Exkurs in die deskriptive Statistik

3

Wie im Einführungsbeispiel in Kap. 1 beginnt praktisch jede statistische Auswertung mit einer Zusammenfassung von Beobachtungen. Dabei geht es darum, die Beobachtungen in einer Grafik oder wenigen Kennzahlen, sogenannten Statistiken, kurz und unmissverständlich zusammenzufassen. Ziel ist es dabei, einen Überblick über die Beobachtungen zu erhalten: Welche Werte werden überhaupt angenommen? ‚Wo' etwa liegen die Beobachtungen? Wie stark ‚streuen' sie? Welche Form hat ihre ‚Verteilung'? usw. Die Beschreibung von Beobachtungen wird auch als *deskriptive Statistik* bezeichnet. Beispielsweise nutzen wir gerne den Mittelwert der Beobachtungen als eine Statistik, die uns Auskunft über die Lage der Beobachtungen gibt.

Neben ihrer Aufgabe, Beobachtungen zu beschreiben, werden Statistiken auch später im Kontext der *statistischen Modellierung* auftauchen. Diese Modellierung wird sich Konzepten aus der Stochastik bedienen, und daher ist es nicht verwunderlich, dass eine Statistik auch immer ein theoretisches Analogon in der Stochastik besitzt. Um beide Welten zu unterscheiden, bezeichnen wir eine Statistik auch häufig als eine *empirische* Kenngröße. So verstehen wir etwa den Mittelwert als eine Statistik, die als empirisches Analogon zum Erwartungswert fungiert.

Skalenniveaus Zur Zusammenfassung von Beobachtungen unterscheidet man zunächst verschiedene Skalenniveaus. *Kategorielle* Beobachtungen nehmen Werte in verschiedenen Kategorien an. Im Einführungsbeispiel in Kap. 1 hatten wir es etwa mit den zwei Kategorien *ja* und *nein* zu tun. Zwischen den Kategorien besteht keine Ordnung – die Aussage „ *ja* ist kleiner als *nein*" ist nicht sinnvoll. *Ordinale* Beobachtungen kann man in eine Ordnung bringen, d. h. bezüglich einer Dimension sortieren. Ein Beispiel wären die Antwortmöglichkeiten bei der Evaluation einer Lehrveranstaltung: ‚Sie waren mit dem Arbeitsklima in der Veranstaltung ‚sehr unzufrieden', ‚unzufrieden', ‚zufrieden', ‚sehr zufrieden'?' Offenbar ist ‚unzufrieden' in gewissem Sinne weniger als ‚zufrieden', allerdings kann man die Abstände zwischen den Kategorien nicht bemessen. *Metrische* Beobachtungen nehmen Werte in den

M. Messer und G. Schneider, *Statistik,* https://doi.org/10.1007/978-3-662-59339-4_3

reellen Zahlen an, wobei die Beziehungen zwischen den Werten durch den euklidischen Abstandsbegriff beschrieben werden können. In diesem Buch werden vor allem Verfahren für metrische, seltener für kategorielle oder ordinale Beobachtungen behandelt.

Im Folgenden diskutieren wir eine Reihe einfacher Statistiken und Darstellungsformen an Beispielen metrischer Beobachtungen. Wir werden sehen, dass die Wahl der Verfahren bis zu gewissem Grad von den Beobachtungen abhängt.

Stripchart und Histogramm Einen Beobachtungsvektor mit metrischen Beobachtungen $\mathbf{x} := \mathbf{x}_n = (x_1, \ldots, x_n)^t \in \mathbb{R}^n$, beispielsweise

$$\mathbf{x} = (3.65, 5.14, 4.11, 4.42, \ldots, 5.23)^t \quad \text{mit } n = 100, \tag{3.1}$$

kann man zum Beispiel in einem *Stripchart* darstellen, siehe Abb. 3.1, in dem jeder Punkt einer Beobachtung x_i entspricht. Zur besseren Sichtbarkeit sind die Punkte vertikal leicht gestreut, der y-Wert hat keine inhaltliche Bedeutung.

