

*My sources are unreliable, but their information is fascinating.  
Ashleigh E. Brilliant*

## Am Ende dieses Kapitels ...

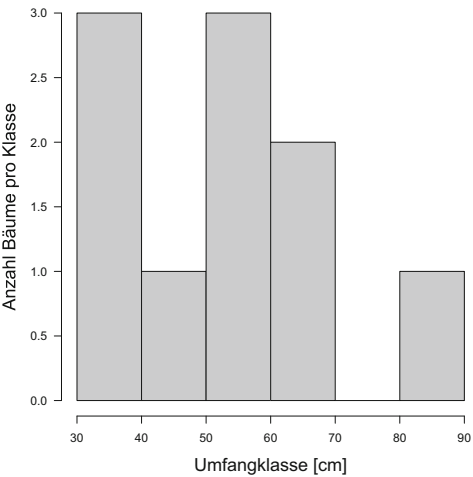
- ... sollte Dir klar sein, was eine Stichprobe ist.
- ... sollten Dir die Begriffe Mittelwert, Median, Standardabweichung, Varianz, Standardfehler so geläufig sein, dass Du die Formeln dafür auswendig kannst. Schiefe, Kurtosis, Interquartilabstand sollten bekannt klingen.
- ... sollte Dir ein Histogramm als einfache und nützliche Abbildung vertraut sein.
- ... solltest Du verstehen, was eine Häufigkeitsverteilung ist.
- ... solltest Du den Unterschied zwischen Häufigkeit und Dichte eines Messwerts kennen.
- ... solltest Du nun aber wirklich einmal etwas praktisch machen wollen.

Ausgangspunkt aller statistischen Analysen ist ein Datensatz, bestehend aus einzelnen Datenpunkten. Wenn wir z. B. den Umfang eines Parkbaumes messen, so kann dies ein Datenpunkt sein. *Muss* dieser Baum diesen Umfang haben? Natürlich nicht. Wenn er früher oder später gepflanzt worden wäre, weniger oder mehr Licht oder Dünger erhalten hätte, weniger Menschen ihre Initialen in die Borke geritzt hätten, dann wäre der Baum dicker oder dünner. Aber gleichzeitig hätte er nicht beliebig dick oder dünn sein können: eine Esche wird nun mal keine 4 m dick. Worauf ich hinaus will ist, dass ein Datenpunkt *eine* Realisierung von vielen möglichen ist. Was wir aus vielen Messungen herausbekommen können ist eine Erwartung, wie dick typischerweise eine 50 Jahre alte Parkesche ist. Aber jede einzelne Parkesche ist natürlich nicht genau *so* dick.

Mit vielen Worten habe ich hier die Idee der *Zufallsvariable* beschrieben: eine Zufallsvariable ist eine zufällige Größe, die verschiedene Werte haben kann. Jeder Datensatz

**Tab. 1.1** Brusthöhendurchmesser von 100 Eschen in cm, gemessen von 10 verschiedenen Gruppen

Gruppe 1	Gruppe 2	Gruppe 3	Gruppe 4	Gruppe 5	Gruppe 6	Gruppe 7	Gruppe 8	Gruppe 9	Gruppe 10
37	54	33	82	57	34	60	65	62	44
80	58	38	6	72	49	50	69	66	62
68	66	51	10	62	49	47	21	40	58
77	48	58	49	22	42	42	49	72	65
47	45	64	61	36	36	57	65	48	68
58	38	57	27	79	90	43	29	61	47
98	49	64	51	35	54	14	79	53	93
60	36	62	31	25	56	41	50	51	38
39	47	74	20	62	57	71	44	57	55
39	74	73	64	82	61	24	39	26	41



