5 Kenngrößen univariater empirischer Verteilungen

5.1 Lagemaße

Häufigkeitsverteilungen für ungruppierte oder gruppierte Daten vermitteln einen Eindruck von der Gestalt der Verteilung eines Datensatzes. Die Histogramme in Abbildung 4.5 zur Verteilung von Bruttoverdiensten in zwei südeuropäischen Staaten zeigen z. B., dass die Verteilung der Daten in beiden Fällen eine deutliche Asymmetrie aufweist, also eine gewisse „Schiefe“ der Verteilung zu beobachten ist. Ferner sieht man bei beiden Teilgrafiken, dass das „Zentrum“ (oder der „Schwerpunkt“) der Einkommensverteilung für Portugal im Bereich kleinerer Werte liegt und auch die „Streuung“ hier geringer ist. Die Begriffe „Zentrum“, „Schwerpunkt“, „Streuung“ oder „Schiefe“ einer Verteilung sind zunächst unscharf und bedürfen der Präzisierung. Lage- und Streuungsparameter dienen dem Zweck, solche Befunde zu präzisieren und zu objektivieren. Es geht darum, die in einem Datensatz steckende Information zu wenigen Kenngrößen zu verdichten. Eine solche Informationsverdichtung ermöglicht eine unmissverständliche Beschreibung von Charakteristika eines Datensatzes, ist aber grundsätzlich mit Informationsverlust verbunden. So können zwei sehr unterschiedliche Datensätze einen ähnlichen Schwerpunkt oder eine vergleichbare Streuung aufweisen. Kenngrößen zur Beschreibung empirischer Verteilungen sind aber dennoch überaus wichtig. Sie liefern für einen gegebenen Datensatz nämlich wertvolle zusätzliche Informationen, die sich visuell aus der grafischen Darstellung einer empirischen Verteilung nicht immer ohne weiteres erschließen.

Wofür werden Kenngrößen von Verteilungen benötigt?

Zur Charakterisierung des „Zentrums“ einer Verteilung werden Lageparameter herangezogen. Ein besonders leicht zu bestimmender Lageparameter ist der **Modus** oder **Modalwert** x_{mod}. Dieser lässt sich immer anwenden, also auch bei auch Merkmalen, deren Ausprägungen nur Kategorien sind (qualitative Merkmale). Er ist definiert als die Merkmalsausprägung mit der größten Häufigkeit.

Beispiel 5.1 **Modus beim Datensatz zum ZDF-Politbarometer**

Beim Beispiel 4.1 (ZDF-Politbarometer vom 16. Oktober 2009, Merkmal „Parteipräferenz“) war die Ausprägung a_1 (Präferenz für die CDU/CSU) mit der größten Häufigkeit verbunden, d. h. hier ist $x_{mod} = a_1$. Anhand von Abbildung 4.4 lässt sich der Modus leicht bestimmen, weil die Häufigkeit $h(a_1)$ deutlich größer als alle anderen Häufigkeiten war. Wären zwei Häufigkeiten, z. B. $h(a_1)$ und $h(a_2)$ gleich groß, hätte man eine zweigipflige Häufigkeitsverteilung und es gäbe zwei Modalwerte (Modi). Der Modus ist also nur dann eindeutig erklärt, wenn die Häufigkeitsverteilung ein eindeutig bestimmtes Maximum aufweist.

Ein weiterer Lageparameter ist der **Median** $\widetilde{x}$ (lies: *x-Schlange*), der gelegentlich mit

x_{med} abgekürzt wird und für den man auch die Bezeichnung **Zentralwert** findet. Der Median ist nur bei mindestens ordinalskalierten Merkmalen anwendbar, also bei Merkmalen, für deren Werte eine natürliche Rangordnung erklärt ist. Betrachtet sei also ein – noch nicht notwendigerweise geordnet vorliegender – Datensatz $x_1, x_2, \ldots, x_n$ für ein solches Merkmal. Um zwischen dem ursprünglichen und dem geordneten Datensatz unterscheiden zu können, sei letzterer mit $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ bezeichnet.[1] Der Median ist dann, grob gesprochen, der „mittlere" Wert des geordneten Datensatzes. Bei ungeradem n ist dies der eindeutig bestimmte Wert $x_{(\frac{n+1}{2})}$. Bei geradem n gibt es hingegen zwei Werte $x_{(\frac{n}{2})}$ und $x_{(\frac{n}{2}+1)}$, die die Mitte des Datensatzes repräsentieren. In diesem Falle ist der Median bei einem ordinalskalierten Merkmal nicht eindeutig bestimmt, sofern sich die beiden Werte $x_{(\frac{n}{2})}$ und $x_{(\frac{n}{2}+1)}$ voneinander unterscheiden. Bezieht sich der Datensatz hingegen auf ein metrisch skaliertes Merkmal, so bildet man aus den beiden zentralen Werten den Mittelwert. Der Median ist dann also definiert durch

$$\widetilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2} \cdot (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{falls } n \text{ gerade.} \end{cases} \tag{5.1}$$

Der bekannteste Lageparameter ist der **Mittelwert**, der auch **arithmetisches Mittel** genannt und mit $\overline{x}$ abgekürzt wird (lies: *x-quer*). Er ist nur bei metrisch skalierten Merkmalen anwendbar und ergibt sich, indem man alle Werte $x_1, x_2, \ldots, x_n$ eines Datensatzes addiert und die resultierende Summe durch n dividiert:[2]

$$\overline{x} := \frac{1}{n} \cdot (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i. \tag{5.2}$$

Der Mittelwert berücksichtigt demnach alle Werte eines Datensatzes mit gleichem Gewicht $\frac{1}{n}$, während in die Berechnung eines Medians nur ein oder zwei zentrale Elemente eines Datensatzes eingehen. Wenn man also bei einem Datensatz den größten Wert $x_{max} = x_{(n)}$ deutlich vergrößert, hat dies nur auf den Mittelwert einen Effekt. Der Mittelwert reagiert demnach, anders als der Median, empfindlich gegenüber extremen Werten. Man spricht in diesem Zusammenhang von einer höheren *Sensitivität* oder auch von einer geringeren *Robustheit* des Mittelwerts gegenüber Ausreißern, d. h. gegenüber auffällig großen oder kleinen Beobachtungswerten.

Wenn man von jedem der Elemente $x_1, x_2, \ldots, x_n$ eines Datensatzes den Mittelwert subtrahiert und aufsummiert, resultiert 0, d. h. die Summe der Abweichungen $x_i - \overline{x}$ verschwindet:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0. \tag{5.3}$$

Gleichung (5.3) beinhaltet, dass sich der Mittelwert als Schwerpunkt des Datensatzes interpretieren lässt.

[1] Man kann auf die Notation $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ verzichten, wenn man von der Annahme ausgeht, dass der Datensatz $x_1, x_2, \ldots, x_n$ schon geordnet vorliegt.

[2] Das Summenzeichen Σ und andere mathematische Symbole sind in Tabelle 20.3 erklärt. Unter dem Summenzeichen wird für den – in (5.2) mit „i" bezeichneten – ganzzahligen Laufindex der Startwert angegeben, über dem Summenzeichen der letzte zu berücksichtigende Wert des Laufindexes.