Eine andere Form der grafischen Darstellung ist das *Histogramm* (Abb. 3.1b). Die Balkenhöhe entspricht der absoluten Häufigkeit bzw. der Anzahl aller Beobachtungen, die in das entsprechende Intervall fallen. Zum Verständnis sind hier zusätzlich die rohen Beobachtungen $\mathbf{x}$ als Punkte dargestellt. In einem Histogramm gehen nur wenige Informationen verloren, nämlich die genaue Lage der Punkte innerhalb der Balken. Fallen Beobachtungen auf Balkenränder, so sollte man spezifizieren, ob sie dem linken oder rechten Balken zugeordnet werden, oder gegebenenfalls die Balkenränder anders positionieren. In der Abbildung fällt keine Beobachtung auf einen Rand. Wir stellen fest, dass sich die Beobachtungen etwa glockenförmig, d. h. eingipflig und ungefähr symmetrisch, verteilen.

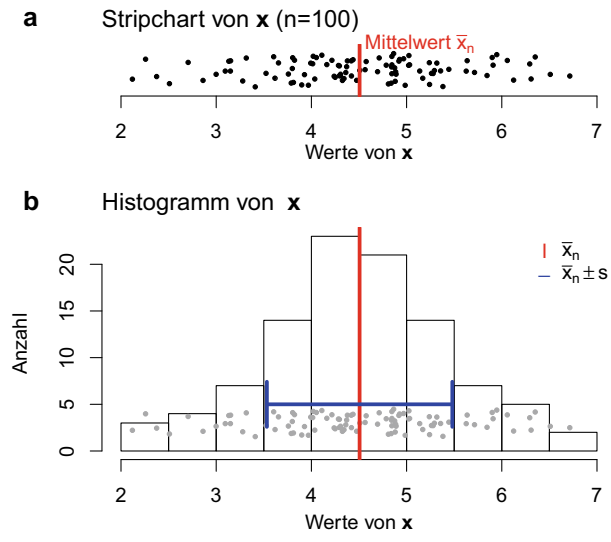

Abb. 3.1 Darstellung des Beobachtungsvektors $\mathbf{x}$. **a** Stripchart. Der rote Balken markiert den Mittelwert $\bar{x}_n$. **b** Histogramm. In Rot ist der Mittelwert $\bar{x}_n$ markiert. Der blaue Bereich markiert eine Standardabweichung um den Mittelwert $[\bar{x}_n - s(\mathbf{x}), \bar{x}_n + s(\mathbf{x})]$

Mittelwert und Standardabweichung Zur Zusammenfassung der Beobachtungen $\mathbf{x}$ nutzt man geeignete *Statistiken.* Diese verstehen wir als Funktionen von $\mathbf{x}$. Hier ist $\mathbf{x} \in \mathbb{R}^n$, und in diesem Fall denken wir bei einer Statistik an eine Abbildung vom $\mathbb{R}^n$ in die reellen Zahlen.

Zur Zusammenfassung von Beobachtungen, die sich wie in Abb. 3.1 etwa glockenförmig verteilen, eignen sich der Mittelwert und die empirische Standardabweichung. Den *Mittelwert*

$$\bar{x}_n := \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.2}$$

kann man geometrisch wegen $\sum_i (x_i - \bar{x}_n) = 0$ als Schwerpunkt der Beobachtungen interpretieren: Stellen wir uns zum Beispiel in Abb. 3.1 die x-Achse als eine Waage vor, auf der alle Punkte gleiches Gewicht besitzen, dann ist $\bar{x}_n$ derjenige Drehpunkt, bei dem die Waage im Gleichgewicht ist. Wir können den Mittelwert also nicht nur mit Gl. (3.2) berechnen, sondern auch direkt per Auge aus der Grafik abschätzen. Hier erkennen wir auch den Zusammenhang zur Welt des Zufalls, denn für eine integrierbare Zufallsvariable, die eine Dichte oder Gewichte besitzt, ist der Erwartungswert derjenige Drehpunkt, der die Dichte bzw. die Gewichte im Gleichgewicht hält. Daher verstehen wir den Mittelwert als empirisches Analogon des Erwartungswertes.