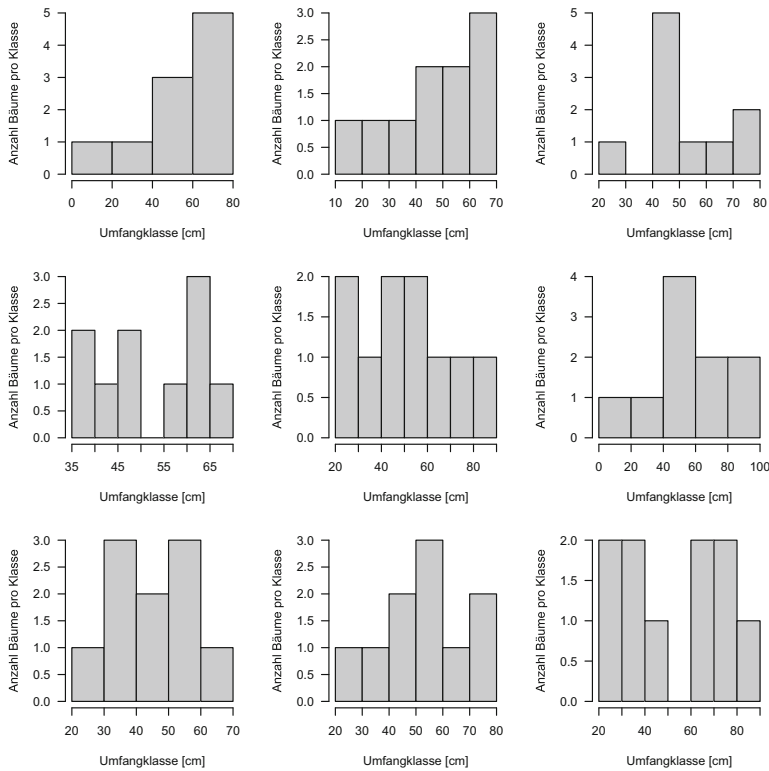
**Abb. 1.1** Histogramm der Eschenumfänge von Gruppe 1

ist eine mögliche Realisierung einer solchen Zufallsvariablen. Schicken wir also einmal 10 Gruppen imaginärer StudentInnen in einen Park. Dort soll jede Gruppe 10 Eschen in Brusthöhe vermessen. Diese 10 Eschen sind dann 10 zufällige Beobachtungen der Zufallsvariablen „Eschenumfänge“. Wir bezeichnen solche Gruppe an Ausprägungen als *Stichprobe*.<sup>1</sup>

Die Ergebnisse der 10 mal 10 Messungen (= Beobachtungen) sind in Tab. 1.1 angegeben.

Wir haben also 100 Realisierungen der Zufallsvariablen „Eschenumfang“. Eine häufige und nützliche Form der Darstellung solcher Daten ist ein Histogramm (Abb. 1.1).

<sup>1</sup> Ausprägung, Beobachtung und Realisierung sind hier Synonyme. Realisierung ist am technischsten, Beobachtung vielleicht am intuitivsten. In jedem Fall ist die Zufallsvariable ein hypothetisches Konstrukt, von dem wir nur seine Realisierungen beobachten und stichprobenhaft messen können.



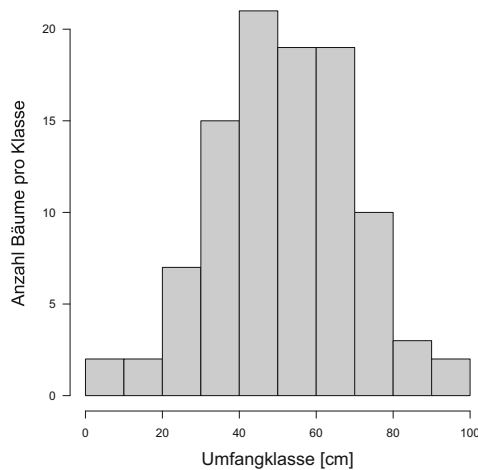
**Abb. 1.2** Histogramm der Eschenumfänge von Gruppen 2 bis 10

In diesem Histogramm sind die Bäume in Umfangsklassen eingeteilt worden, hier in 10 cm weite Klassen.<sup>2</sup> Die Klassen sind alle gleich groß. Wären sie es nicht, könnte man die Höhe der Säulen nicht unmittelbar vergleichen. Wieviele Bäume sich in jeder Klasse befinden ist dann auf der y-Achse aufgetragen. Diese Form von Histogramm nennt man ein Häufigkeithistogramm, da gezählt wurde, wie viele Elemente jede Klasse enthält. Viele Statistikprogramme nehmen die Einteilung in *bins* automatisch vor, aber man kann das natürlich auch selbst steuern. Zu viele Klassen haben keinen Sinn, da dann in jeder nur 0 oder 1 Baum steckt. Zu wenige Klassen sind auch informationsarm, weil sich dann in den Klassen jeweils ein weiter Bereich unterschiedlicher Bäume befinden würde. Verlassen wir uns also auf den Algorithmus der Software.<sup>3</sup>

Die Histogramme der anderen Gruppen sehen deutlich anders aus (Abb. 1.2).