Beispiel 5.2 **Median und Mittelwert für Daten zum Energieverbrauch**

In der Wochenzeitung „Die Zeit“, Ausgabe vom 11. 4. 2002, fand man in Ergänzung des Beitrags „Big Oil regiert“ von Th. Fischermann die nachstehende Tabelle mit umweltrelevanten Kennzahlen für die USA, Deutschland, Japan, China und Indien. Die Daten beziehen sich auf das Jahr 1999 und stammen von der Internationalen Energieagentur.

Land	Erdölverbrauch (in t/Kopf)	Stromverbrauch (in 1000 kWh/Kopf)	CO_2-Emissionen (in t/Kopf)
USA	8,32	13,45	20,46
Deutschland	4,11	6,48	10,01
Japan	4,07	8,13	9,14
China	0,87	0,91	2,40
Indien	0,48	0,42	0,91

Tab. 5.1: *Umweltrelevante Daten für fünf Staaten*

Man erkennt, dass die USA vergleichsweise großzügig Energie verbrauchen und CO_2 emittieren. Gedanklich stelle man sich 5 Personen vor, je eine Person aus den Ländern USA, Deutschland, Japan, China und Indien, für die jeweils die in Tabelle 5.1 angegebenen Verbrauchs- und Emissionswerte zutreffen, die also bezüglich der drei Merkmale als typische Vertreter ihrer Länder gelten können. Für diese kleine Personengruppe lässt sich dann der „mittlere“ Pro-Kopf-Verbrauch für Öl und Strom bzw. eine „mittlere“ CO_2-Emission ermitteln, wobei man den Median oder den Mittelwert des jeweiligen Datensatzes heranziehen kann.

Es seien hier die Daten für das metrisch skalierte Merkmal „Stromverbrauch / Kopf“ (in 1000 kWh) in der mittleren Spalte von Tabelle 5.1 betrachtet. Um den Median zu errechnen, sind die Werte $x_1 = 13,45$, $x_2 = 6,48$, $x_3 = 8,13$, $x_4 = 0,91$, $x_5 = 0,42$ zunächst nach Größe zu ordnen. Aus der resultierenden Folge $x_{(1)} = 0,42$, $x_{(2)} = 0,91$, $x_{(3)} = 6,48$, $x_{(4)} = 8,13$, $x_{(5)} = 13,45$ ergibt sich der Median für den hier vorliegenden Fall $n = 5$ nach (5.1) als $\widetilde{x} = x_{(3)} = 6,48$. Würde man bei dem ursprünglichen Datensatz den Wert $x_5 = 0,42$ für Indien unberücksichtigt lassen, den Median also nur auf der Basis der Datenreihe $x_1, \ldots, x_4$ ermitteln, erhielte man für $\widetilde{x}$ den Wert $\widetilde{x} = \frac{1}{2} \cdot (x_{(2)} + x_{(3)}) = 7,305$.

Bestimmt man mit denselben Ausgangsdaten den Mittelwert, so erhält man nach (5.2) den Wert $\overline{x} = \frac{1}{5} \cdot 29,39 = 5,878$. Würde man für x_1 anstelle von $13,45$ z. B. den 10-fach größeren Wert $134,50$ einsetzen, bliebe der Median unverändert bei $\widetilde{x} = 6,48$, während sich für den Mittelwert nun $\overline{x} = \frac{1}{5} \cdot 150,44 = 30,088$ ergäbe.

Die Berechnung des Mittelwerts kann etwas einfacher bewerkstelligt werden, wenn Merkmalswerte mehrfach auftreten. Hat man für ein diskretes Merkmal X mit den Aus-

Verwendung von Häufigkeitsverteilungen bei der Berechnung des Mittelwerts

prägungen $a_1, \ldots, a_k$ insgesamt n Beobachtungswerte $x_1, \ldots, x_n$ $(n > k)$, so würde die Anwendung von (5.2) implizieren, dass n Werte zu addieren sind. Anstelle der Urliste kann man hier für die Berechnung des Mittelwerts auch die relative Häufigkeitsverteilung $f(a_1), \ldots, f(a_k)$ verwenden und $\overline{x}$ nach

$$\overline{x} := a_1 \cdot f_1 + a_2 \cdot f_2 + \ldots + a_k \cdot f_k = \sum_{i=1}^{k} a_i \cdot f_i \tag{5.4}$$

als Summe von nur k Termen berechnen. Der Mittelwert $\overline{x}$ lässt sich also alternativ als Summe der mit den relativen Häufigkeiten f_i gewichteten Ausprägungen a_i ermitteln $(i = 1, 2, \ldots, k)$.

Die Formel (5.4) lässt sich in leicht modifizierter Fassung auch zur Berechnung des Mittelwerts bei *gruppierten Daten* verwenden. Man hat nur die Ausprägungen a_i durch die Mitte m_i der Klassen zu ersetzen und die Häufigkeiten f_i sind dann die relativen Klassenbesetzungshäufigkeiten.

Beispiel 5.3 **Bestimmung des Mittelwerts bei einem Würfelexperiment**

In Abbildung 4.9 wurde das Ergebnis eines 10 Würfe umfassenden Würfelexperiments veranschaulicht, bei dem vier Mal die 1, zwei Mal die 4, drei Mal die 5 und einmal die 6 beobachtet wurde. Nach (5.2) erhält man für $\overline{x}$ den Wert

$$\overline{x} = \frac{1}{10} \cdot (1 + 1 + 1 + 1 + 4 + 4 + 5 + 5 + 5 + 6) = \frac{1}{10} \cdot 33 = 3,3.$$

Zieht man bei der Berechnung des Mittelwerts (5.4) heran, resultiert mit den neben Abbildung 4.9 tabellierten relativen Häufigkeiten $f_i = f(a_i)$

$$\overline{x} = 1 \cdot 0,4 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot 0,2 + 5 \cdot 0,3 + 6 \cdot 0,1 = 3,3.$$

Die Vorteile der Formel (5.4) verstärken sich, wenn für n ein im Vergleich zu k noch größerer Wert gewählt wird, z. B. bei einem Würfelexperiment $n = 1000$ Würfe.

Gibt es einen „besten“ Lageparameter?

Welchen der vorgestellten Lageparameter sollte man aber verwenden? Hierzu gibt es keine allgemeingültige Ausage. Die Antwort hängt sowohl von der Skalierung des Merkmals ab als auch von der jeweiligen Fragestellung. Bei einem nominalskalierten Merkmal kann man nur den Modalwert verwenden. Bei einem metrisch skalierten Merkmal hat man schon drei Alternativen, nämlich den Modalwert, den Median und den Mittelwert und es ist zu überlegen, wie robust die zu berechnende Kenngröße gegenüber Extremwerten sein soll. Bei einem kleinen Datensatz für das Merkmal „Bruttoverdienst“ (in Euro / Stunde) kann z. B. ein einziger Extremwert den Mittelwert erheblich beeinflussen. Hier kann dann der Median aussagekräftiger sein, während der Modalwert i. a. wenig Information liefert, vor allem wenn die Verdienste auf Cent genau ausgewiesen werden. Bei metrisch skalierten Daten wird oft nicht nur ein Lageparameter berechnet, weil ein zweiter Parameter, etwa der Median zusätzlich neben dem Mittelwert, noch zusätzliche Information über die empirische Verteilung eines Datensatzes liefern kann. Bei einer

Einkommensverteilung kann man z. B. $\overline{x}$ und $\widetilde{x}$ vergleichen und hieraus Aussagen zur Symmetrie oder Asymmetrie der Verteilung ableiten.

