end{pmatrix}. \tag{2.19}$$

Beispiel 15. (fortgesetzt) Es gilt

$$\mathbf{R} = \begin{pmatrix} 1 & 0.916 & 0.953 \\ 0.916 & 1 & 0.932 \\ 0.953 & 0.932 & 1 \end{pmatrix}.$$

Es fällt auf, dass alle Elemente der empirischen Korrelationsmatrix positiv sind. □

Man kann die empirische Korrelationsmatrix auch mit Hilfe der Matrix der standardisierten Merkmale (2.14) bestimmen. Es gilt

$$\mathbf{R} = \frac{1}{n-1} \mathbf{X}^{*\prime} \mathbf{X}^{*}. \tag{2.20}$$

Das Element in der i-ten Zeile und j-ten Spalte von $\mathbf{X}^{*\prime}\mathbf{X}^{*}$ erhält man dadurch, dass man das innere Produkt aus den Vektoren bildet, die in der i-ten und der j-ten Spalte von $\mathbf{X}^{*}$ stehen. Dieses ist

$$\sum_{k=1}^{n} \frac{x_{ki} - \bar{x}_i}{s_i} \frac{x_{kj} - \bar{x}_j}{s_j} = \frac{1}{s_i\, s_j} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)\,(x_{kj} - \bar{x}_j).$$

Dividiert man diesen Ausdruck durch $n-1$, so erhält man den empirischen Korrelationskoeffizienten zwischen dem i-ten und j-ten Merkmal, wie man durch einen Vergleich mit Gleichung (2.18) erkennt. In der empirischen Korrelationsmatrix sind die Zusammenhänge zwischen allen Paaren von Merkmalen zusammengefasst. Eine hierzu analoge graphische Darstellung ist die *Streudiagrammmatrix*. Hier werden die Streudiagramme aller Paare von Merkmalen in einer Matrix zusammengefasst.

Beispiel 15. (fortgesetzt) Die Streudiagrammmatrix ist in Abbildung 2.8 zu finden. Wir sehen hier auf einen Blick, dass alle Merkmale miteinander positiv korreliert sind. □

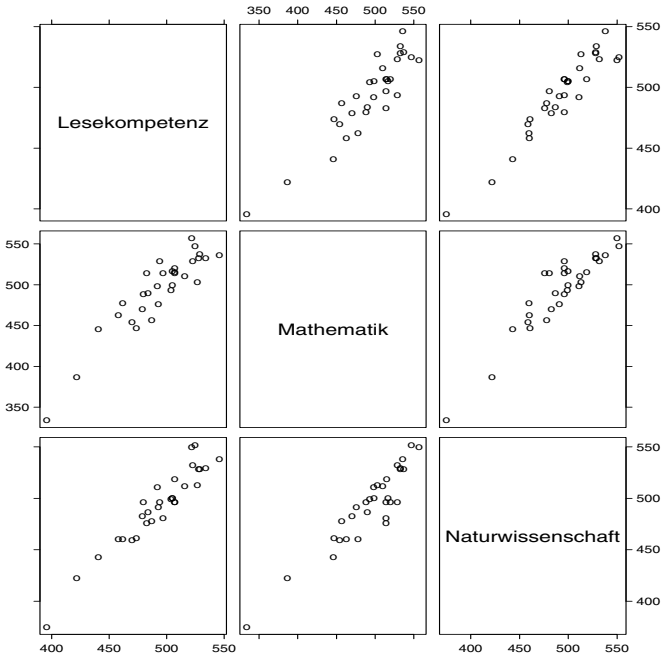

Abb. 2.8. Streudiagrammmatrix der drei Merkmale im Rahmen der PISA-Studie

Bisher haben wir mit Hilfe von Streudiagrammen versucht herauszufinden, welcher Zusammenhang zwischen zwei Merkmalen besteht. Die Objekte, an denen die Merkmale erhoben wurden, waren nicht von Interesse. Mit diesen wollen wir uns nun aber auch beschäftigen.

Beispiel 15. (fortgesetzt) Abbildung 2.9 zeigt das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung`, wobei wir aber an die Koordinaten jedes Landes den Namen des Landes schreiben. Wir sehen nun sehr schön, wo die einzelnen Länder liegen. Will man eine graphische Darstellung hinsichtlich aller drei Merkmale, so könnte man eine dreidimensionale Graphik erstellen. Bei mehr als drei Merkmalen ist eine direkte graphische Darstellung der Objekte hinsichtlich aller Merkmale nicht mehr möglich. Wir werden aber Verfahren kennenlernen, die eine interpretierbare Darstellung von Objekten in einem zweidimensionalen Streudiagramm ermöglichen. □

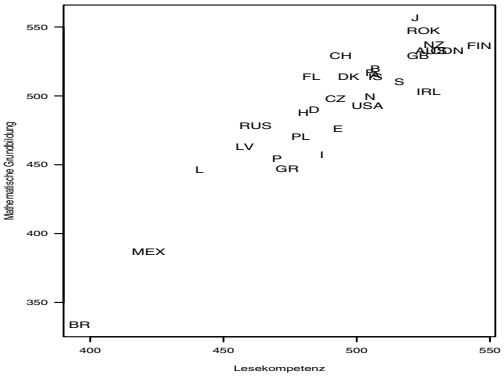

Abb. 2.9. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

2.2.2 Beschreibung und Darstellung von Datenmatrizen qualitativer Merkmale

Beispiel 16. Im Beispiel 2 auf Seite 3 wurde eine Reihe qualitativer Merkmale erhoben. Die Datenmatrix ist in (2.2) auf Seite 14 zu finden. Wir wählen von dieser die Spalten 1, 2 und 4 mit den Merkmalen `Geschlecht`, `MatheLK` und `Abitur88` aus. □

Bei nur einem Merkmal haben wir eine Häufigkeitstabelle erstellt. Dies wird auch der erste Schritt bei mehreren qualitativen Merkmalen sein. Die klassische Form der Darstellung einer (n, p)-Datenmatrix, die nur qualitative Merkmale enthält, ist die *Kontingenztabelle*. Eine Kontingenztabelle ist nichts anderes als eine Häufigkeitstabelle mehrerer qualitativer Merkmale. Schauen wir uns diese zunächst für zwei qualitative Merkmale A und B an. Wir bezeichnen die Merkmalsausprägungen von A mit $A_1, A_2, \ldots, A_I$ und die Merkmalsausprägungen von B mit $B_1, B_2, \ldots, B_J$. Wie im univariaten Fall bestimmen wir absolute Häufigkeiten, wobei wir aber die beiden Merkmale gemeinsam betrachten. Sei n_{ij} die Anzahl der Objekte, die beim Merkmal A die Ausprägung A_i und beim Merkmal B die Ausprägung B_j aufweisen. Tabelle 2.4 zeigt den allgemeinen Aufbau einer zweidimensionalen Kontingenztabelle.

Beispiel 16. (fortgesetzt) Sei A das Merkmal `Geschlecht` und B das Merkmal `MatheLK`. Fassen wir bei beiden Merkmalen die 0 als erste Merkmalsausprägung und die 1 als zweite Merkmalsausprägung auf, so gilt

$$n_{11} = 5, \quad n_{12} = 5, \quad n_{21} = 4, \quad n_{22} = 6.$$

Tabelle 2.5 zeigt die Kontingenztabelle. □

Tabelle 2.4. Allgemeiner Aufbau einer zweidimensionalen Kontingenztabelle

B / A	B_1	B_2	$\dots$	B_J
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}

Tabelle 2.5. Kontingenztabelle der Merkmale Geschlecht und MatheLK

`MatheLK` 0	1
`Geschlecht`	
0	5 5
1	4 6

Die absoluten Häufigkeiten der Merkmalsausprägungen der univariaten Merkmale erhalten wir durch Summierung der Elemente der Zeilen bzw. Spalten. Wir bezeichnen die absolute Häufigkeit der Merkmalsausprägung A_i mit $n_{i.}$ und die absolute Häufigkeit der Merkmalsausprägung B_j mit $n_{.j}$. Es gilt

$$n_{i.} = \sum_{j=1}^{J} n_{ij}$$

und

$$n_{.j} = \sum_{i=1}^{I} n_{ij}.$$

Beispiel 16. (fortgesetzt) Es gilt $n_{1.} = 10$, $n_{2.} = 10$, $n_{.1} = 9$ und $n_{.2} = 11$. □

Es ist von Interesse, ob zwischen den beiden Merkmalen ein Zusammenhang besteht. Hierzu schaut man sich zunächst die *bedingten relativen Häufigkeiten* an. Dies bedeutet, dass man unter der Bedingung, dass die einzelnen Kategorien des Merkmals A gegeben sind, die Verteilung des Merkmals B bestimmt.