<sup>2</sup> Diese Histogrammklassen heißen im Englischen *bins* und das Einteilen heißt tatsächlich *to bin*.

<sup>3</sup> Der in R benutzte Algorithmus geht auf die Sturges-Formel zurück: es gibt  $k$  *bins*, mit  $k = \lceil 1 + \log_2 n \rceil$ , für  $n$  Datenpunkte ( $\lceil \cdot \rceil$  bedeuten Aufrunden). R variiert dies allerdings, anscheinend werden statt  $n$  die Anzahl eindeutiger Werte benutzt. Für weniger als 30 Datenpunkte kann das Ergebnis schon mal hässlich sein. Dann gibt es u. a. noch Scotts Vorschlag:  $k = \frac{3.5\sigma}{n^{1/3}}$ . Da wir die Standardabweichung  $\sigma$  aber noch nicht kennengelernt haben, führt das hier zu weit.



**Abb. 1.3** Histogramm aller 100 Eschenumfänge

Wir sehen auf den ersten Blick, dass die Säulen alle etwas anders angeordnet sind. Auf den zweiten Blick sehen wir, dass die Säulen unterschiedlich hoch sind (die y-Achsen sind jeweils anders skaliert) und die Umfangsklassen sind unterschiedlich breit. Was wir also merken ist, dass jeder dieser 10 Datensätze etwas anders ist und einen etwas anderen Eindruck über die Eschendicken vermittelt.

Was wir hoffentlich auch sehen ist, dass Histogramme eine sehr gute und schnelle Methode sind, um die Daten zusammenzufassen. Wir sehen auf einen Blick den Wertebereich, wo die Mitte liegt, usw. So können wir im links-oberen Histogramm von Abb. 1.2 sehen, dass es anscheinend einen Baum mit weniger als 20 cm Umfang gab. In Tab. 1.1 sehen wir, dass der Wert tatsächlich 6 cm ist. So entdecken wir außergewöhnliche Werte, die z. B. durch Fehler bei der Dateneingabe oder beim Messen im Feld oder beim Diktieren ins Messbuch entstanden sein können. In diesem Fall ist es einfach ein dünnes Bäumchen: es war wirklich so!

Nun war der Park groß, die Gruppen liefen sich nicht über den Weg und schrieben nicht voneinander ab.<sup>4</sup> Das heißt, wir haben eigentlich Messwerte von 100 Eschen, nicht nur von 10. Das Histogramm aller Daten sieht so aus (Abb. 1.3).

Jetzt sieht das Ganze schon anders aus. Die mittleren Umfänge sind deutlich häufiger als die extremen. Die Säulen steigen hübsch monoton an und fallen dann monoton ab, kein Rauf und Runter wie in Abb. 1.2. (Eine Alternative zum Histogramm, den Boxplot, werden wir weiter unten kennenlernen, da er auf Stichprobenstatistiken aufbaut.)

Die Visualisierung von Daten ist das mächtigste Verfahren, uns diese vor Augen zu führen. Natürlich können wir ganz verschiedene beschreibende Werte berechnen (mehr dazu gleich), aber eine Abbildung sagt mehr als 100 Werte. Hier erkennen wir, ob Daten schief verteilt sind, welche Werte wie häufig sind, ob Wertebereiche fehlen oder unterrepräsentiert sind, wie weit die Werte im Extrem streuen und vieles mehr. Unser Hirn ist famos

<sup>4</sup> Unrealistisch, ich weiß, aber es sind ja imaginäre StudentInnen.

darin, solche Muster in Windeseile abzuchecken. Das Produzieren von Abbildungen ist (vor allem in der explorativen Phase) dem von Zahlen und Statistiken weit vorzuziehen!