Beispiel 5.4 **Irreführende Presseberichte zum realen Haushaltseinkommen**

Im März 2005 veröffentlichte das *Institute for Fiscal Studies* (IFS), ein unabhängiges Wirtschaftsforschungsinstitut in Großbritannien, einen Bericht „Poverty and Inequality in Britain“, in dem u. a. angeführt wurde, dass das mittlere verfügbare Hauseinkommen („average take-home income“) im Land im Zeitraum 2003/04 gegenüber dem Vorjahreszeitraum abgenommen habe, zum ersten Mal seit Beginn der 90-er Jahre, und zwar um $0,2$ % auf nunmehr 408 Britische Pfund. Dieser Befund wurde von der Presse sehr kritisch kommentiert, so dass schließlich Gordon Brown, der damalige Schatzkanzler und spätere Premierminister, unter Druck geriet und Stellung beziehen musste.

Die von den Medien aufgegriffene Information bezog sich auf den *Mittelwert* der Variablen „verfügbares Hauseinkommen“. Der Bericht führte aber auch an, ohne dass dies allerdings von den Journalisten aufgegriffen wurde, dass der *Median* im fraglichen Zeitraum um $0,5$ % gestiegen war und jetzt 336 Britische Pfund betrug. Der Median wäre aber zur Charakterisierung des „durchschnittlichen“ Haushaltseinkommens weitaus geeigneter als der Mittelwert, weil Einkommensverteilungen asymmetrisch sind und der Mittelwert hier durch extrem hohe und für die Grundgesamtheit eher untypische Werte stark beeinflusst werden kann. Man erkennt dies z. B. am Beispiel der Abbildung 4.5. Diese zeigte zwei Einkommensverteilungen und zusätzlich – oberhalb der Grafiken – den aus den Individualdaten errechneten Mittelwert sowie drei Dezile, von denen eines der dort mit D5 bezeichnete Median war. Bloßes Betrachten der Abbildungen macht schon deutlich, dass der Mittelwert für die betrachteten Grundgesamtheiten weniger repräsentativ als der Median ist. Der Anstieg des Medians um $0,5$ % war also bei dem IFS-Bericht die weitaus aussagekräftigere und positiv zu bewertende Information. Sie beinhaltete nämlich, dass der Wert, der die unteren 50 % der Haushaltseinkommen von den oberen 50 % trennte, sich leicht nach oben verschoben hatte, d. h. die Ungleichheit der Verteilung der Haushaltseinkommen hatte leicht abgenommen. [3]

Gordon Brown. Quelle: World Economic Forum

Dass die Journalisten den Report negativ kommentierten, lag entweder daran, dass sie zwischen Mittelwert und Median nicht recht zu unterscheiden wussten oder aber unterstellten, dass dies für die Leser zutrifft. Statistische Methodenkompetenz ist offenbar eine Voraussetzung dafür, besser gegenüber unscharfen oder manipulativen Darstellungen statistischer Sachverhalte in den Medien gefeit zu sein.

---

[3] Dieser Befund schlug sich im Bericht in einer leichten Zunahme des *Gini-Koeffizienten* nieder, der neben dem Quotienten von Dezilen, etwa $\frac{D9}{D1}$, als Maß für Einkommensungleichheiten Verwendung findet (vgl. hierzu Kapitel 6).

Exkurs 5.1 **Weitere Lageparameter**

Mittelwert und Median sind Lösungen unterschiedlicher Minimierungsprobleme. Der Mittelwert hat die Eigenschaft, für einen gegebenen Datensatz $x_1, x_2, \ldots, x_n$ denjenigen Wert z zu repräsentieren, der die Summe der quadrierten Abweichungen $(x_i - z)^2$ minimiert:

$$z = \overline{x}: \qquad \sum_{i=1}^{n} (x_i - z)^2 \to Min.$$

Der Median minimiert hingegen die Summe der absoluten Abweichungen $|x_i - z|$:

$$z = \widetilde{x}: \qquad \sum_{i=1}^{n} |x_i - z| \to Min.$$

Einen Beweis dieser Aussagen findet man z. B. bei Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 2.2.1) oder Burkschat / Cramer / Kamps (2004, Abschnitt 3.2).

Neben den vorgestellten Kenngrößen zur Charakterisierung der Lage empirischer Verteilungen gibt es für metrisch skalierte Merkmale noch einige weitere Lageparameter. Zu nennen ist hier vor allem das **gewichtete arithmetische Mittel**, bei dem die Werte $x_1, x_2, \ldots, x_n$ eines Datensatzes, anders als beim ungewichteten „gewöhnlichen“ Mittelwert (5.2), mit unterschiedlichen Gewichten versehen werden. Will man z. B. anhand der Stromverbrauchsdaten aus Tabelle (5.1) den mittleren Stromverbrauch für alle Einwohner der 5 in der Tabelle aufgeführten Länder berechnen, also nicht nur für eine modellhafte Gruppe von fünf Ländervertretern, so bezöge sich die Mittelwertbildung auf einen Datensatz, dessen Umfang n durch die Summe $n_1+n_2+n_3+n_4+n_5$ der Bevölkerungszahlen aller 5 Länder gegeben wäre. Damit Länder mit sehr unterschiedlichen Bevölkerungszahlen, etwa China und Deutschland, bei der Bildung des Mittelwerts angemessen berücksichtigt werden, wird der Wert x_i für ein Land jeweils mit dem als Gewichtungsfaktor fungierenden Wert n_i multipliziert.

Zu erwähnen ist ferner das **getrimmte arithmetische Mittel**. Dieses lässt einen kleineren Anteil der Randdaten $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ eines nach aufsteigender Größe geordneten Datensatzes unberücksichtigt. Wenn dieser Anteil α beträgt, spricht man auch von einem α-getrimmten Mittelwert und kürzt diesen mit $\overline{x}_\alpha$ ab. Bei der Berechnung von $\overline{x}_\alpha$ werden die unteren und oberen $\frac{\alpha}{2} \cdot 100\%$ des geordneten Datensatzes vor der Mittelwertberechnung eliminiert. Dies führt dazu, dass getrimmte Mittelwerte, ähnlich wie der Median, robuster gegenüber Extremwerten (Ausreißerdaten) sind.