Beispiel 16. (fortgesetzt) Wir betrachten zunächst nur die Männer. Von den 10 Männern haben 5 den Mathematik-Leistungskurs besucht, also 50 Prozent. Von den 10 Frauen haben 6 den Mathematik-Leistungskurs besucht, also 60 Prozent. Wir sehen, dass sich diese Häufigkeiten unterscheiden. Es stellt sich die Frage, ob dieser Unterschied signifikant ist. Wir werden diese Frage im Kapitel 10 über loglineare Modelle beantworten. □

Für die bedingte relative Häufigkeit der Merkmalsausprägung B_j unter der Bedingung, dass die Merkmalsausprägung A_i gegeben ist, schreiben wir $h_{j|i}$. Offensichtlich gilt

$$h_{j|i} = \frac{n_{ij}}{n_{i.}}.$$

Den allgemeinen Aufbau einer Tabelle mit bedingten relativen Häufigkeiten zeigt Tabelle 2.6. Die Zeilen dieser Tabelle bezeichnet man auch als *Profile*.

Tabelle 2.6. Allgemeiner Aufbau einer Kontingenztabelle mit bedingten relativen Häufigkeiten

A \ B	B_1	B_2	$\dots$	B_J
A_1	$h_{1\|1}$	$h_{2\|1}$	$\cdots$	$h_{J\|1}$
A_2	$h_{1\|2}$	$h_{2\|2}$	$\cdots$	$h_{J\|2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_I	$h_{1\|I}$	$h_{2\|I}$	$\cdots$	$h_{J\|I}$

Beispiel 16. (fortgesetzt) Für das Beispiel erhalten wir die bedingten relativen Häufigkeiten in Tabelle 2.7. □

Tabelle 2.7. Kontingenztabelle der Merkmale Geschlecht und MatheLK mit bedingten relativen Häufigkeiten

`Geschlecht` \ `MatheLK`	0	1
0	0.5	0.5
1	0.4	0.6

Man kann natürlich auch die Verteilung von A unter der Bedingung bestimmen, dass die einzelnen Kategorien von B gegeben sind. Dies wollen wir aber nicht im Detail ausführen.

Die Kontingenztabelle von zwei qualitativen Merkmalen ist ein Rechteck. Nimmt man ein weiteres Merkmal hinzu, so erhält man einen Quader. Diesen stellt man nun nicht dreidimensional, sondern mit Hilfe von Schnitten zweidimensional dar. Gegeben seien also die qualitativen Merkmale A, B und C mit den Merkmalsausprägungen $A_1, \dots, A_I$, $B_1, \dots, B_J$ und $C_1, \dots, C_K$. Dann ist n_{ijk} die absolute Häufigkeit des gemeinsamen Auftretens von A_i, B_j und C_k, $i = 1, \dots, I$, $j = 1, \dots, J$ und $k = 1, \dots, K$. Tabelle 2.8 beinhaltet den allgemeinen Aufbau einer dreidimensionalen Kontingenztabelle.

Tabelle 2.8. Allgemeiner Aufbau einer dreidimensionalen Kontingenztabelle

		B		
C	A	B_1	$\dots$	B_J
C_1	A_1	n_{111}	$\dots$	n_{1J1}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_I	n_{I11}	$\dots$	n_{IJ1}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_K	A_1	n_{11K}	$\dots$	n_{1JK}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_I	n_{I1K}	$\dots$	n_{IJK}

Beispiel 16. (fortgesetzt) Die dreidimensionale Kontingenztabelle der Merkmale `Geschlecht`, `MatheLK` und `Abitur88` ist in Abbildung 2.9 zu finden. □

Tabelle 2.9. Dreidimensionale Kontingenztabelle der Merkmale Geschlecht, MatheLK und Abitur88

		`MatheLK`	
`Abitur88`	`Geschlecht`	0	1
0	0	5	4
	1	1	3
1	0	0	1
	1	3	3

Aus einer dreidimensionalen Tabelle kann man durch Summation über die Häufigkeiten eines Merkmals drei zweidimensionale Kontingenztabellen erhalten.

Beispiel 16. (fortgesetzt) Wir haben die Kontingenztabelle der Merkmale `Geschlecht` und `MatheLK` bereits erstellt. Sie ist in Tabelle 2.5 auf Seite 37 zu finden. Die beiden anderen Tabellen sind in den Abbildungen 2.10 und 2.11 zu finden. Schaut man sich die entsprechenden bedingten relativen Häufigkeiten an, so sieht es so aus, als ob zwischen den Merkmalen `Geschlecht` und `Abitur88` ein Zusammenhang besteht, während zwischen den Merkmalen `Abitur88` und `MatheLK` kein Zusammenhang zu bestehen scheint. □

Wir haben bisher nur die zweidimensionalen Kontingenztabellen betrachtet, die man aus einer dreidimensionalen Kontingenztabelle gewinnen kann und

Tabelle 2.10. Kontingenztabelle der Merkmale Geschlecht und Abitur88

Abitur88	0	1
Geschlecht		
0	9	1
1	4	6

Tabelle 2.11. Kontingenztabelle der Merkmale Abitur88 und MatheLK

MatheLK	0	1
Abitur88		
0	6	7
1	3	4

deskriptiv auf Zusammenhänge untersucht. In dreidimensionalen Kontingenztabellen können aber noch komplexere Zusammenhänge existieren. Mit diesen werden wir uns detailliert im Kapitel 10 im Zusammenhang mit loglinearen Modellen beschäftigen.

2.3 Datenbehandlung in S-PLUS

2.3.1 Univariate Datenanalyse

Quantitative Merkmale Wir wollen nun lernen, wie man in S-PLUS Daten elementar analysiert. S-PLUS bietet eine interaktive Umgebung, Befehlsmodus genannt, in der man die Daten direkt eingeben und analysieren kann. Durch das Bereitschaftszeichen > wird angezeigt, dass eine Eingabe erwartet wird. Der Befehlsmodus ist ein mächtiger Taschenrechner. Wir können hier die Grundrechenarten Addition, Subtraktion, Multiplikation und Division mit den Operatoren +, -, * und / durchführen:

```
> 3+4
[1] 7
> 3-4
[1] -1
> 3*4
[1] 12
> 3/4
[1] 0.75
```

Zum Potenzieren benutzen wir ^ :

```
> 3^4
[1] 81
```

Man kann aber auch komplizierte Analysen durchführen. Wir wollen die Punkte aller Länder im Bereich Mathematische Grundbildung aus dem Beispiel 14 auf Seite 17 analysieren, die wir hier noch einmal wiedergeben:

```
533 520 334 514 490 536 517 447 529 503 514 457 557
533 547 463 514 446 387 537 499 515 470 454 478 510
529 476 498 488 493 .
```

Die Standarddatenstruktur in S-PLUS ist der Vektor. Ein Vektor ist eine Zusammenfassung von Objekten zu einer endlichen Folge. Einen Vektor erstellt man mit der Funktion c. Diese macht aus einer Folge von Zahlen, die durch Kommata getrennt sind, einen Vektor, dessen Komponenten die einzelnen Zahlen sind. Die Zahlen sind die Argumente der Funktion c. Argumente einer Funktion stehen in runden Klammern hinter dem Funktionsnamen und sind durch Kommata voneinander getrennt. Der Aufruf

```
> c(533,520,334,514,490,536,517,447,529,503,514,457,557,
    533,547,463,514,446,387,537,499,515,470,454,478,510,
    529,476,498,488,493)
```

liefert am Bildschirm folgendes Ergebnis:

```
[1] 533 520 334 514 490 536 517 447 529 503 514 457 557
    533 547 463 514 446 387 537 499 515 470 454 478 510
    529 476 498 488 493
```