## 1.1 Stichprobenstatistiken

Zur Beschreibung von Stichproben mit Hilfe von Statistiken haben sich einige verschiedene Standards eingebürgert (siehe Tab. 1.2). Grundsätzlich unterscheiden wir Statistiken für den „zentralen“ Wert (*location*) und für die Streuung der Stichprobe (*spread*). Übliche Maße für den zentralen Wert sind z. B. der Mittelwert und der Median. Für die Streuung der Werte einer Stichprobe können wir z. B. deren Standardabweichung oder Varianzkoeffizienten berechnen. Diese Maße sind essentiell und jeder, der in irgendeiner Form Statistik macht, muss sie kennen. Die weiteren hier vorgestellten Maße für Zentralität und Streuung sind für bestimmte Zwecke nützlich und werden deshalb von bestimmten Disziplinen angewandt. Sie sind vor allem der Vollständigkeit halber hier aufgeführt.

### 1.1.1 Zentralitätsmaße

Um eine typische Esche zu beschreiben, ist ein Wert besonders wichtig: der **Mittelwert** (englisch: *mean*). Er ist die häufigste und wichtigste Stichprobenstatistik. Sein Wert berechnet sich für eine Reihe von  $n$  gemessenen Werten ( $x_i$ ) wie folgt:

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

(1.1)

**Tab. 1.2** Eine Zusammenstellung möglicher Stichprobenstatistiken, ihrer deutschen und englischen Namen (nach Quinn und Keough, 2002, ergänzt)

Deutscher Name	Englischer Name	Abk.
Mittelwert	<i>mean</i>	$\bar{x}$
Median	<i>median</i>	
Modus	<i>mode</i>	
Hubers Mittelwert Schätzer	<i>Huber's mean estimate</i>	
Varianz	<i>variance</i>	$s^2$
Standardabweichung	<i>standard deviation</i>	$s$
Standardfehler des Mittelwertes	<i>standard error of the mean</i>	
Mittlere absolute Abweichung	<i>mean absolute deviation</i>	
Varianzkoeffizient	<i>coefficient of variation</i>	CV
95 % Konfidenzintervall	<i>confidence interval</i>	CI
95 % Quantilen	<i>95 % quantiles</i>	
Interquartilabstand	<i>interquartile range</i>	IQR
Spanne	<i>range</i>	
Schiefte	<i>skewness</i>	
Kurtosis	<i>kurtosis</i>	

Für die ersten 10 Eschenwerte (Stichprobenumfang  $n = 10$ ) wäre der Mittelwert  $\bar{x} = (37 + 54 + 33 + 82 + 57 + 34 + 60 + 65 + 62 + 44)/10 = 52.8$ .

Während der Mittelwert den Durchschnittswert repräsentiert, ist der **Median** der mittlere Wert in dem Sinn, dass die Hälfte aller Werte größer und die Hälfte kleiner ist. Abhängig davon, ob es eine gerade oder ungerade Anzahl Datenpunkte gibt, wird er wie folgt berechnet:

$$\text{Median} = \begin{cases} x'_{(n+1)/2}, & \text{für ungerade } n; \\ (x'_{n/2} + x'_{(n/2)+1})/2, & \text{für gerade } n. \end{cases} \quad (1.2)$$

Dabei bedeutet  $x'_n$  den  $n$ -ten Wert der Größen-sortierten (= geordneten) Zahlenwerte von  $x$ .  $x'_{(n+1)/2}$  ist also gerade der mittlere Wert.<sup>5</sup>

Der Median ist also ein gemessener Datenwert (für ungerade  $n$ ) bzw. das Mittel der zwei mittleren Werte (für gerade  $n$ ).

Der Median ist ein Spezialfall des *Quantils*. Ein  $p$ -Quantil teilt eine Stichprobe in zwei Teile, den mit den Werten die kleiner sind als  $100 \cdot (1 - p)$  Prozent der beobachteten Werte, und die die größer oder gleich diesem Wert sind. Das 10 %-Quantil ist also der Wert, an dem gerade 10 % der Stichprobenwerte kleiner sind.<sup>6</sup> Der Median ist entsprechend das 0.5-Quantil.