Als weiterer Lageparameter ist das **geometrische Mittel** $\overline{x}_g$ zu nennen. Dieses wird für Datensätze $x_1, x_2, \ldots, x_n$ verwendet, die Veränderungsraten repräsentieren, z. B. zur Quantifizierung von Wachstumsraten bei Unternehmensgewinnen oder von Lernzuwächsen, die anhand lernpsychologischer Experimente bestimmt werden. Das geometrische Mittel errechnet sich als

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

◇

5.2 Streuungsmaße

Ein Datensatz definiert eine empirische Verteilung eines Merkmals. Im vorigen Abschnitt wurde illustriert, dass eine solche Verteilung ein „Zentrum“ besitzt, das man anhand einer oder mehrerer Kenngrößen charakterisieren kann. Bei einem metrisch skalierten Merkmal stehen vor allem der Modalwert, der Median und der Mittelwert zur Verfügung, wobei man hier i. a. den Mittelwert oder den Median verwenden wird. Die Kenntnis des Schwerpunktes reicht aber nicht aus, um einen Datensatz zu beschreiben. Zwei Datensätze können in den Lageparametern übereinstimmen und sich dennoch bezüglich der Variation der Merkmalswerte deutlich unterscheiden. Hat man z. B. einen Datensatz $x_1, x_2, \ldots, x_n$ mit Mittelwert $\overline{x}$, so lässt die alleinige Kenntnis von $\overline{x}$ offen, ob die einzelnen Elemente des Datensatzes alle sehr nahe am Mittelwert liegen, mit ihm gar alle übereinstimmen oder von $\overline{x}$ stark nach oben und unten abweichen und sich nur „ausmitteln“. Zur Charakterisierung von Merkmalen, für die Abstände zwischen Merkmalsausprägungen erklärt sind, also bei quantitativen Merkmalen (metrische Merkmalsskalierung), muss man somit noch Kenngrößen heranziehen, die die Streuung innerhalb des Datensatzes messen.

Warum braucht man auch Kenngrößen für die Streuung von Datensätzen?

Ein besonders einfaches Streuungsmaß für metrisch skalierte Merkmale ist die **Spannweite** R eines Datensatzes.[4] Um diese zu berechnen, ordnet man – wie bei der Berechnung des Medians $\widetilde{x}$ – den Datensatz zunächst nach aufsteigender Größe. Die Spannweite ergibt sich dann aus dem geordneten Datensatz $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ als Differenz aus dem größten Wert $x_{(n)}$ und dem kleinsten Wert $x_{(1)}$:

$$R := x_{(n)} - x_{(1)}. \tag{5.5}$$

Die Spannweite hat den Nachteil, dass sie eine hohe Empfindlichkeit bzw. eine geringe Robustheit gegenüber Ausreißern besitzt. Ändert man in einem Datensatz den maximalen oder den minimalen Wert stark, wirkt sich dies auch massiv auf den Wert von R aus.

Ein sehr häufig verwendetes Maß für die Streuung eines Datensatzes ist die **Varianz** oder **Stichprobenvarianz** s^2, die auch **empirische Varianz** genannt wird.[5] In die Varianz gehen die Abweichungen $x_i - \overline{x}$ der Merkmalswerte vom Mittelwert $\overline{x}$ ein; $i = 1, 2, \ldots, n$. Wegen (5.3) kommt die Verwendung des Mittelwerts aus allen Abweichungen $x_i - \overline{x}$ nicht in Betracht. Die Varianz bildet statt dessen den Mittelwert aus den quadrierten Abweichungen $(x_i - \overline{x})^2$, d. h. es gilt

$$s^2 := \frac{1}{n} \cdot \left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2)\right] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{5.6}$$

Wenn man die Varianz ohne Rechner per Hand ermittelt, kann die nachstehende, auch

[4] Die Abkürzung „R“ leitet sich aus dem englischen Wort „range“ für Spannweite her. Die Verwendung eines Großbuchstabens „R“ trägt dazu bei, dass Verwechslungen mit dem Korrelationskoeffizienten r nach Bravais-Pearson (s. Abschnitt 9.2) vermieden werden.

[5] Das Verhalten von Zufallsvariablen wird in Kapitel 11 - 12 anhand von Modellen (Wahrscheinlichkeitsverteilungen) charakterisiert. Hier spricht man von *theoretischen Verteilungen* und diese lassen sich ebenfalls anhand von Lage- und Streuungsparametern beschreiben, z. B. anhand des Erwartungswerts μ (lies: mü) und der theoretischen Varianz σ^2 (lies: sigma-Quadrat). Kenngrößen empirischer und theoretischer Verteilungen sollten jedenfalls mit unterschiedlichen Notationen belegt sein.

als **Verschiebungssatz** bezeichnete Zerlegungsformel nützlich sein, bei der $\overline{x^2}$ das arithmetische Mittel der quadrierten Elemente $x_1^2, \ldots, x_n^2$ des Datensatzes bezeichnet:[6]

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2. \tag{5.7}$$

Die Darstellung (5.7) geht aus (5.6) hervor, wenn man dort den quadrierten Term $(x_i - \overline{x})^2$ hinter dem Summenzeichen ausmultipliziert (binomische Formel) und die Summierung dann gliedweise vornimmt.

Die Varianz s^2 ist ein *quadratisches* Streuungsmaß. Sind die Originaldaten z. B. Werte in *cm* oder in *sec*, so wird die Varianz in cm^2 bzw. in sec^2 gemessen. Die Kenngröße (5.6) geht in ein *lineares* Streuungsmaß über, wenn man die Wurzel zieht. Man erhält so die **Standardabweichung** oder, genauer, die **empirische Standardabweichung**

$$s := \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{\overline{x^2} - \overline{x}^2} \tag{5.8}$$

des Datensatzes. Diese wird – wie auch Median $\widetilde{x}$ und Mittelwert $\overline{x}$ – in der Einheit angegeben, in der die Ausgangsdaten gemessen werden. Die Standardabweichung ist daher im Vergleich zur Varianz ein wesentlich anschaulicheres Streuungsmaß.

Vorsicht: Uneinheitliche Definition von Varianz und Standardabweichung

Die Bezeichnungen für Varianz und Standardabweichung eines Datensatzes sind in der Lehrbuchliteratur leider nicht einheitlich. Häufig wird für die Varianz anstelle von (5.6) eine Formel verwendet, bei der vor dem Summenterm anstelle von $\frac{1}{n}$ der Term $\frac{1}{n-1}$ steht. Das dann resultierende und hier mit s^{*2} abgekürzte Streuungsmaß

$$s^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1} \cdot s^2. \tag{5.9}$$

wird **korrigierte Varianz** oder **korrigierte Stichprobenvarianz** genannt (vgl. auch Mosler / Schmid (2009, Abschnitt 5.1.4)). Durch Wurzelziehen geht aus (5.9) die **korrigierte Standardabweichung** s^* hervor.