Die *empirische Varianz* $s^2 := s_n^2$ ist für $n > 1$ definiert durch

$$s^2(\mathbf{x}) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2. \tag{3.3}$$

Ihre Wurzel s heißt *empirische Standardabweichung.* Die empirische Varianz verstehen wir als Analogon der Varianz einer Zufallsvariablen. Aufgrund des Faktors $1/(n-1)$ anstatt $1/n$ spricht man oft auch von der *korrigierten* empirischen Varianz bzw. Standardabweichung. Die Korrektur hat einen gewissen theoretischen Vorteil, wie wir in Kap. 5 sehen werden, sie macht aber bei großem Stichprobenumfang wenig aus.

Sind die Beobachtungen näherungsweise glockenförmig verteilt, so lassen sie sich sinnvoll durch den Mittelwert und die empirische Varianz zusammenfassen, siehe Abb. 3.1. Der Mittelwert $\bar{x}_n$ gibt eine gute Vorstellung von der ungefähren Lage einer ‚typischen Beobachtung', und die empirische Standardabweichung $s(\mathbf{x})$ hat die schöne Interpretation einer ‚typischen Abweichung' vom Mittelwert $\bar{x}_n$ und lässt sich auch naiv aus der Grafik schätzen. Denn analog zur glockenförmigen Normalverteilung, wo sich etwa 2/3 der Masse in der Umgebung einer Standardabweichung um den Erwartungswert sammeln, liegen etwa 2/3 der Beobachtungen eine empirische Standardabweichung vom Mittelwert entfernt, also in dem Intervall $[\bar{x}_n - s(\mathbf{x}), \bar{x}_n + s(\mathbf{x})]$. So finden wir in Abb. 3.1 per Auge, dass $\bar{x}_n \approx 4.5$ und $s(\mathbf{x}) \approx 1$, denn die imaginäre Waage ist ungefähr bei 4.5 im Gleichgewicht, und etwa zwei Drittel der Beobachtungen lassen sich im Intervall [3.5, 5.5] einfangen.

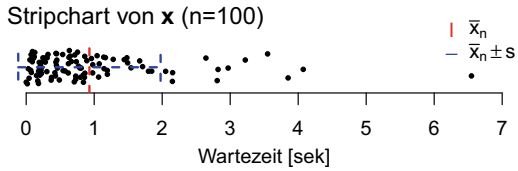

Abb. 3.2 Darstellung der Wartezeiten $\mathbf{x}$ im Stripchart, mit Mittelwert $\bar{x}_n$ (rot) und Standardabweichung $[\bar{x}_n - s(\mathbf{x}), \bar{x}_n + s(\mathbf{x})]$ (blau). Hier sind $\bar{x}_n$ und $s(\mathbf{x})$ keine sinnvollen Statistiken zur Beschreibung der Beobachtungen

Zur Zusammenfassung *nicht*-glockenförmig verteilter Beobachtungen sind Mittelwert und Standardabweichung nur begrenzt geeignet. Wir betrachten dazu ein zweites Beispiel: Es wurde einhundert Mathematikern je eine Additionsaufgabe im Zehnerbereich präsentiert und die Wartezeit bis zur Lösung der Aufgabe gemessen. Die Wartezeiten sind im Vektor $\mathbf{x} = (x_1, \ldots, x_{100})^t$ zusammengefasst und in Abb. 3.2 dargestellt. Die Verteilung ist asymmetrisch, die meisten Beobachtungen sind kleiner als der Mittelwert. Wenige große Werte ziehen den Mittelwert nach rechts. Man kann also nicht mehr davon sprechen, dass der Mittelwert etwa die Größe einer typischen Beobachtung hat. Zudem wird die Standardabweichung $s(\mathbf{x})$ durch die wenigen großen Werte vergleichsweise groß, sodass nicht mehr 2/3, sondern fast alle Werte innerhalb des Intervalls $[\bar{x}_n - s(\mathbf{x}), \bar{x}_n + s(\mathbf{x})]$ liegen. Wiederum ist $s(\mathbf{x})$ also keine ‚typische Abweichung' mehr vom Mittelwert, denn fast alle Abweichungen sind kleiner.