Schließlich gibt es noch als Zentralitätsmaß den **Modus** (*mode*). Das ist schlicht der am häufigsten auftauchende Wert (bzw. die am häufigsten auftretenden Werte) einer Stichprobe.

Weitaus seltener benutzt werden robuste Schätzer (Huber, 1981). Sie ersetzen den Mittelwert, indem sie Ausreißer heruntergewichten. Robuste Statistiken sind etwas konservativer, also nicht so leicht von starken Abweichungen beeinflusst. Trotz ihrer Nützlichkeit sind sie in der angewandten statistischen Literatur kaum vorhanden.

### 1.1.2 Maße für Streuung

Neben einem Maß für die Zentralität brauchen wir auch Statistiken, die die Streuung beschreiben. Das Wichtigste, analog zum Mittelwert  $\hat{x}$ , ist die **Standardabweichung**:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (1.3)$$

<sup>5</sup> Ein Beispiel: Unsere Messwerte sind  $x = (4; 2; 7; 1; 9)$ . Geordnet sind sie dann  $x' = (1; 2; 4; 7; 9)$ .  $x'_1$  hätte dann den Wert 1,  $x'_4 = 7$ . Der Median ist dann entsprechend  $x'_{(5+1)/2} = x'_3 = 4$ .

<sup>6</sup> Die Berechnung erfordert dabei häufig Interpolationen, und die Statistikprogramme unterscheiden sich z. T. sehr stark in der Art und Weise, wie Quantilen berechnet werden; siehe nächstes Kapitel für Beispiele dazu.

Wenn die Daten einer normalverteilten Zufallsvariable entsprechen, oder anders formuliert, wenn „die Daten normalverteilt sind“, dann liegen etwa 68 % der Datenwerte  $\pm 1$  Standardabweichung und etwa 95 %  $\pm 2$  Standardabweichungen um den Mittelwert. Bei schiefen Histogrammen (die ja gerade *nicht* normalverteilt sind) müsste die Standardabweichung in einer Richtung größer sein, als in der anderen. Deshalb gelten diese Prozentsätze nur für normalverteilte Daten.

Ihr Pendant ist die **Varianz**,  $s^2$ , die einfach das Quadrat der Standardabweichung ist.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (1.4)$$

Die Varianz wird selten zur Beschreibung von Stichproben genommen, weil ihre Dimensionen nicht intuitiv sind. Die Standardabweichung hat die gleiche Dimension wie der Mittelwert (also etwa cm für unsere Eschenumfänge), während die Varianz deren Quadrat hat (cm<sup>2</sup>).

Eine wirkliche Vergleichbarkeit erreicht man, wenn man die Standardabweichung  $s$  mit dem Mittelwert  $\bar{x}$  standardisiert.

$$CV = \frac{s}{\bar{x}} \quad (1.5)$$

Dieses Maß, der **Varianzkoeffizient** (*coefficient of variation*, CV), ist direkt zwischen Datensätzen vergleichbar, weil unabhängig von den absoluten Werten der Stichprobe. Für kleine Stichproben hat der  $CV$  einen systematischen Fehler (engl.: *biased*). Die korrigierte Version ist:

$$CV^* = \left(1 + \frac{1}{4n}\right) \frac{s}{\bar{x}} \quad (1.6)$$

Die robuste, aber selten benutzte Variante beruht auf dem Verhältnis von Interquartilabstand und Median:

$$\text{Quartilen-Dispersionkoeffizient} = \frac{Q_3 - Q_1}{Q_3 + Q_1}, \quad (1.7)$$

wobei  $Q_1$  und  $Q_3$  die erste und dritte Quartile sind (= 25 und 75 % Quantil).