Die korrigierte Varianz wird beim Schätzen und Testen anstelle von (5.7) bevorzugt verwendet, weil sie – wie mit (14.8) und (14.9) noch gezeigt wird – günstigere Eigenschaften besitzt. Die Division durch $n - 1$ wird jedenfalls erst im Kontext der schließenden Statistik nachvollziehbar; sie lässt sich im Rahmen der beschreibenden Statistik nicht motivieren. Wichtig ist aber, dass man bei Verwendung eines Taschenrechners oder ei-

[6] Sind mehrere Merkmale im Spiel, etwa X und Y, so kann man zwischen den empirischen Varianzen und Standardabweichungen durch Verwendung tiefgestellter Indizes differenzieren, etwa s_x^2 und s_y^2 im Falle der Varianzen.

ner Statistiksoftware weiß, welche Formel der Berechnungsprozedur zugrunde lag.[7] In diesem Manuskript werden die Bezeichnungen „Varianz“ und „Standardabweichung“ für Kenngrößen eines Datensatzes stets auf (5.6) bzw. (5.7) bezogen und mit s^2 bzw s abgekürzt. Aus der Varianz s^2 kann man wegen $s^{*2} = \frac{n}{n-1} \cdot s^2$ leicht die korrigierte Varianz s^{*2} berechnen und umgekehrt. Die Unterschiede zwischen beiden Größen verschwinden mit zunehmendem n, können aber bei kleinem n durchaus ins Gewicht fallen.

Beispiel 5.5 **Spannweite und Standardabweichung (Stromverbrauchsdaten)**

Geht man erneut vom Datensatz zum Pro-Kopf-Strom-Verbrauch in den USA, Deutschland, Japan, China resp. Indien aus (mittlere Spalte in Tabelle 5.1), so ist dieser für die Berechnung von R zunächst in die geordnete Folge $x_{(1)} = 0,42$, $x_{(2)} = 0,91$, $x_{(3)} = 6,48$, $x_{(4)} = 8,13$, $x_{(5)} = 13,45$ zu überführen. Es errechnet sich dann $R = 13,45 - 0,42 = 13,03$. Würde man bei dem ursprünglichen Datensatz den Wert $13,45$ für die USA z. B. auf den Wert $8,13$ von Japan herabsetzen, hätte dies für die Spannweite einen erheblichen Effekt. Es resultierte nun für R der Wert $R = 8,13 - 0,42 = 7,71$.

Bei der Berechnung der empirischen Varianz nach (5.6) werden die Originaldaten um den Mittelwert $\overline{x} = 5,878$ vermindert und die resultierenden Mittelwertabweichungen quadriert, aufsummiert und durch $n = 5$ dividiert. Man erhält so bei Rundung auf drei Nachkommastellen

$$s^2 = \frac{1}{5} \cdot \left[7,572^2 + 0,602^2 + 2,252^2 + (-4,968)^2 + (-5,458)^2\right] \approx 23,448.$$

Geht man alternativ von (5.7) aus, erhält man, wenn man wieder auf drei Dezimalstellen rundet und auf den in Beispiel 5.2 errechneten Mittelwert $\overline{x} = 5,878$ zurückgreift die etwas kürzere Rechnung

$$s^2 = \frac{1}{5} \cdot 289,9943 - 5,878^2 \approx 57,999 - 34,551 = 23,448.$$

Für die Standardabweichung folgt mit (5.8)

$$s = \sqrt{\frac{1}{5} \cdot [7,572^2 + 0,602^2 + 2,252^2 + (-4,968)^2 + (-5,458)^2]} \approx 4,842.$$

Die korrigierte empirische Varianz errechnet sich nach (5.9) als $s^{*2} = \frac{5}{4} \cdot s^2 \approx 29,310$. Der Unterschied zu $s^2 \approx 23,448$ ist deutlich, weil der Umfang n des Datensatzes hier klein ist ($n = 5$).

[7] In EXCEL wird über „Einfügen / Funktion / STABWNA“ eine Prozedur zur Berechnung der empirischen Standardabweichung s gemäß (5.7) angeboten und unter „Einfügen / Funktion / STABWA“ zusätzlich eine Berechnung für die korrigierte Standardabweichung s^*. Bei der Statistiksoftware SPSS wird hingegen bei der Berechnung von Varianz und Standardabweichung eines Datensatzes stets durch $n-1$ dividiert. SPSS bezeichnet ein in den *Sozialwissenschaften* und auch in der *Psychologie* häufig verwendetes Statistik-Softwarepaket (die Abkürzung stand anfangs für *S*tatistical *P*ackage for the *S*ocial *S*ciences), das ab Herbst 2010 in der Version 19 vorliegt. Als Alternative zu kommerzieller Statistiksoftware wird bei der statistischen Analyse von Daten zunehmend *R* eingesetzt – eine kostenfreie und inzwischen überaus leistungsfähige Statistik-Software und Programmierumgebung.

Verwendung von Häufigkeitsverteilungen bei der Berechnung der Varianz

Wie bei der Berechnung des Mittelwertes $\overline{x}$ kann man auch bei der Ermittlung der Varianz im Falle mehrfach auftretender Merkmalswerte auf relative Häufigkeiten zurückgreifen. Liegt für ein diskretes Merkmal X mit den Ausprägungen $a_1, \ldots, a_k$ eine größere Anzahl n von Beobachtungswerten $x_1, \ldots, x_n$ vor $(n > k)$, so wären bei der Anwendung von (5.7) n Mittelwertabweichungen $x_i - \overline{x}$ zu quadrieren. Statt der Abweichungen $x_i - \overline{x}$ der Urwerte vom Mittelwert kann man alternativ die Abweichungen $a_i - \overline{x}$ der Merkmalsausprägungen vom Mittelwert heranziehen und deren Quadrate mit den Elementen f_i der relativen Häufigkeitsverteilung $f_1 = f(a_1), \ldots, f_k = f(a_k)$ gewichten. Man erhält so für s^2 die zu (5.4) analoge alternative Berechnungsformel

$$s^2 = (a_1 - \overline{x})^2 \cdot f_1 + (a_2 - \overline{x})^2 \cdot f_2 + \ldots + (a_k - \overline{x})^2 \cdot f_k = \sum_{i=1}^{k} (a_i - \overline{x})^2 \cdot f_i, \qquad (5.10)$$

bei der sich die Summenbildung auf nur k Terme bezieht. Auch diese Formel lässt sich zur Varianzberechnung bei *gruppierten Daten* heranziehen, wenn man die Ausprägungen a_i durch die Mitte m_i der Klassen ersetzt. Die Häufigkeiten f_i entprechen dann wieder den relativen Besetzungshäufigkeiten der einzelnen Klassen.

Beispiel 5.6 **Varianz bei einem Würfelexperiment**

Es sei noch einmal der Datensatz $\{1, 1, 1, 1, 4, 4, 5, 5, 5, 6\}$ zugrunde gelegt, der den Ausgang des in Abbildung 4.9 veranschaulichten Würfelexperiments beschreibt (Augenzahlen bei 10 Würfen mit einem Würfel). In Beispiel 5.3 war auf der Basis dieser 10 Werte der Mittelwert $\overline{x} = 3,3$ berechnet worden und zwar anhand der Urwerte und alternativ unter Verwendung der relativen Häufigkeiten.