Empirische Verteilungsfunktion und empirische Quantile Um die Beobachtungen aus Abb. 3.2 zu beschreiben, führen wir die Begriffe der empirischen Verteilungsfunktion und der empirischen Quantile ein. Auch diese verstehen wir als empirische Analogien der entsprechenden Begriffe einer Zufallsvariablen oder einer Verteilung, siehe Gl. (2.5) und Definition 2.3.

Definition 3.1 (Empirische Verteilungsfunktion)
Es sei $\mathbf{x} = (x_1, \ldots, x_n)^t \in \mathbb{R}^n$. *Die empirische Verteilungsfunktion* $\hat{F} := \hat{F}^{(\mathbf{x})}$ *von* $\mathbf{x}$ *ist eine Abbildung* $\hat{F} : \mathbb{R} \to \{0, 1/n, \ldots, (n-1)/n, 1\}$ *via*

$$\hat{F}(z) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, z]}(x_i).$$

Die Größe $\hat{F}(z)$ beschreibt also den Anteil der Beobachtungen von $\mathbf{x}$, die nicht größer als z sind. $\hat{F}$ ist rechtsstetig, da das Intervall $(-\infty, z]$ rechts abgeschlossen ist.

Beispiel 3.2 (Empirische Verteilungsfunktionen)

i. Es sei $\mathbf{x} = (0, 1, 1, 2)^t$. *Dann ist die empirische Verteilungsfunktion* $\hat{F}$ *von* $\mathbf{x}$ *gerade die Verteilungsfunktion* F *einer Zufallsvariablen* X *mit* $X \sim b(2, 1/2)$, *vgl. Abb. 2.2b.*

ii. Allgemein gilt: Es sei $\mathbf{x} = (x_1, \ldots, x_n)^t$ *ein Beobachtungsvektor. Dann ist eine diskrete Verteilung* $\nu_{\mathbf{x}}$ *mit Gewichten in* $\{x_1, \ldots, x_n\}$ *gegeben durch* $\nu_{\mathbf{x}}((a, b]) := (1/n) \sum_{i=1}^{n} \mathbb{1}_{(a,b]}(x_i)$. *Die empirische Verteilungsfunktion* $\hat{F}$ *von* $\mathbf{x}$ *ist dann gerade die Verteilungsfunktion* F *von* $\nu_{\mathbf{x}}$.

Diese Gleichheit von $\hat{F}$ und F hat zur Folge, dass die folgenden Begriffe rund um empirische Quantile in direkter Analogie zu den Begriffen bzgl. der Quantile einer Zufallsvariable stehen.

Definition 3.3 (Empirisches Quantil)

Es sei $\mathbf{x} = (x_1, \ldots, x_n)^t$ *ein Beobachtungenvektor und* $p \in (0, 1)$. *Eine reelle Zahl* q_p *heißt ein p-Quantil (der empirischen Verteilung) von* $\mathbf{x}$, *wenn gilt*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, q_p]}(x_i) \geq p \quad \text{und} \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[q_p, \infty)}(x_i) \geq 1 - p.$$

Wir nennen q_p auch kurz ein empirisches p-Quantil von $\mathbf{x}$. Die Interpretation entspricht der des Quantils einer Zufallsvariablen: Der Anteil der x_i, die kleiner oder gleich q_p sind, ist mindestens p, und der Anteil der x_i, die größer oder gleich q_p sind, ist mindestens $1 - p$. Auch bezüglich eines Beobachtungsvektors $\mathbf{x}$ bildet die Menge $\hat{Q}_p$ aller p-Quantile ein Intervall $\hat{Q}_p(\mathbf{x}) := [\hat{q}_p^-(\mathbf{x}), \hat{q}_p^+(\mathbf{x})]$ mit $\hat{q}_p^-(\mathbf{x}) := \sup\{z \in \mathbb{R} | \hat{F}(z) < p\}$ und $\hat{q}_p^+(\mathbf{x}) := \inf\{z \in \mathbb{R} | \hat{F}(z) > p\}$, vgl. Gl. (2.6). Analog zu $\mathcal{P}_q$ können wir anhand der empirischen Verteilungsfunktion $\hat{F}$ die Menge $\hat{\mathcal{P}}_q$ aller $p \in (0, 1)$ definieren, für die q ein empirisches p-Quantil bzgl. $\mathbf{x}$ ist. Dafür ersetze in Gl. (2.7) F durch $\hat{F}$.