Ein anderes, häufig berichtetes Stichprobenmaß ist der **Standardfehler des Mittelwertes**, sem.<sup>7</sup> Hierbei handelt es sich **nicht** um ein Maß für die Streuung der Stichprobe! Der Standardfehler des Mittelwertes beschreibt vielmehr die Genauigkeit der Berechnung des Stichproben-Mittelwertes. Wenn wir viele Datenpunkte haben, dann können wir uns auch

---

<sup>7</sup> Im Englischen ist es schon länger üblich nicht mehr vom *standard error* zu sprechen, sondern nur noch vom *standard error of the mean*. Das sollten wir uns im Deutschen auch angewöhnen.

des Mittelwertes ziemlich sicher sein, bei wenigen hingegen nicht. Der Standardfehler des Mittelwertes quantifiziert dies:

$$\text{sem} = \frac{s}{\sqrt{n}} \quad (1.8)$$

Er wird ähnlich interpretiert wie die Standardabweichung, aber eben nicht in Bezug auf die Stichprobe, sondern in Bezug auf den daraus berechneten Mittelwert. Der wahre (aber unbekannte) Mittelwert liegt mit 95 % Wahrscheinlichkeit  $\pm 2$  Standardfehler um  $\bar{x}$ . Das ist *nicht* das Gleiche wie zu behaupten, dass 95 % der *Datenpunkte*  $\pm 2$  Standardabweichungen um den Mittelwert liegen! Insofern passt der Standardfehler nicht zu den Streuungsmaßen, sondern gehört eigentlich zum Mittelwert als dessen Genauigkeitsmaß.

Da dieser Punkt häufig falsch verstanden wird, hier ein kurzes Beispiel. Unsere 10 Gruppen haben jeweils 10 Eschen gemessen. Jede Gruppe kann einen Mittelwert berechnen, seine Standardabweichung und den Standardfehler des Mittelwertes. Der Mittelwert aller 100 Messungen soll unseren wahren Mittelwert darstellen. Für Gruppe 1 ist  $\bar{x}_1 = 60.3$ ,  $s_1 = 20.49$  und  $se_1 = 20.49/\sqrt{10} = 6.48$ . Wir erwarten mit 95 % Wahrscheinlichkeit, dass der wahre Mittelwert ( $\bar{x} = 52.2$ ) in das Intervall  $[60.3 - 2 \cdot 6.48, 60.3 + 2 \cdot 6.48] = [47.34, 73.26]$  fällt.

Also noch einmal: Der Standardfehler des Mittelwertes beschreibt den Wertebereich, in dem wir mit 65 %-iger Wahrscheinlichkeit den wahren Mittelwert erwarten (und analog für 2 Standardfehler 95 %).<sup>8</sup>

Der **Interquartilsabstand** (*interquartile range*, IQR) ist für den Median, was die Standardabweichung für den Mittelwert ist. Sie gibt den Wertebereich an, in dem 50 % der Datenpunkte liegen (nämlich zwischen 25 % und 75 %).

Die mittlere absolute Abweichung (*median absolute difference*, mad) ist eine robuste Variante des Interquartilsabstands. Ihre Herleitung ist nicht trivial und hier nicht wichtig. Mad ist robuster als IQR, weil es die Werte beiderseits des Median kollabiert und davon den Median berechnet, also die Mitte der kollabierten Werte. Diese werden mit einer Konstanten multipliziert, damit der mad bei normalverteilten Daten mit der Standardabweichung zusammenfällt:<sup>9</sup>

$$\text{mad} = \text{Median}(|x - \text{Median}_x|) \cdot 1.4826 \quad (1.9)$$

Die senkrechten Striche bedeuten „Betrag“, also das Wegnehmen von Minuszeichen.  $x$  repräsentiert hier eine Vektor mit den einzelnen Messwerten:  $x = (x_1, x_2, \dots, x_n)$ .

<sup>8</sup> Vollkommen analog gibt es auch für jede andere Stichprobenstatistik einen Standardfehler. Der Standardfehler der Standardabweichung einer Stichprobe ist beispielsweise  $se_\sigma = \sigma^2 \sqrt{\frac{2}{n-1}}$ , und wir benutzen  $s$  als Schätzer für  $\sigma$ .

<sup>9</sup> Der Faktor ist äquivalent zu Wahrscheinlichkeitsdichte der Standardnormalverteilung bei 0.75 ( $1/\text{qnorm}(3/4)$ ). Hintergründe werden in Tukey (1977) beschrieben.