Aufgabe 5.1

Wenn man die Varianz s^2 unter Rückgriff auf die Urwerte berechnet, kann man (5.6) oder (5.7) verwenden. Bei Verwendung von (5.7) ergibt sich

$$s^2 = \frac{1}{10} \cdot 147 - 3,3^2 = 14,70 - 10,89 = 3,81.$$

Zieht man bei der Berechnung der Varianz des Datensatzes (5.10) heran, resultiert

$$\begin{aligned} s^2 :&= \left[(-2,3)^2 \cdot 0,4 + (-1,3)^2 \cdot 0 + (-0,3)^2 \cdot 0 + 0,7^2 \cdot 0,2 + 1,7^2 \cdot 0,3 + 2,7^2 \cdot 0,1\right] \\ &= 2,116 + 0,098 + 0,867 + 0,729 = 3,81. \end{aligned}$$

Standardisierung von Datensätzen

Wenn man Datensätze $x_1, x_2, \ldots, x_n$, die sich auf Messungen in unterschiedlichen Grundgesamtheiten beziehen oder die mit unterschiedlichen Messinstrumenten gewonnen wurden, direkt vergleichbar machen will, kann man von jedem Element eines Datensatzes jeweils dessen Mittelwert $\overline{x}$ subtrahieren und die Differenz noch durch die Standardabweichung s oder die korrigierte Standardabweichung s^* dividieren. Es resultieren neue Datensätze $y_1, y_2, \ldots, y_n$ mit Mittelwert $\overline{y} = 0$ und Standardabweichung $s = 1$ resp. $s^* = 1$. Solche Transformationen sind z. B. sinnvoll, wenn man Intelligenzmessungen in

unterschiedlichen Grundgesamtheiten durchführen oder schulische Leistungen anhand unterschiedlicher Fragebögen messen will. Die beschriebene Transformation wird in der *Psychologie* und in den *Sozialwissenschaften* auch **z-Transformation** genannt. Sie ist das empirische Analogon zu der in Abschnitt 12.2 dieses Manuskripts noch ausführlicher behandelten Transformation (12.11), die zur Standardisierung von Zufallsvariablen herangezogen wird.

Exkurs 5.2 **Verhalten der Kenngrößen bei Lineartransformation**

Varianz s^2 und Standardabweichung s sind Streuungsmaße, die sich auf Abweichungen $x_i - \overline{x}$ vom *Mittelwert* eines Datensatzes für ein metrisch skaliertes Merkmal beziehen. Ein alternatives Streuungsmaß ist die **mittlere absolute Abweichung vom Median** . Dieses oft mit d abgekürzte Maß basiert auf Abweichungen $x_i - \widetilde{x}$ vom Median, bildet aber nicht den Mittelwert aus den Quadraten, sondern aus den Absolutbeträgen dieser Abweichungen:

$$d := \frac{1}{n} \sum_{i=1}^{n} |x_i - \widetilde{x}|$$

Wenn man die Daten x_i für ein quantitatives Merkmal einer Lineartransformation $y_i = a + b \cdot x_i$ unterzieht, so werden Median und Mittelwert sowie die Standardabweichung in gleicher Weise transformiert, d. h es gilt z. B. für den Mittelwert y der transformierten Daten die Beziehung $\overline{y} = a + b \cdot \overline{x}$. Auf die Varianz und die Standardabweichung wirkt sich die Niveauverschiebung a nicht aus; nur der Wert von b ist hier relevant. Bezeichnet man die empirische Varianz des ursprünglichen Merkmals X mit s_x^2 und die des transformierten Merkmals Y mit s_y^2, so gilt $s_y^2 = b^2 \cdot s_x^2$ und $s_y = |b| \cdot s_x$.

Mediane, Mittelwerte und Standardabweichungen von Datensätzen sind also vom Maßstab abhängig. Für quantitative Merkmale mit nicht-negativen Ausprägungen wird oft der durch

$$v := \frac{s}{\overline{x}}$$

definierte **Variationskoeffizient** verwendet (*maßstabsunabhängiges* Streuungsmaß).

◇

5.3 Quantile und Boxplots

Der für ein metrisch oder mindestens ordinalskaliertes Merkmal erklärte Median $\widetilde{x}$ hat die Eigenschaft, dass mindestens 50% der nach Größe geordneten Elemente $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ eines Datensatzes kleiner oder gleich und mindestens 50% größer oder gleich $\widetilde{x}$ sind. Bei den 5 Werten in der mittleren Spalte von Tabelle 5.1 war der Median z. B. durch $\widetilde{x} = x_{(3)} = 6,48$ gegeben und je 3 der 5 Elemente in dieser Spalte, d. h. 60% der Werte, waren kleiner oder gleich resp. größer oder gleich $\widetilde{x}$. Bei ordinalskaliertem Merkmal ist $\widetilde{x}$ nicht immer eindeutig bestimmt. Bei metrischer Skalierung gilt dies im Prinzip auch; hier lässt sich aber über (5.1) eine eindeutige Festlegung erreichen.

Verallgemeinerung des Medians

Der Median markiert also die „Mitte“ eines Datensatzes. Eine Verallgemeinerung des

Medians ist das **p-Quantil**. Auch dieses setzt wieder ein metrisch oder zumindest ordinalskaliertes Merkmal voraus. Ein p-Quantil wird mit x_p abgekürzt und hat die Eigenschaft, dass mindestens $p \cdot 100\%$ der Elemente der geordneten Folge $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ kleiner oder gleich und mindestens $(1-p) \cdot 100\%$ größer oder gleich x_p sind.[8] Abbildung 5.1 veranschaulicht diese Definition.

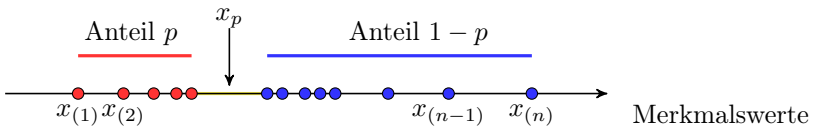

Abb. 5.1: *Veranschaulichung des p-Quantils*

Auch das p-Quantil ist bei einem ordinalskalierten Merkmal i. d. R. nicht eindeutig bestimmt. Bei metrischer Merkmalsskalierung kann, analog zur Definition des Medians, eine eindeutige Bestimmbarkeit erreicht werden, wenn das arithmetische Mittel derjenigen zwei Merkmalsausprägungen herangezogen wird, zwischen denen das p-Quantil liegt. Bezeichne $[np]$ die größte ganze Zahl, die kleiner oder gleich np ist. Es ist dann $[np]+1$ die kleinste ganze Zahl, die größer als np ist.[9]Mit dieser Notation kann x_p bei einem metrisch skalierten Merkmal in Verallgemeinerung von (5.1) definiert werden durch (vgl. z. B. Burkschat / Cramer / Kamps (2004, Abschnitt 3.2).

$$x_p = \begin{cases} x_{([np]+1)} & \text{falls } np \text{ nicht ganzzahlig} \\ \frac{1}{2} \cdot (x_{(np)} + x_{(np+1)}) & \text{falls } np \text{ ganzzahlig.} \end{cases} \tag{5.11}$$

Spezielle Quantile

Der Median ist demnach ein spezielles Quantil, nämlich das $0,5$-Quantil. Weitere wichtige Quantile sind das $0,25$-Quantil und das $0,75$-Quantil, die **unteres Quartil** resp. **oberes Quartil** genannt werden. Abbildung 5.2 veranschaulicht diese drei Spezialfälle.