Beispiel 3.4 (Empirische Quantile)

i. Es sei $\mathbf{x} = (0, 1, 1, 2)^t$. *Dann gilt für alle* $p \in (0, 1)$, *dass die Menge* $\hat{Q}_p$ *bzgl.* $\mathbf{x}$ *gerade der Menge* Q_p *bzgl.* $b(2, 1/2)$ *gleicht. Dies ist eine unmittelbare Konsequenz aus Beispiel 3.2i.*

ii. Analog folgt: Es sei $\mathbf{x} = (x_1, \ldots, x_n)^t$ *ein Beobachtungenvektor und* $\nu_{\mathbf{x}}$ *die diskrete Verteilung aus Beispiel 3.2ii. Aus der Gleichheit von* $\hat{F}$ *und* F *folgt dann, dass die Quantile der empirischen Verteilung bzgl.* $\mathbf{x}$ *gerade die Quantile von* $\nu_{\mathbf{x}}$ *sind.*

Sprechen wir von *dem* p-Quantil $\hat{q}_p(\mathbf{x})$, so meinen wir hier den eindeutigen Mittelwert des Intervalls $\hat{Q}_p(\mathbf{x})$, d. h.

$$\hat{q}_p(\mathbf{x}) := (\hat{q}_p^-(\mathbf{x}) + \hat{q}_p^+(\mathbf{x}))/2.$$

Diese Definition ist aber nicht einheitlich; so wird manchmal zum Beispiel auch $\hat{q}_p^-$ als eindeutiges empirisches p-Quantil festgelegt. Eine sehr prominente Statistik ist *der empirische Median* M von $\mathbf{x}$, hier gegeben durch

$$M(\mathbf{x}) := \frac{\hat{q}_{1/2}^-(\mathbf{x}) + \hat{q}_{1/2}^+(\mathbf{x})}{2} = \begin{cases} x_{((n+1)/2)}, & \text{falls } n \text{ ungerade,} \\ [x_{(n/2)} + x_{((n/2)+1)}]/2, & \text{falls } n \text{ gerade.} \end{cases} \tag{3.4}$$

Dabei bezeichnet $(x_{(1)}, \ldots, x_{(n)})^t$ die *Ordnungsstatistik* von $\mathbf{x}$, d. h. die der Größe nach geordnete Stichprobe $x_{(1)} \leq \ldots \leq x_{(n)}$.

Analog definieren wir *das* empirische erste Quartil $\hat{q}_{0.25}(\mathbf{x})$, und *das* empirische dritte Quartil $\hat{q}_{0.75}(\mathbf{x})$. Im Rahmen der empirischen Quantile verstehen wir den Median und den Interquartilsabstand $(\hat{q}_{0.75}(\mathbf{x}) - \hat{q}_{0.25}(\mathbf{x}))$ als Maße für die Lage bzw. die Variabilität der Beobachtungen.

Boxplot und Q-Q-Plot Auch zur grafischen Darstellung nicht-glockenförmig verteilter Beobachtungen könnten wir wieder einen Stripchart oder ein Histogramm heranziehen. Eine andere Form der grafischen Darstellung, die sich der empirischen Quantile bedient, ist der *Boxplot* (auch Box-and-Whisker-Plot genannt, vgl. Abb. 3.3). Die Box fängt die mittleren 50 % der Beobachtungen ein. Genauer bildet der linke Rand das erste Quartil $\hat{q}_{0.25}(\mathbf{x})$ und der rechte Rand das dritte Quartil $\hat{q}_{0.75}(\mathbf{x})$. Der vertikale Balken innerhalb der Box ist der Stichprobenmedian. Die vertikalen Linien außerhalb der Box heißen *Whisker* und können unterschiedlich positioniert werden. In Variante Grün markieren sie die minimale bzw. maximale Beobachtung. In einer anderen häufig verwendeten Variante (Rot) ist die maximale Länge der Whisker begrenzt durch den 1,5-fachen Interquartilsabstand, $1.5 \cdot (\hat{q}_{0.75}(\mathbf{x}) - \hat{q}_{0.25}(\mathbf{x}))$, bzw. genauer durch die extremste Beobachtung innerhalb dieses Bereichs. Offenbar ist der Boxplot eine einfache Form der grafischen Darstellung, die auf nur fünf Statistiken beruht. Sie wird gerne zur Beschreibung asymmetrisch verteilter Beobachtungen herangezogen.