Die Elemente des Vektors werden ausgegeben. Am Anfang steht `[1]`. Dies zeigt, dass die erste Zahl gleich der ersten Komponente des Vektors ist. Um mit den Werten weiterhin arbeiten zu können, müssen wir sie in einer Variablen speichern. Dies geschieht mit dem Zuweisungsoperator `<-`, den man durch die Zeichen `<` und `-` erhält. Auf der linken Seite steht der Name der Variablen, der die Werte zugewiesen werden sollen, auf der rechten Seite steht der Aufruf der Funktion `c`. Die Namen von Variablen dürfen beliebig lang sein, dürfen aber nur aus Buchstaben, Ziffern und dem Punkt bestehen, wobei das erste Zeichen ein Buchstabe oder der Punkt sein muss. Beginnt ein Name mit einem Punkt, so dürfen nicht alle folgenden Zeichen Ziffern sein. Hierdurch erzeugt man nämlich eine Zahl. Wir nennen die Variable `Mathe`. `S-PLUS` unterscheidet Groß- und Kleinschreibung. Die Variablennamen `Mathe` und `mathe` beziehen sich also auf unterschiedliche Objekte. Wir geben ein

```
> Mathe<-c(533,520,334,514,490,536,517,447,529,503,514,
           457,557,533,547,463,514,446,387,537,499,515,
           470,454,478,510,529,476,498,488,493)
```

Den Inhalt einer Variablen kann man sich durch Eingabe des Namens anschauen. Der Aufruf

```
> Mathe
```

liefert das Ergebnis

```
 [1] 533 520 334 514 490 536 517 447 529 503 514 457 557
     533 547 463 514 446 387 537 499 515 470 454 478 510
     529 476 498 488 493
```

Man kann gleichlange Vektoren mit Operatoren verknüpfen. Dabei wird der Operator auf die entsprechenden Komponenten der Vektoren angewendet. Man kann aber auch einen Skalar mit einem Vektor über einen Operator verknüpfen. Dabei wird der Skalar mit jeder Komponente des Vektors über den Operator verknüpft. Will man also wissen, wie sich jede Komponente des Vektors `Mathe` von der Zahl 500 unterscheidet, so gibt man ein

```
> Mathe-500
 [1] 33 20 -166 14 -10 36 17 -53 29 3 14 -43 57 33 47 -37
     14 -54 -113 37 -1 15 -30 -46 -22 10 29 -24 -2 -12 -7
```

Auf Komponenten eines Vektors greift man durch Indizierung zu. Hierzu gibt man den Namen des Vektors gefolgt von eckigen Klammern ein, zwischen denen die Nummer der Komponente steht, auf die man zugreifen will. Will man also die Punkte des zweiten Landes wissen, so gibt man ein

```
> Mathe[2]
```

und erhält als Ergebnis

```
[1] 520
```

Will man auf die letzte Komponente zugreifen, so benötigt man die Länge des Vektors. Diese liefert die Funktion `length`:

```
> length(Mathe)
[1] 31
```

Die letzte Komponente des Vektors `Mathe` erhalten wir also durch

```
> Mathe[length(Mathe)]
[1] 493
```

Auf mehrere Komponenten eines Vektors greift man zu, indem man einen Vektor mit den Nummern der Komponenten bildet und mit diesem indiziert. So erhält man die Punkte der ersten drei Länder durch

```
> Mathe[c(1,2,3)]
[1] 533 520 334
```

Wir können auf Komponenten, die hintereinander stehen, einfacher zugreifen. Sind i und j natürliche Zahlen mit $i < j$, so liefert in `S-PLUS` der Ausdruck

```
i:j
```

die Zahlenfolge $i, i+1, \ldots, j-1, j$. Ist $i > j$, so erhalten wir die Zahlenfolge $i, i-1, \ldots, j+1, j$. Wollen wir also auf die ersten drei Komponenten von `Mathe` zugreifen, so geben wir ein

```
> Mathe[1:3]
[1] 533 520 334
```

Wollen wir den Vektor `Mathe` in umgekehrter Reihenfolge ausgeben, so geben wir ein

```
> Mathe[length(Mathe):1]
 [1] 493 488 498 476 529 510 478 454 470 515 499 537 387
     446 514 463 547 533 557 457 514 503 529 447 517 536
     490 514 334 520 533
```

Mit der Funktion `rev` hätten wir das gleiche Ergebnis erhalten.

Oft will man Komponenten eines Vektors selektieren, die bestimmte Eigenschaften besitzen. Hierzu benötigt man Vergleichsoperatoren, mit denen man auf Gleichheit mit `==`, Ungleichheit mit `!=`, kleiner mit `<`, kleiner gleich mit `<=`, größer mit `>` oder größer gleich mit `>=` überprüfen kann. Das Ergebnis des Vergleichs ist vom Typ `logical`, ist also entweder `T` oder `F`, wobei `T` für true und `F` für false steht:

```
> 3<4
[1] T
```

Man kann natürlich auch einen Vektor der Länge n und einen Skalar mit einem Vergleichsoperator verknüpfen:

```
> 1:5 <= 3
[1] T T T F F
```

Indiziert man einen Vektor der Länge n mit einem Vektor vom Typ `logical` der Länge n, so werden die Komponenten ausgewählt, bei denen im Vektor vom Typ `logical` ein `T` steht. Wollen wir die Punktezahlen der Länder wissen, die weniger als 480 Punkte erreicht haben, so geben wir ein

```
> Mathe[Mathe<480]
 [1] 334 447 457 463 446 387 470 454 478 476
```

Die Nummern der Länder erhalten wir durch

```
> (1:length(Mathe))[Mathe<480]
 [1]  3  8 12 16 18 19 23 24 25 28
```

Die Funktion `sum` bestimmt die Summe der Komponenten eines Vektors. Sind diese vom Typ `logical`, so wird `F` in 0 und `T` in 1 umgewandelt. Der Aufruf

```
> sum(Mathe<480)
[1] 10
```

liefert also die Anzahl der Länder mit weniger als 480 Punkten. Sollen mehrere Bedingungen erfüllt sein, so kann man die logischen Operatoren `&` und `|` verwenden. Der Operator `&` entspricht dem logischen `"und"` und der Operator `|` entspricht dem logischen `"oder"`. Die Indizes der Länder, die mindestens 490 und höchstens 510 Punkte erreicht haben, erhalten wir durch

```
> (1:length(Mathe))[Mathe>=490 & Mathe<=510]
[1]  5 10 21 26 29 31
```

In `S-PLUS` gibt es eine Vielzahl von Funktionen. Von diesen haben wir die Funktionen `c`, `sum`, `length` und `rev` kennengelernt. Mit den Funktionen `sum` und `length` können wir den Mittelwert folgendermaßen bestimmen:

```
> sum(Mathe)/length(Mathe)
[1] 493.1613
```

In `S-PLUS` gibt es zur Bestimmung des Mittelwerts die Funktion `mean`. Für die Variable `Mathe` erhalten wir

```
> mean(Mathe)
[1] 493.1613
```

Mit der Funktion `mean` kann man aber nicht nur den Mittelwert bestimmen.

Schauen wir uns die Funktion an:

```
> mean
 function(x, trim = 0, na.rm = F) {
     if(na.rm) {
         wnas <- which.na(x)
         if(length(wnas))
             x <- x[ - wnas]
     }
     if(mode(x) == "complex") {
         if(trim > 0)
             stop("trimming not allowed for complex data")
         return(sum(x)/length(x))
     }
     x <- as.double(x)
     if(trim > 0) {
         if(trim >= 0.5)
             return(median(x, na.rm = F))
         if(!na.rm && length(which.na(x)))
             return(NA)
         n <- length(x)
         i1 <- floor(trim * n) + 1
         i2 <- n - i1 + 1
         x <- sort(x, unique(c(i1, i2)))[i1:i2]
     }
     sum(x)/length(x)
}
```

Wie jede Funktion in S-PLUS besteht `mean` aus einem Kopf und einem Körper.

Der Funktionskopf besteht aus dem Namen der Funktion gefolgt von den Argumenten der Funktion, die in runden Klammern stehen und durch Kommata getrennt sind. Die Funktion `mean` hat die drei Argumente `x`, `trim` und `na.rm`. Die Argumente `trim` und `na.rm` sind in dem Sinne fakultativ, dass ihnen beim Aufruf der Funktion vom Benutzer keine Werte zugewiesen werden müssen. Sie sind vorbelegt durch `0` im Falle von `trim` und von `F` im Falle von `na.rm`. Das Argument `x` hingegen muss der Funktion `mean` übergeben werden. Hierzu haben wir zwei Möglichkeiten. Wir können die Funktion aufrufen mit

```
> mean(x=Mathe)
```

Hierdurch werden der lokalen Variablen `x` beim Aufruf von `mean` die Werte der Variablen `Mathe` zugewiesen. Wie wir weiter oben gesehen haben, können wir aber auch eingeben

```
> mean(Mathe)
```

Hierbei müssen die Argumente nur an der richtigen Position stehen. Da das erste Argument der Funktion `mean` der Datenvektor `x` ist, werden die Werte von `Mathe` für diesen eingesetzt.