Die häufig mit Q abgekürzte Differenz der beiden Quartile $x_{0,75}$ und $x_{0,25}$, also

$$Q := x_{0,75} - x_{0,25}, \tag{5.12}$$

Flash-Animation „Quantile“

wird **Quartilsabstand** genannt. Sie wird in manchen Lehrbüchern auch als **Interquartilsabstand** IQR angesprochen (engl: *i*nter*q*uartile *r*ange). Ferner sind noch die Dezile zu nennen, die sich bei Wahl von $p = 0,1, p = 0,2, \ldots, p = 0,9$ ergeben und oft mit D1, D2, ..., D9 abgekürzt werden. Der Median $\widetilde{x} = x_{0,5}$ stimmt also mit dem Dezil D5 überein.

Wie erkennt man eine asymmetrische empirische Verteilung?

In Abbildung 4.5 waren für spanische und portugiesische Arbeitnehmer Bruttojahresverdienste in Form von Histogrammen visualisiert, wobei über den Histogrammen jeweils die aus den Originaldaten (ungruppierte Daten) errechneten Dezile D1 und D9 sowie der Median D5 $= \widetilde{x}$ und der Mittelwert $\overline{x}$ wiedergegeben war. Das ebenfalls aus-

[8] Die Notation für Quantile ist in der Literatur nicht ganz einheitlich. Man findet z. B. auch die Schreibweise $\widetilde{x}_p$ anstelle von x_p; vgl. z. B. Steland (2010, Abschnitt 1.6.4) oder Toutenburg / Heumann (2009, Abschnitt 3.1.2)

[9] Die auf Carl Friedrich Gauss zurückgehende Funktion $f(x) = [x]$ wird *Gauß-Klammer-Funktion* oder *Abrundungsfunktion* genannt. Sie ist eine für alle reellen Zahlen erklärte Treppenfunktion mit Sprungstellen bei jeder ganzen Zahl (Sprunghöhe 1). Es ist z. B. $[3,8] = 3$.

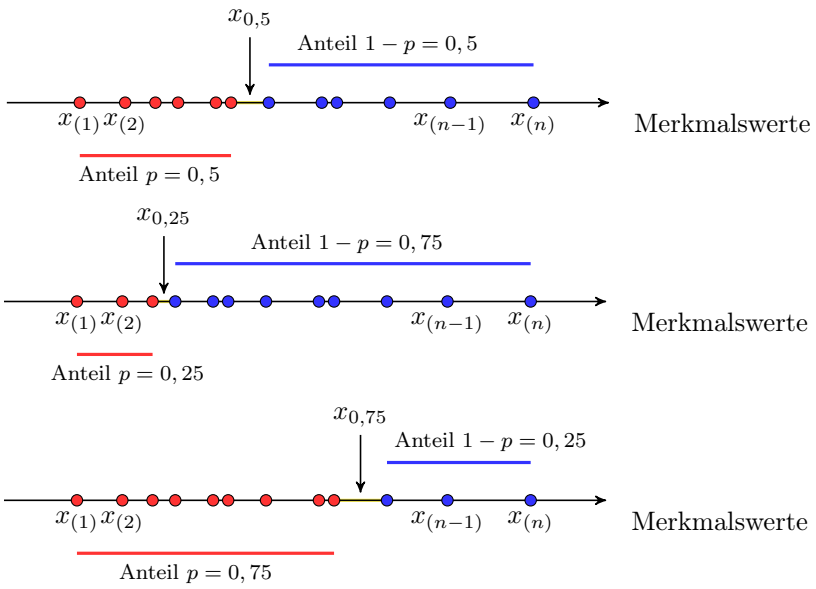

Abb. 5.2: *Median $x_{0,5}$, unteres Quartil $x_{0,25}$ und oberes Quartil $x_{0,75}$ als spezielle Quantile*

gewiesene Verhältnis $\frac{D9}{D1}$ der extremen Dezile liefert eine Information über den Grad der Ungleichheit der Verdienste in der betrachteten Grundgesamtheit von Arbeitnehmern – hohe Werte des Quotienten sprechen für eine ausgeprägte Ungleichheit. Man erkennt natürlich schon anhand der Grafiken, dass sich der überwiegende Teil der in Abbildung 4.5 veranschaulichten Verdienste, insbesondere bei der Grafik für Portugal, in den unteren Einkommensbereichen bewegen, d. h. der überwiegende Teil der Daten ist linksseitig konzentriert – hier sind höhere Klassenbesetzungshäufigkeiten und damit ein steilerer Abfall der Verteilung zu beobachten. Man spricht dann von einer **linkssteilen Verteilung**. Eine **rechtssteile Verteilung** würde hingegen an der rechten Flanke steiler abfallen. In beiden Fällen liegt eine **asymmetrische Verteilung** vor. Die Nicht-Übereinstimmung von Median und Mittelwert einer empirischen Verteilung ist ebenfalls schon ein Indiz für eine Asymmetrie der betreffenden Verteilung.

Boxplots:

- Basisvariante

Ein sehr aussagekräftiges grafisches Instrument zur Beurteilung einer empirischen Verteilung (Zentrum, Streuung, Asymmetrie) ist der sog. **Boxplot** („Schachtelzeichnung“). Dieser fasst in seiner einfachsten Form fünf Charakteristika eines Datensatzes zusammen, nämlich die beiden Extremwerte $x_{min} = x_{(1)}$ und $x_{max} = x_{(n)}$, deren Differenz $x_{(n)} - x_{(1)}$ nach (5.5) die Spannweite R darstellt, die beiden Quartile $x_{0,25}$ und $x_{0,75}$ sowie den Median $x_{0,5}$.

Die beiden Quartile definieren die Länge einer Box („Schachtel“), in der noch der Median in Form eines Strichs oder Punktes markiert ist. Die Box wird mit den Extremwerten durch Linien verbunden (sog. „whisker“, übersetzt: Schnurrhaare), deren Ende durch einen Strich markiert wird. Die Länge der Box entspricht also dem Quartilsabstand Q. Abbildung 5.3 veranschaulicht die Konstruktion. Innerhalb der Box liegen etwa 50% der Daten, unterhalb und oberhalb der Box jeweils ca. 25%. Der Median liefert eine Information zum Zentrum des Datensatzes. Manchmal wird neben dem Median auch noch der Mittelwert innerhalb der Box dargestellt. Bei einer symmetrischen Verteilung liegt

Aufgabe 5.2

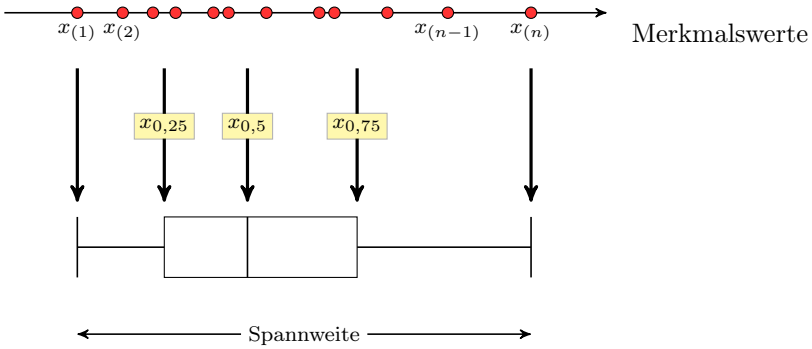

Abb. 5.3: *Aufbau eines Boxplots (Basisversion)*

der Median genau in der Mitte der Box.