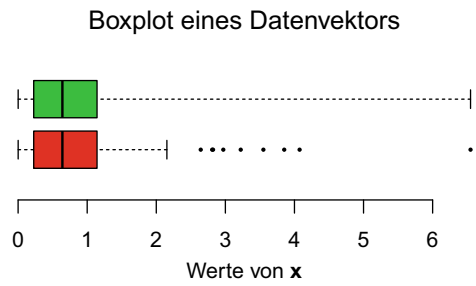

Abb. 3.3 Darstellung des Beobachtungsvektors $\mathbf{x}$ aus Abb. 3.2 im Boxplot

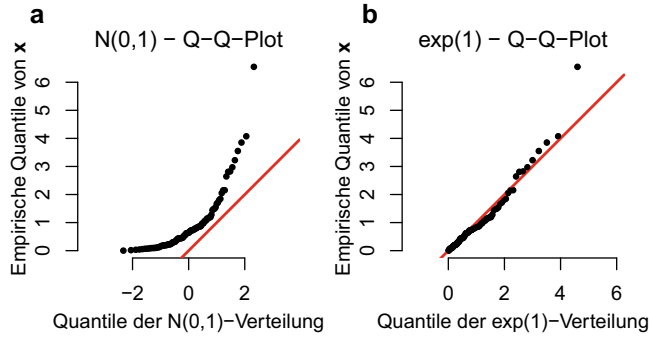

Abb. 3.4 Q-Q-Plot zum Vergleich der Beobachtungen **x** aus Abb. 3.2 mit einer theoretischen Verteilung, a: $N(0, 1)$ und b: $exp(1)$. Die rote Gerade markiert die Hauptdiagonale

Eine andere Form der grafischen Darstellung, die auch auf Quantilen beruht, bietet der sogenannte *Q-Q-Plot*. Hierbei werden die Beobachtungen $\mathbf{x} \in \mathbb{R}^n$ verglichen entweder mit einer reellwertigen Verteilung ν oder mit anderen Beobachtungen $\mathbf{y} \in \mathbb{R}^m$. Dafür wird das empirische p-Quantil von **x** gegen das theoretische p-Quantil von ν bzw. gegen das empirische p-Quantil von **y** aufgetragen. Dabei durchläuft p meist eine Menge äquidistanter Punkte zwischen 0 und 1. Die Idee ist, dass die so entstandene Menge von Punkten etwa auf der Hauptdiagonalen liegen sollte, falls die den Beobachtungen zugrunde liegende Verteilung mit der verglichenen Verteilung identisch ist. In Abb. 3.4 sind zwei Q-Q Plots dargestellt, in denen die Beobachtungen aus Abb. 3.2 mit der $N(0, 1)$- (Abb. 3.4a) und der $exp(1)$-Verteilung (Abb. 3.4b) verglichen werden. Sowohl die konvexe Struktur als auch die Abweichung von der Hauptdiagonalen (rot) in Abb. 3.4a signalisiert eine Unverträglichkeit der Verteilung der Beobachtungen **x** mit der Standardnormalverteilung, was auch zur Asymmetrie der Verteilung von **x** in Abb. 3.2 passt. Weniger unverträglich sind die Beobachtungen mit der $exp(1)$-Verteilung, denn in Abb. 3.4b liegen die Punkte sehr nahe an der Hauptdiagonalen. Umfangreiche Diskussionen von Q-Q-Plots finden sich zum Beispiel bei Welch und Gnanadesikan (1968) oder Dümbgen (2015).