Der Funktionskörper besteht aus den Anweisungen. Die Anweisungen stehen in geschweiften Klammern. Wir sehen, dass die Funktion `mean` aus einer Reihe von Anweisungen besteht. Wir wollen diese hier nicht alle diskutieren, sondern nur bemerken, dass durch das Argument `trim` ein getrimmter Mittelwert bestimmt werden kann und dass mit Hilfe des Arguments `na.rm` gesteuert werden kann, was mit fehlenden Beobachtungen bei der Berechnung des Mittelwerts geschehen soll. Für jede fehlende Beobachtung gibt man den Wert `NA` ein. Liegen keine fehlenden Beobachtungen vor und soll auch nicht getrimmt werden, so wird nur der letzte Befehl der Funktion `mean` ausgeführt:

```
sum(x)/length(x)
```

Diesen kennen wir bereits. Da `S-PLUS` das Ergebnis des letzten Ausdrucks einer Funktion als Ergebnis der Funktion zurückgibt, liefert die Funktion `mean` den Mittelwert als Ergebnis.

Wir haben schon erwähnt, dass man mit der Funktion `mean` auch getrimmte Mittelwerte bestimmen kann. Man muss in diesem Fall nur dem Argument `trim` den gewünschten Trimmanteil α zuweisen. Der Aufruf

```
> mean(Mathe,0.05)
```

liefert das Ergebnis

```
[1] 496.4483
```

Man kann mit der Funktion `mean` auch den Median bestimmen. Man muss nur das Argument `trim` auf 0.5 setzen:

```
> mean(Mathe,0.5)
[1] 503
```

Schaut man sich den Inhalt der Funktion `mean` an, so sieht man, dass in diesem Fall die Funktion `median` aufgerufen wird. Wir können den Median also direkt bestimmen durch

```
> median(Mathe)
[1] 503
```

Schauen wir uns die Stelle in der Funktion `mean` an, an der die Funktion `median` aufgerufen wird. Sie lautet

```
if(trim >= 0.5)
    return(median(x, na.rm = F))
```

Hierbei handelt es sich um eine bedingte Anweisung. Es wird die Bedingung `trim>=0.5` überprüft. Ist das Argument von `if` gleich `T`, so wird die Anweisungsfolge ausgeführt, die hinter dem Ausdruck `trim>=0.5` steht. Dabei besteht eine Anweisungsfolge in `S-PLUS` aus einer Folge von Anweisungen,

die von geschweiften Klammern umgeben sind. Liegt nur ein Befehl vor, so kann man auf die Klammern verzichten. Ist das Argument von `if` gleich `F`, so wird die Anweisungsfolge übersprungen, die hinter dem Ausdruck `trim>=0.5` steht, und der hinter dieser Befehlsfolge stehende Befehl wird ausgeführt. Ist `trim` also größer oder gleich 0.5, so wird der Befehl `return(median(x, na.rm = F))` ausgeführt. Es wird der Median berechnet und als Ergebnis der Funktion `mean` zurückgegeben. Der Ausdruck `return(x)` bewirkt, dass die Ausführung einer Funktion beendet wird, und `x` als Ergebnis der Funktion zurückgegeben wird.

Kehren wir zu den Funktionen zurück, mit denen man Daten analysieren kann. Die Varianz einer Variablen erhält man mit der Funktion `var`:

```
> var(Mathe)
[1] 2192.873
```

Für die Standardabweichung gibt es keine eigene Funktion in `S-PLUS`. Man kann sich aber eine eigene Funktion schreiben. Die Standardabweichung ist die Wurzel aus der Varianz. Die Funktion `sqrt` bestimmt die Wurzel. Wir erhalten die Standardabweichung also durch

```
> sqrt(var(Mathe))
[1] 46.82812
```

Wir wollen nun eine Funktion `std` schreiben, die die Standardabweichung der Elemente eines Objekts `x` bestimmt. Eine Funktion wird durch folgende Befehlsfolge deklariert:

```
fname<-function(Argumente)
{
   Koerper der Funktion
   return(Ergebnis)
}
```

Wir geben also ein

```
std<-function(x)
{
   return(sqrt(var(x)))
}
```

Wir können die Funktion sofort benutzen:

```
> std(Mathe)
[1] 46.82812
```

Der Zweck der Funktion `std` ist ersichtlich, aber auch hier verbessert die Verwendung von Kommentaren die Lesbarkeit.

Hier ist die kommentierte Version von `std`:

```
std<-function(x)
{
# Standardabweichung der Elemente von x
   return(sqrt(var(x)))
}
```

Bei der Beschreibung eines univariaten Merkmals haben wir auch die Fünf-Zahlen-Zusammenfassung betrachtet. Die Funktion `summary` bestimmt das Minimum $x_{(1)}$, das untere Quartil $x_{0.25}$, den Median $x_{0.5}$, das obere Quartil $x_{0.75}$ und das Maximum $x_{(n)}$. Der Aufruf

```
> summary(Mathe)
```

liefert das Ergebnis

```
Min. 1st Qu. Median  Mean 3rd Qu. Max.
 334     473    503 493.2   524.5  557
```

Wir sehen, dass neben den 5 Zahlen auch noch der Mittelwert bestimmt wird. Die Zahlen stimmen mit denen überein, die wir weiter oben bestimmt haben. `S-PLUS` liefert aber nicht für jeden Stichprobenumfang die Quartile so, wie es auf Seite 19 beschrieben wird. Der Aufruf

```
> summary(1:6)
```

liefert das Ergebnis

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
   1    2.25    3.5  3.5    4.75    6
```

Bei Tukey nimmt das untere Quartil den Wert 2 an. Hyndman & Fan (1996) geben an, wie `S-PLUS` die Quartile bestimmt. Wir wollen hierauf aber nicht eingehen. Im Anhang ist auf Seite 465 eine Funktion `quartile` zu finden, die die Quartile so bestimmt, wie es auf Seite 19 beschrieben wird.:

```
> quartile(1:6)
[1] 2 5
```

Mit Hilfe der 5 Zahlen kann man einen Boxplot erstellen. Der Aufruf

```
> boxplot(Mathe)
```

liefert die Abbildung 2.10. Der Boxplot sieht nicht so aus wie in Abbildung 2.3 auf Seite 20. Die Beschriftung der Ordinate unterscheidet sich in beiden Abbildungen. Wir sind eine Beschriftung wie in Abbildung 2.3 gewohnt. Diese erreichen wir, indem wir den Graphikparameter `las` auf den Wert 1 setzen. Damit der Boxplot wie in Abbildung 2.3 aussieht, müssen wir einige Argumente der Funktion `boxplot` mit speziellen Werten aufrufen. Der folgende Aufruf liefert den Boxplot in Abbildung 2.3:

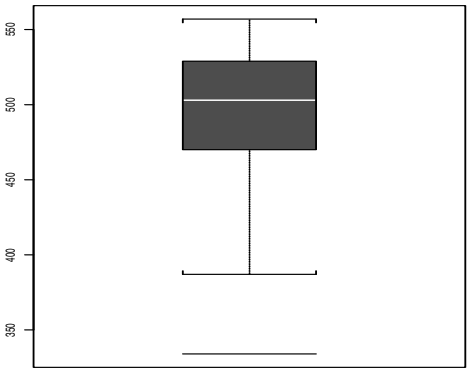

Abb. 2.10. Boxplot des Merkmals Mathematische Grundbildung

```
> par(las=1)
> boxplot(Mathe,names="Mathematische Grundbildung",boxcol=0,
   medline=T,medcol=1,outline=F,outpch="*",medlwd=0.5,col=1)
```

Das Argument `names` ist eine Zeichenkette. Eine Zeichenkette ist eine Folge von Zeichen, die in Hochkommata stehen. Wir werden uns gleich mit Zeichenketten beschäftigen. Schauen wir uns vorher die Befehlsfolge an, die das Histogramm in Abbildung 2.2 auf Seite 18 liefert:

```
>  hist(Mathe,prob=T,xlab="Mathematische Grundbildung")
```

Durch `prob=T` stellen wir sicher, dass die Fläche unter dem Histogramm gleich 1 ist. Setzen wir `prob` auf `F`, so haben die Rechtecke die Höhe der absoluten Häufigkeiten. `S-PLUS` wählt standardmäßig gleich große Klassen. Die Anzahl der Klassen ist proportional zu $\ln n$.

Qualitative Merkmale Wir wollen die Analyse des Merkmals `MatheLK` aus Beispiel 13 auf Seite 15 in `S-PLUS` nachvollziehen. Das Merkmal `MatheLK` kann die Werte `j` und `n` annehmen. Hier sind noch einmal die Werte der 20 Studenten:

```
n n n n n n n n n j j j j j j j j j j
```

Wir wollen diese Werte der Variablen `MatheLK` zuweisen. Hierzu erzeugen wir uns einen Vektor der Länge 20, dessen Komponenten Zeichenketten sind. Die ersten 9 Komponenten sollen die Zeichenkette `"n"` und die letzten 11 Komponenten die Zeichenkette `"j"` enthalten. Um uns die Eingabe zu erleichtern, verwenden wir die Funktion `rep`. Der Aufruf

```
rep(x,times)
```

erzeugt einen Vektor, in dem das Argument `x` `times`-mal wiederholt wird:

```
> rep("n",9)
[1] "n" "n" "n" "n" "n" "n" "n" "n" "n"
```

Wir erzeugen den Vektor `MatheLK` also durch

```
> MatheLK<-c(rep("n",9),rep("j",11))
```

Schauen wir uns `MatheLK` an:

```
> MatheLK
 [1] "n" "j" "n" "n" "n" "n" "n" "n" "n" "n"
     "j" "j" "j" "j" "j" "j" "j" "j" "j" "j"
```

Das Merkmal `MatheLK` ist nominalskaliert. Ein nominalskaliertes qualitatives Merkmal ist in `S-PLUS` ein *Faktor*. Ein Faktor wird erzeugt mit der Funktion `factor`:

```
> MatheLK<-factor(MatheLK)
> MatheLK
 [1] n n n n n n n n n j j j j j j j j j j
```

Ein ordinalskaliertes qualitatives Merkmal ist in `S-PLUS` ein *geordneter Faktor*. Diesen erzeugt man mit der Funktion `ordered`. Die absoluten Häufigkeiten der Merkmalsausprägungen erhalten wir mit der Funktion `table`. Der Aufruf

```
> table(MatheLK)
```

liefert das Ergebnis

```
 j n
11 9
```

Die relativen Häufigkeiten erhalten wir, indem wir das Ergebnis der Funktion `table` durch die Anzahl der Beobachtungen teilen:

```
> table(MatheLK)/length(MatheLK)
    j    n
0.55 0.45
```

Mit der Funktion `barplot` erstellen wir das Stabdiagramm. Der Aufruf

```
> barplot(table(MatheLK)/length(MatheLK))
```

liefert die Abbildung 2.11. Wir sehen, dass bei diesem Stabdiagramm im Gegensatz zum Stabdiagramm in Abbildung 2.1 auf Seite 16 die Stäbe sehr breit sind. Außerdem fehlt die Achsenbeschriftung. Die Breite der Stäbe wird mit dem Parameter `space` festgelegt. Dieser gibt das Verhältnis aus dem Zwischenraum zwischen den Balken zur Breite der Balken an. Setzt man im Beispiel das Argument `space` auf den Wert 20, so erhält man die Balkenbreite in Abbildung 2.1. Die Achsenbeschriftung erhalten wir, indem wir dem Parameter `names` die Merkmalsausprägungen als Zeichenkettenvektor übergeben.

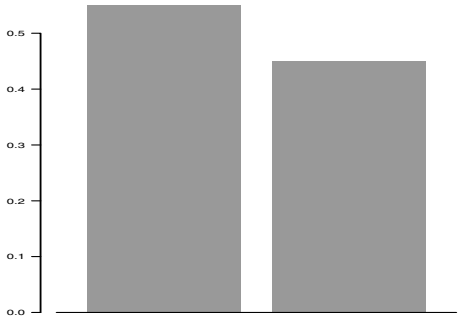

Abb. 2.11. Stabdiagramm des Merkmals MatheLK mit breiten Balken und ohne Achsenbeschriftung

2.3.2 Multivariate Datenanalyse

Quantitative Merkmale Nun wollen wir die Merkmale `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` aus dem Beispiel 15 auf Seite 22 gemeinsam analysieren. Hierzu geben wir die Daten in Form einer Matrix ein. In `S-PLUS` erzeugt man eine Matrix mit der Funktion `matrix`. Der Aufruf von `matrix` ist

```
matrix(data,nrow=1,ncol=1,byrow=F)
```

Dabei ist `data` der Vektor mit den Elementen der Matrix. Das Argument `nrow` gibt die Anzahl der Zeilen und das Argument `ncol` die Anzahl der Spalten der Matrix an. Standardmäßig wird eine Matrix spaltenweise eingegeben. Sollen die Zeilen aufgefüllt werden, so muss das Argument `byrow` auf den Wert `T` gesetzt werden. Wir weisen die Punkte der 31 Länder der Matrix `PISA` zu, wobei wir hier die Daten verkürzt wiedergeben. Die drei Punkte stehen für die restlichen 87 Beobachtungen:

```
> PISA<-matrix(c(528,507,396,...,511,496,499),31,3)
```

Wir wollen nun noch den Zeilen und Spalten der Matrix `PISA` Namen geben. Dies geschieht mit der Funktion `dimnames`. Der Aufruf von `dimnames` für eine Matrix `mat` ist

```
> dimnames(mat)<-list(ZN,SN)
```

Dabei sind `ZN` und `SN` Vektoren mit den Namen der Zeilen beziehungsweise Spalten der Matrix `mat`. In der Regel werden dies Vektoren sein, die Zeichenketten enthalten. Die Funktion `list` verbindet ihre Argumente zu einer Liste. Eine Liste besteht aus Komponenten, die unterschiedliche `S-PLUS`-Objekte sein können. In einer Liste kann man zum Beispiel Vektoren und Matrizen

zu einem Objekt zusammenfassen. Schauen wir uns dies für das Beispiel an. Wir erzeugen zunächst einen Vektor `laender` mit den Namen der Länder, wobei wir die Ländernamen durch die Autokennzeichen abkürzen:

```
> laender<-c("AUS","B","BR","DK","D","FIN","F","GR","GB",
             "IRL","IS","I","J","CDN","ROK","LV","FL","L",
             "MEX","NZ","N","A","PL","P","RUS","S","CH",
             "E","CZ","H","USA")
```

Dann erzeugen wir einen Vektor `bereiche` mit den drei Bereichen:

```
> bereiche<-c("Lesekompetenz","Mathematik",
              "Naturwissenschaft")
```

Wir weisen diese beiden Vektoren einer Liste mit Namen `namen.PISA` zu:

```
> namen.PISA<-list(laender,bereiche)
```

Schauen wir uns `namen.PISA` an:

```
> namen.PISA
[[1]]:
 [1] "AUS" "B"   "BR"  "DK"  "D"   "FIN" "F"   "GR"
     "GB"  "IRL" "IS"  "I"   "J"   "CDN" "ROK" "LV"
     "FL"  "L"   "MEX" "NZ"  "N"   "A"   "PL"
     "P"   "RUS" "S"   "CH"  "E"   "CZ"  "H"   "USA" [[2]]:
  [1] "Lesekompetenz" "Mathematik" "Naturwissenschaft"
```

Auf Komponenten einer Liste greift man mit doppelten eckigen Klammern zu:

```
> namen.PISA[[2]]
[1] "Lesekompetenz" "Mathematik" "Naturwissenschaft"
```

Nun geben wir den Zeilen und Spalten von `pisa` Namen:

```
> dimnames(PISA)<-namen.PISA
```

Einzelne Elemente der Matrix erhält man durch Indizierung, wobei man die Nummer der Zeile und die Nummer der Spalte in eckigen Klammern durch Komma getrennt eingeben muss. Die Punktezahl von Deutschland im Bereich Mathematik erhält man also durch

```
> PISA[5,2]
[1] 490
```

Die Punkte von Deutschland in allen Bereichen erhält man durch

```
> PISA[5,]
 Lesekompetenz Mathematik Naturwissenschaft
           484        490               487
```

Wendet man die Funktion `mean` auf eine Matrix an, so wird der Mittelwert aller Elemente dieser Matrix bestimmt:

```
> mean(PISA)
[1] 493.0753
```

Dieser interessiert aber in der Regel wenig, da man die einzelnen Variablen getrennt analysieren will. Will man die Mittelwerte aller Spalten einer Matrix bestimmen, so muss man die Funktion `apply` aufrufen. Der allgemeine Aufruf von `apply` ist

```
apply(X, MARGIN, FUN)
```

Dabei sind `X` die Matrix und `MARGIN` die Dimension der Matrix, bezüglich der die Funktion angewendet werden soll. Dabei steht 1 für die Zeilen und 2 für die Spalten. Das Argument `FUN` ist der Name der Funktion, die auf `MARGIN` von `X` angewendet werden soll. Der Aufruf `apply(PISA,1,mean)` bestimmt den Vektor der Mittelwerte der Zeilen der Datenmatrix `PISA` und der Aufruf `apply(PISA,2,mean)` bestimmt den Vektor der Mittelwerte der Spalten der Datenmatrix `PISA`. So sind die mittleren Punktezahlen in den Bereichen:

```
> apply(PISA,2,mean)
 Lesekompetenz Mathematik Naturwissenschaft
      493.4516   493.1613          492.6129
```

Die zentrierte Datenmatrix kann man auf drei Arten erhalten. Man kann die Funktion `scale` anwenden, die neben der Datenmatrix `m` noch die beiden Argumente `center` und `scale` besitzt. Diese sind standardmäßig auf `T` gesetzt. Ruft man die Funktion `scale` nur mit der Datenmatrix als Argument auf, so liefert diese die Matrix der standardisierten Variablen. Von jedem Wert jeder Variablen wird der Mittelwert subtrahiert und anschließend durch die Standardabweichung der Variablen dividiert. Setzt man das Argument `scale` auf `F`, so erhält man die Matrix der zentrierten Variablen. Der Aufruf

```
> scale(PISA,scale=F)
```

liefert also die zentrierte Datenmatrix. Man kann aber auch die Funktion `sweep` aufrufen. Der Aufruf von `sweep` für eine Matrix ist

```
sweep(M, MARGIN, STATS, FUN)
```

Dabei sind `M` die Matrix und `MARGIN` die Dimension der Matrix, bezüglich der die Funktion angewendet werden soll. Dabei steht 1 für die Zeilen und 2 für die Spalten. Das Argument `STATS` ist ein Vektor, dessen Länge der Größe der Dimension entspricht, die im Argument `MARGIN` gewählt wurde, und das

Argument FUN ist der Name der Funktion, die auf MARGIN von M angewendet werden soll. Standardmäßig wird die Subtraktion gewählt. Die Funktion sweep bewirkt, dass die Funktion FUN angewendet wird, um die Komponenten des Vektors aus der gewählten Dimension von M im wahrsten Sinne des Wortes herauszufegen. Stehen zum Beispiel in STATS die Mittelwerte der Spalten von M, und ist FUN gleich "-", so liefert der Aufruf

```
> sweep(M,2,STATS,FUN="-")
```

die zentrierte Datenmatrix. Die Komponenten von STATS können wir mit Hilfe von apply bestimmen, sodass der folgende Aufruf für das Beispiel die Matrix der zentrierten Variablen liefert:

```
> sweep(PISA,2,apply(PISA,2,mean),FUN="-")
```

Man kann die zentrierte Datenmatrix aber auch mit der Gleichung (2.10) auf Seite 23 gewinnen. Die Matrix $\mathbf{M}$ liefert folgender Ausdruck:

```
> n<-dim(PISA)[1]
> M<-diag(n)-outer(rep(1,n),rep(1,n))/n
```

Die Funktion outer wird auf Seite 460 beschrieben. Die zentrierte Datenmatrix erhalten wir durch

```
> M%*%PISA
```

Um die Stichprobenvarianzen der drei Variablen zu bestimmen, benutzen wir wiederum die Funktion apply:

```
> apply(PISA,2,var)
 Lesekompetenz Mathematik Naturwissenschaft
      1109.389   2192.873          1418.978
```

Um ein Streudiagramm zu erstellen, verwendet man in S-PLUS die Funktion plot. Übergibt man dieser als Argumente zwei gleich lange Vektoren, so erstellt sie ein Streudiagramm, wobei die Komponenten des ersten Vektors der Abszisse und die des zweiten Vektors der Ordinate zugeordnet werden. Um das Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung zu erstellen, geben wir also ein

```
> plot(PISA[,1],PISA[,2])
```

Wir erhalten die Abbildung 2.12. Die Graphik kann man nun noch verbessern. Die Achsen können noch geeignet beschriftet werden durch die Argumente xlab und ylab. Die Beschriftung wird der Funktion plot als Argument in Form einer Zeichenkette übergeben. Die folgende Befehlsfolge erzeugt die Abbildung 2.4:

```
> plot(PISA[,1],PISA[,2],xlab="Lesekompetenz",
       ylab="Mathematische Grundbildung")
```

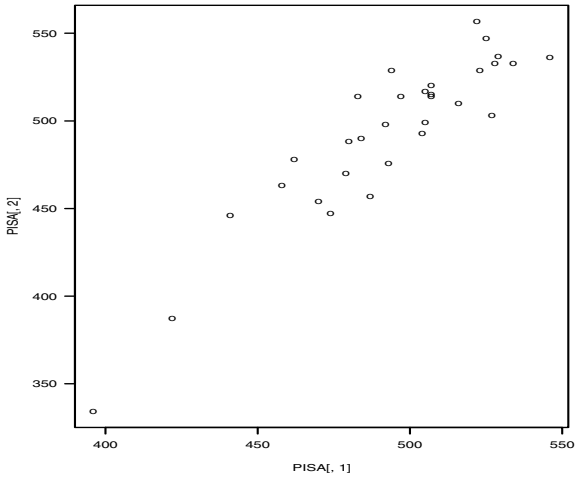

Abb. 2.12. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

In Abbildung 2.9 haben wir die Punkte im Streudiagramm mit Kürzeln der Ländernamen versehen. Um dies zu erreichen, weisen wir beim Aufruf der Funktion `plot` dem Argument `type` den Wert `"n"` zu. In diesem Fall werden keine Punkte gezeichnet:

```
> plot(PISA[,1],PISA[,2],xlab="Lesekompetenz",
       ylab="Mathematische Grundbildung",type="n")
```

Nun müssen wir nur noch mit der Funktion `text` die Namen an den entsprechenden Stellen hinzufügen:

```
> text(PISA[,1],PISA[,2],laender)
```

Wendet man die Funktion `var` auf eine Datenmatrix an, so erhält man die empirische Varianz-Kovarianz-Matrix:

```
> var(PISA)
                 Lesekompetenz Mathematik Naturwissenschaft
    Lesekompetenz      1109.389   1428.325          1195.614
       Mathematik      1428.325   2192.873          1644.031
Naturwissenschaft      1195.614   1644.031          1418.978
```

Um die Struktur besser erkennen zu können, runden wir mit der Funktion `round` auf eine Stelle nach dem Komma:

```
> round(var(PISA),1)
                 Lesekompetenz Mathematik Naturwissenschaft
    Lesekompetenz        1109.4     1428.3            1195.6
       Mathematik        1428.3     2192.9            1644.0
Naturwissenschaft        1195.6     1644.0            1419.0
```

In S-PLUS bestimmen wir die empirische Korrelationsmatrix mit der Funktion `cor`:

```
> cor(PISA)
                 Lesekompetenz Mathematik Naturwissenschaft
    Lesekompetenz     1.0000000  0.9157527         0.9529302
       Mathematik     0.9157527  1.0000000         0.9319989
Naturwissenschaft     0.9529302  0.9319989         1.0000000
```

Eine Streudiagrammmatrix liefert die Funktion `pairs`. Um die Abbildung 2.8 zu erhalten, geben wir ein

```
> pairs(PISA)
```

Nun fehlt uns aus dem Bereich der quantitativen Merkmale noch die konvexe Hülle. Die Indizes der Länder auf der konvexen Hülle aller Beobachtungen erhält man mit der Funktion `chull`. Der allgemeine Aufruf von `chull` ist:

```
chull(x, y, peel=F, maxpeel=<<see below>>, onbdy=peel,
      tol=.0001)
```

Die drei letzten Argumente sind für uns im Folgenden nicht wichtig. Schauen wir uns die anderen an. Das Argument `x` ist ein Vektor mit den ersten Koordinaten der Punkte und das Argument `y` ein Vektor mit den zweiten Koordinaten der Punkte. Das Argument `peel` ist eine logische Variable, über die gesteuert wird, ob eine Folge konvexer Hüllen erzeugt werden soll. Wenn das Argument `peel` gleich `F` ist, erhält man als Ergebnis einen Vektor mit den Indizes der Punkte auf der konvexen Hülle. Um diesen für das Beispiel zu erhalten, geben wir also ein

```
> chull(PISA[,1],PISA[,2])
```

und erhalten das Ergebnis

```
[1] 10  3 18 17 27 13  6
```

Der folgende Befehl liefert den 0.23-getrimmten Mittelwert:

```
> apply(PISA[-chull(PISA[,1],PISA[,2]),1:2],2,mean)
 Lesekompetenz Mathematik
      495.3333   494.5417
```

Um Abbildung 2.5 auf Seite 27 zu erhalten, benötigen wir die Funktion `polygon`. Der Aufruf

```
> polygon(x,y)
```

überlagert eine Graphik mit einem Polygon mit den Eckpunkten `(x,y)`. Wir geben also ein

```
> plot(PISA[,1],PISA[,2],xlab="Lesekompetenz",
       ylab="Mathematische Grundbildung")
> hull <- chull(PISA[,1],PISA[,2])
> polygon(PISA[hull,1],PISA[hull,2],density=0)
```

und erhalten Abbildung 2.5 auf Seite 27. Um den auf der konvexen Hülle beruhenden Median zu erhalten, setzen wir das Argument `peel` der Funktion `chull` auf `T`:

```
> p <- chull(PISA[,1],PISA[,2],peel=T)
```

und erhalten folgendes Ergebnis:

```
> p
 $depth:
 [1] 3 4 1 3 5 1 5 3 4 1 5 2 1 2 2 3
     1 1 2 3 5 5 5 4 2 3 1 4 6 4 4
 $hull:
 [1] 10 3 18 17 27 13 6 19 25 15 14 12 8 16 4 20
     1 26 28 24 31 2 9 30 23 5 7 22 11 21 29
 $count:
 [1] 7 5 6 6 6 1
```

Das Ergebnis ist eine Liste. Die erste Komponente gibt für jeden Punkt die Nummer der konvexen Hülle an, auf der er liegt. Dabei werden die Hüllen von außen nach innen nummeriert. Die zweite Komponente gibt die Indizes der Punkte auf den einzelnen Hüllen an. Die dritte Komponente gibt die Anzahl der Punkte auf jeder Hülle an. Um den Median zu bestimmen, benötigen wir nur die erste Komponente. Wir bestimmen die Punkte, die auf der Hülle mit der höchsten Nummer liegen:

```
> m<-PISA[p[[1]]==max(p[[1]]),1:2]
> m
 Lesekompetenz Mathematik
           492         498
```

Da es sich um einen Punkt handelt, haben wir den Median bereits gefunden. Bei mehr als einem Punkt bestimmen wir den Mittelwert dieser Punkte mit der Funktion `apply`.

Bisher haben wir Vektoren, Listen und Matrizen betrachtet. Von diesen bieten Listen die Möglichkeit, Variablen unterschiedlichen Typs in einem Objekt zu speichern. Die Elemente einer Matrix müssen vom gleichen Typ sein. In `S-PLUS` ist es aber auch möglich, Variablen unterschiedlichen Typs in einem Objekt zu speichern, auf das wie auf eine Matrix zugegriffen werden kann. Diese heißen *Dataframes*. Schauen wir uns exemplarisch die ersten 10 Beobachtungen der Daten in Tabelle 1.2 auf Seite 5 an. Wir erzeugen zunächst die

5 Variablen. Da viele Werte mehrfach hintereinander vorkommen, verwenden wir die Funktion `rep`:

```
> Geschlecht<-c(rep("m",5),rep("w",4),"m")
> Geschlecht<-factor(Geschlecht)
> MatheLK<-c(rep("n",9),"j")
> MatheLK<-factor(MatheLK)
> MatheNote<-c(3,4,4,4,3,3,4,3,4,3)
> MatheNote<-ordered(MatheNote)
> Abitur88<-c(rep("n",6),rep("j",3),"n")
> Abitur88<-factor(Abitur88)
> Punkte<-c(8,7,4,2,7,6,3,7,14,19)
```

Mit der Funktion `data.frame` macht man aus diesen Variablen einen Dataframe:

```
> test<-data.frame(Geschlecht,MatheLK,MatheNote,
                   Abitur88,Punkte)
   Geschlecht MatheLK MatheNote Abitur88 Punkte
1           m       n         3        n      8
2           m       n         4        n      7
3           m       n         4        n      4
4           m       n         4        n      2
5           m       n         3        n      7
6           w       n         3        n      6
7           w       n         4        j      3
8           w       n         3        j      7
9           w       n         4        j     14
10          m       j         3        n     19
```

Die Werte der Variablen `Geschlecht` erhalten wir durch

```
> test[,1]
 [1] m m m m m w w w w m
```

oder durch

```
> test[[1]]
 [1] m m m m m w w w w m
```

Schauen wir uns noch die Beschreibung qualitativer Merkmale an. Wir betrachten wiederum die Merkmale `Geschlecht`, `MatheLK` und `Abitur88` des Beispiels 2 bei den ersten 10 Studenten. Wir bilden eine Matrix `qual` mit den Merkmalen

```
> qual<-test[,c(1,2,4)]
```

Schauen wir uns `qual` an:

```
> qual
   Geschlecht MatheLK Abitur88
 1          m       n        n
 2          m       n        n
 3          m       n        n
 4          m       n        n
 5          m       n        n
 6          w       n        n
 7          w       n        j
 8          w       n        j
 9          w       n        j
10          m       j        n
```

Mit Hilfe der Funktion `table` erstellen wir die zweidimensionale Kontingenztabelle der Merkmale `Geschlecht` und `MatheLK`:

```
> table(qual[,1],qual[,2])
   j n
 m 1 5
 w 0 4
```

In den Zeilen stehen die Ausprägungen des Merkmals `Geschlecht` und in den Spalten die Ausprägungen des Merkmals `MatheLK`. Die Matrix der bedingten relativen Häufigkeiten bestimmen wir mit der Funktion `sweep`. Hierzu weisen wir den obigen Aufruf einer Variablen zu:

```
> e<-table(qual[,1],qual[,2])
```

und rufen dann `sweep` auf:

```
> sweep(e,1,apply(e,1,sum),"/")
           j         n
 m 0.1666667 0.8333333
 w 0.0000000 1.0000000
```

Eine dreidimensionale Kontingenztabelle liefert der Aufruf von `table` mit drei Argumenten. Die dreidimensionale Tabelle der Merkmale `Geschlecht`, `MatheLK` und `Abitur88` erhalten wir durch

```
> e<-table(qual[,1],qual[,2],qual[,3])
> e
 , , j
   j n
 m 0 0
 w 0 3
 , , n
   j n
```

```
m 1 5
w 0 1
```

Die zweidimensionale Tabelle der Merkmale `Geschlecht` und `MatheLK` erhalten wir durch Anwenden von `apply` auf `e` mit der Funktion `sum`:

```
> apply(e,c(1,2),sum)
   j n
m 1 5
w 0 4
```

Entsprechend erhält man die beiden anderen zweidimensionalen Tabellen. Oft liegen die Daten in Form einer dreidimensionalen Kontingenztabelle vor. Dies ist im Beispiel 9 der Fall. Man kann diese in `S-PLUS` mit der Funktion `array` eingeben. Diese wird folgendermaßen aufgerufen:

```
array(data = NA, dim, dimnames = NULL)
```

Dabei ist `data` der Vektor mit den Daten, `dim` ein Vektor mit den Dimensionsangaben und `dimnames` eine Liste mit Namen der Dimensionen. In welcher Reihenfolge wird die dreidimensionale Tabelle nun aufgefüllt? Stellen wir uns die Tabelle als Schichten von Matrizen vor, so wird zuerst die Matrix der ersten Schicht spaltenweise aufgefüllt. Dann wird die Matrix jeder weiteren Schicht spaltenweise aufgefüllt. Wir geben also ein