- Modifikation (Visualisierung von Ausreißern)

Abbildung 5.3 zeigt nur die einfachste Boxplot-Variante. Häufig wird eine andere, hier nur der Vollständigkeit halber erwähnte Version mit gleichem Aufbau der Box, aber anderer Begrenzung der an der Box angebrachten Linien verwendet. Statt die Linien stets genau bis zu den Extremwerten zu führen, kann man auch so verfahren, dass man die Linien nur dann bis zu den Extremwerten zeichnet, wenn deren Abstand zur Box nicht größer ist als das $1,5$-fache des Interquartilabstands IQR. Die an der Box angesetzten Linien werden andernfalls auf die Länge $1,5$ IQR begrenzt und weiter entfernt liegende Werte separat eingezeichnet. So lassen sich auffällige Datenpunkte („Ausreißer“) hervorheben.

Beispiel 5.7 **Boxplots zu Bruttoverdiensten in Europa**

Java-Applet „Bruttoverdienste in Europa 2002“ (View-Option „Boxplots“)

Abbildung 4.1 zeigte Bruttostundenverdienste des Europäischen Amts für Statistik (Eurostat) in 27 europäischen Staaten für das Referenzjahr 2002 anhand eines Säulendiagramms. Die Darstellung bezog sich auf den Bereich „Industrie und Dienstleistungen“, in dem 9 Wirtschaftszweige zusammengefasst sind. Die in Abbildung 4.1 veranschaulichten Werte sind Mittelwerte aus den Verdiensten in diesen Branchen (gewichtete Mittel mit der Anzahl der in einem Wirtschaftszweig Beschäftigten als Gewichte). Wenn man ein etwas differenziertes Bild gewinnen will und z. B. auf einen Blick erfassen möchte, wie die Verdienste in den einzelnen Ländern von Branche zu Branche streuen, kann man für jedes Land einen Boxplot heranziehen, der den aus 9 Branchenverdiensten bestehenden Datensatz für jedes Land zu 5 Charakteristika aggregiert.

Der Boxplot für Deutschland ist in der Grafik betont. Der die obere Begrenzung des Boxplots definierende maximale Wert des Datensatzes, also die Branche, in der in Deutschland die Verdienste am höchsten sind, ist ebenfalls hervorgehoben. Es ist dies der Finanzsektor „Kreditinstitute und Versicherungen“, der nach der „*n*omenclature générale des *ac*tivités *é*conomique“ (amtliche Klassifikation NACE für Wirtschaftszweige; Stand 2006) mit “J“ codiert wurde.

Man erkennt anhand des Niveaus der Mediane, wie extrem das mittlere Verdienstni-

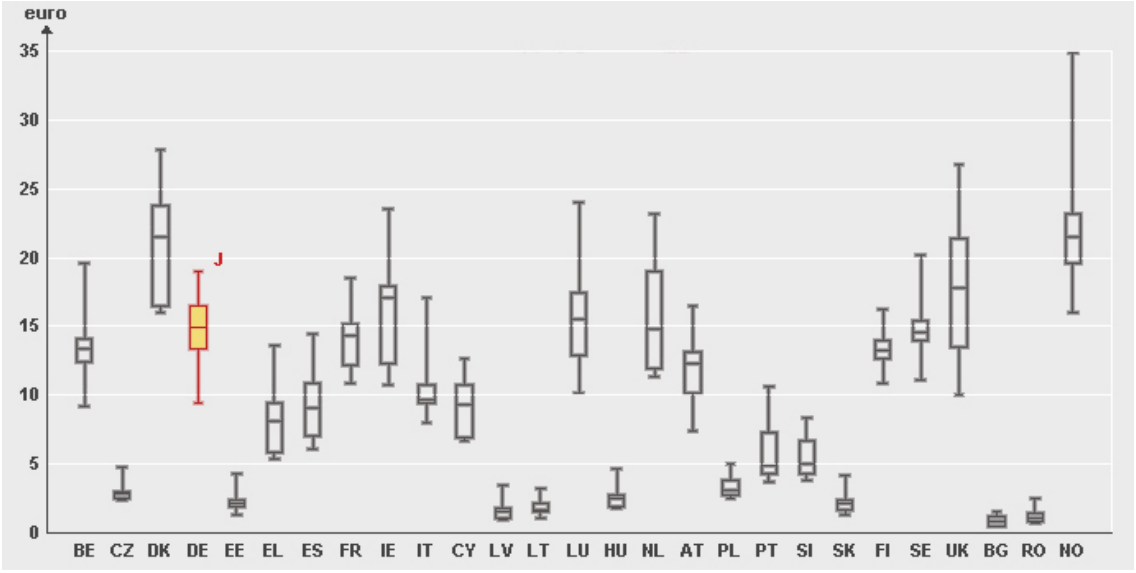

Abb. 5.4: *Streuung von Bruttoverdiensten zwischen den Wirtschaftszweigen in Europa*

veau zwischen den Staaten variiert – mit sehr niedrigen Nivaus in Bulgarien (BG) und Rumänien (RO) und hohen Niveaus in Dänemark (DK) oder Norwegen (NO). Die Grafik kann zum Verständnis der fortschreitenden Arbeitsplatzverlagerungen in Niedriglohnländer im Zuge der Globalisierung beitragen. Starke Verdienstniveauunterschiede in Europa ließen sich allerdings schon aus Abbildung 4.1 ableiten. Die Boxplots liefern aber ein wesentlich differenzierteres Bild als Abbildung 4.1. Man erkennt nämlich hier auch, dass die Spannweite zwischen den Branchen mit minimalen und maximalen Verdiensten von Land zu Land recht unterschiedlich ausfällt (z. B. kleinere Spannweite für Dänemark im Vergleich zu Norwegen). Boxplots mit großer Spannweite und kleinem Quartilsabstand (kürzere Boxen) weisen auf wenig ausgeglichene Einkommensverteilungen hin. Abbildung 5.4, hinter der Individualdaten von Millionen europäischer Arbeitnehmer stehen, illustriert, dass man mit geeigneten Visualisierungsinstrumenten zentrale „Botschaften“ und Auffälligkeiten sichtbar machen kann, die sich aus unüberschaubaren „Zahlenfriedhöfen“ alleine nicht ohne weiteres erschließen lassen.[10]

[10]Eine ausführliche Darstellung von Verdienstunterschieden zwischen europäischen Ländern und Regionen für 2002 findet man bei Mittag (2006), Eurostat-Schriftenreihe „Statistics in Focus“.