```
> wahl<-array(c(4,2,12,2,46,4,24,6),c(2,2,2),
                    dimnames=list(c("BWL","VWL"),
                    c("CDU","SPD"),c("w","m")))
```

Schauen wir uns `wahl` an:

```
> wahl
 , , w
     CDU SPD
 BWL   4  12
 VWL   2   2
 , , m
     CDU SPD
 BWL  46  24
 VWL   4   6
```

2.4 Ergänzungen und weiterführende Literatur

In diesem Kapitel haben wir Verfahren kennengelernt, mit denen man die wesentlichen Charakteristika eines Datensatzes mit mehreren Merkmalen beschreiben kann. Hierbei haben wir einige Aspekte nicht berücksichtigt. Das Histogramm ist ein spezieller Dichteschätzer, der aber nicht glatt ist. Mit *Kerndichteschätzern* erhält man eine glatte Schätzung der Dichtefunktion. Univariate und multivariate Dichteschätzer werden bei Härdle (1990b) beschrieben. Dort sind auch Funktionen in `S` zur Dichteschätzung zu finden. Rousseeuw et al. (1999) entwickelten einen zweidimensionalen Boxplot, den sie *Bagplot* nennen. Von Rousseeuw (1984) wurden zwei robuste Schätzer der Varianz-Kovarianz-Matrix vorgeschlagen. Beim *MVE-Schätzer* wird das Ellipsoid mit kleinstem Volumen bestimmt, das h der n Beobachtungen enthält, während man beim *MCD-Schätzer* die h Beobachtungen sucht, deren empirische Varianz-Kovarianz-Matrix die kleinste Determinante besitzt. Bei beiden Schätzern wird die Varianz-Kovarianz-Matrix durch die empirische Varianz-Kovarianz-Matrix der h Beobachtungen geschätzt. Ein Algorithmus zur schnellen Bestimmung des MCD-Schätzers und einige Anwendungen sind bei Rousseeuw & van Driessen (1999) zu finden. Einen weiteren wichtigen Aspekt multivariater Datensätze haben wir nicht berücksichtigt. In der Regel enthalten multivariate Datensätze eine Vielzahl fehlender Beobachtungen. Es gibt eine Reihe von Verfahren zur Behandlung fehlender Beobachtungen. Man kann zum Beispiel alle Objekte aus dem Datensatz entfernen, bei denen mindestens eine Beobachtung fehlt. Eine solche Vorgehensweise führt meistens zu einer drastischen Verringerung des Datenbestandes und ist deshalb nicht sinnvoll. Man wird eher versuchen, die fehlenden Beobachtungen zu ersetzen. Wie man hierbei vorgehen sollte, kann man bei Bankhofer (1995) finden. Bei Schafer (1997) ist ein bayesianischer Zugang zur Behandlung fehlender Beobachtungen in multivariaten Datensätzen zu finden.

2.5 Übungen

Übung 1. Im Rahmen der PISA-Studie wurde das Merkmal `Lesekompetenz` näher untersucht. Dabei wurden die mittleren Punktezahlen der Schüler in den Bereichen `Ermitteln von Informationen`, `Textbezogenes Interpretieren` und `Reflektieren und Bewerten` in jedem der 31 Ländern bestimmt. In Tabelle 2.12 sind die Ergebnisse zu finden.

Tabelle 2.12. Mittelwert der Punkte in den Bereichen der Lesekompetenz im Rahmen der PISA-Studie, vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S.533

Land	Ermitteln von Informationen	Textbezogenes Interpretieren	Reflektieren und Bewerten
Australien	536	527	526
Belgien	515	512	497
Brasilien	365	400	417
Dänemark	498	494	500
Deutschland	483	488	478
Finnland	556	555	533
Frankreich	515	506	496
Griechenland	450	475	495
Grossbritannien	523	514	539
Irland	524	526	533
Island	500	514	501
Italien	488	489	483
Japan	526	518	530
Kanada	530	532	542
Korea	530	525	526
Lettland	451	459	458
Liechtenstein	492	484	468
Luxemburg	433	446	442
Mexiko	402	419	446
Neuseeland	535	526	529
Norwegen	505	505	506
Österreich	502	508	512
Polen	475	482	477
Portugal	455	473	480
Russland	451	468	455
Schweden	516	522	510
Schweiz	498	496	488
Spanien	483	491	506
Tschechien	481	500	485
Ungarn	478	480	481
USA	499	505	507

Benutzen Sie bei der Lösung der folgenden Aufgaben bitte `S-PLUS`.

1. Bestimmen Sie den Mittelwertvektor und die Matrix der zentrierten Merkmale.
2. Bestimmen Sie die Matrix der standardisierten Merkmale.
3. Bestimmen Sie die empirische Varianz-Kovarianz-Matrix und die empirische Korrelationsmatrix der Daten.
4. Erstellen und interpretieren Sie die Streudiagrammmatrix.

Übung 2. Betrachten Sie die Daten in Tabelle 1.1 auf Seite 4.

1. Bestimmen Sie den Mittelwertvektor und die Matrix der zentrierten Merkmale mit `S-PLUS`.
2. Zeigen Sie, dass der Mittelwert der zentrierten Merkmale gleich 0 ist.
3. Prüfen Sie mit `S-PLUS` am Datensatz, dass der Mittelwert der zentrierten Merkmale gleich 0 ist.
4. Bestimmen Sie die Matrix der standardisierten Merkmale mit `S-PLUS`.
5. Zeigen Sie, dass die Stichprobenvarianz der standardisierten Merkmale gleich 1 ist.
6. Prüfen Sie mit `S-PLUS` am Datensatz, dass die Stichprobenvarianz der standardisierten Merkmale gleich 1 ist.
7. Bestimmen Sie die empirische Varianz-Kovarianz-Matrix und die empirische Korrelationsmatrix der Daten mit `S-PLUS`.
8. Bestimmen Sie mit `S-PLUS` die empirische Varianz-Kovarianz-Matrix der Daten mit Hilfe von Gleichung (2.17).
9. Bestimmen Sie mit `S-PLUS` die empirische Korrelationsmatrix der Daten mit Hilfe von Gleichung (2.20).

Übung 3. Im Rahmen einer Weiterbildungsveranstaltung sollten die Teilnehmer einen Fragebogen ausfüllen. Neben dem Merkmal `Geschlecht` mit den Ausprägungsmöglichkeiten `w` und `m` wurde noch eine Reihe weiterer Merkmale erhoben. Die Teilnehmer wurden gefragt, ob sie den Film Titanic gesehen haben. Dieses Merkmal bezeichnen wir mit `Titanic`. Außerdem sollten sie den folgenden Satz fortsetzen:

> Zu Risiken und Nebenwirkungen ...

Wir bezeichnen das Merkmal mit `Satz`. Es nimmt die Ausprägung `j`, wenn der Satz richtig fortgesetzt wurde. Ansonsten nimmt es den Wert `n` an. Die Ergebnisse sind in Tabelle 2.13 zu finden.

1. Erstellen Sie die dreidimensionale Kontingenztabelle.
2. Erstellen Sie die Kontingenztabelle der Merkmale `Geschlecht` und `Titanic` und bestimmen Sie die bedingten relativen Häufigkeiten.
3. Erstellen Sie die Kontingenztabelle der Merkmale `Geschlecht` und `Satz` und bestimmen Sie die bedingten relativen Häufigkeiten.
4. Erstellen Sie die Kontingenztabelle der Merkmale `Satz` und `Titanic` und bestimmen Sie die bedingten relativen Häufigkeiten.
5. Auf welche Zusammenhänge deuten die drei zweidimensionalen Kontingenztabellen hin?

Tabelle 2.13. Ergebnisse einer Befragung in einer Weiterbildungsveranstaltung

Person	Geschlecht	Titanic	Satz	Person	Geschlecht	Titanic	Satz
1	m	n	n	14	w	j	j
2	w	j	n	15	w	j	n
3	w	j	j	16	m	j	n
4	m	n	n	17	m	n	n
5	m	n	n	18	m	j	n
6	m	j	j	19	w	n	n
7	w	j	n	20	w	j	n
8	m	n	n	21	w	j	j
9	w	j	j	22	w	j	j
10	m	n	n	23	w	j	n
11	w	j	j	24	w	j	j
12	m	j	n	25	m	n	j
13	m	j	j				

http://www.springer.com/978-3-642-14986-3

Multivariate Analysemethoden
Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS
Handl, A.
2010, XVI, 491 S. 100 Abb., Softcover
ISBN: 978-3-642-14986-3