Schließlich kann man für eine Stichprobe noch schlicht den kleinsten und größten Wert angeben (also das Intervall der Daten) bzw. deren Differenz, die als Spannweite (engl.: *range*) bezeichnet wird.

Wenn eine Stichprobe nicht so hübsch symmetrisch ist wie unsere Eschenumfänge, dann kann man ihre Schiefe berechnen. Wie „spitz“ eine Stichprobe ist beschreibt hingegen die Wölbung. Diese beiden Werte liegen in unserem Fall bei respektive  $-0.073$  und  $-0.054$ . Werte um die Null (wie diese hier), weisen darauf hin, dass die Schiefe und Wölbung nicht von der einer Normalverteilung abweicht.

**Schiefe** und Wölbung können auf unterschiedliche Art berechnet werden. Grundlage für die Schiefe ist  $g$ , wie sie in älteren Lehrbüchern berechnet wird:

$$g = \sqrt{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(\sum_{i=1}^n (x_i - \bar{x})^2)^{3/2}} \quad (1.10)$$

mit  $n$  = Anzahl Datenpunkte und  $\bar{x}$  = Mittelwert von  $\mathbf{x}$ .

Dieser Wert wird heute üblicherweise mit einer von zwei möglichen Korrektur für kleine Stichproben versehen<sup>10</sup> (Joanes und Gill, 1998). Alle drei Formeln sind brauchbar („*unbiased under normality*“).

$$g_2 = g \cdot \frac{\sqrt{n(n-1)}}{(n-2)} \quad (1.11)$$

$$g_3 = g \cdot ((n-1)/n)^{3/2} = g(1 - 1/n)^{3/2} \quad (1.12)$$

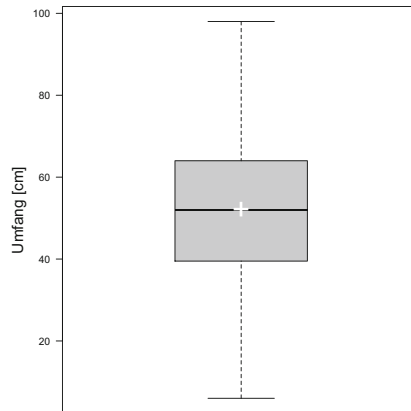
Analog gibt es auch drei Formeln für die **Wölbung** (= Kurtosis). Da nur eine dieser Formeln *unbiased under normality* ist, sei nur diese Formel hier angegeben (siehe Joanes und Gill, 1998):

$$k = \frac{\left( (n+1) \left( \frac{\sum_{i=1}^n x_i^4}{(\sum_{i=1}^n x_i^2)^2} - 3 \right) + 6 \right) (n-1)}{(n-2)(n-3)} \quad (1.13)$$

Stichprobenstatistiken kann man für *jede* Stichprobe von Zahlenwerten berechnen.<sup>11</sup> Ihre Aufgabe ist Zentralität und Streuung zu beschreiben. Für unsere Eschen ist der Mittelwert  $\bar{x} = 52$  cm und die Standardabweichung  $s = 17.9$  cm. Es ist üblich, für die Standardabweichung eine Kommastelle mehr anzugeben als für den Mittelwert. Ebenso ist ein wenig gesunder Menschenverstand gefragt, wenn man entscheidet, wie viele Kommastellen für

<sup>10</sup> Die erste Korrektur ( $g_2$ ) ist etwa bei SAS und SPSS üblich und die Grundeinstellung bei R in **e1071**. Minitab benutzt hingegen Korrektur 2, bzw.  $g_3$ .

<sup>11</sup> Mit „Zahlenwerten“ meinen wir hier *metrische Variablen*, also solche, bei denen der Zahlenwert eine quantitative Aussage macht. Manche Menschen glauben, dass der Berechnung dieser Werte die Annahme der Normalverteilung zugrunde liegt. Das ist nicht so.  $\bar{x}$  und  $s$  sind hier *Stichprobenstatistiken*, keine Verteilungsparameter. Was aber nicht heißt, dass Mittelwert und Standardabweichung immer *sinnvolle* Aussagen für eine Stichprobe machen.



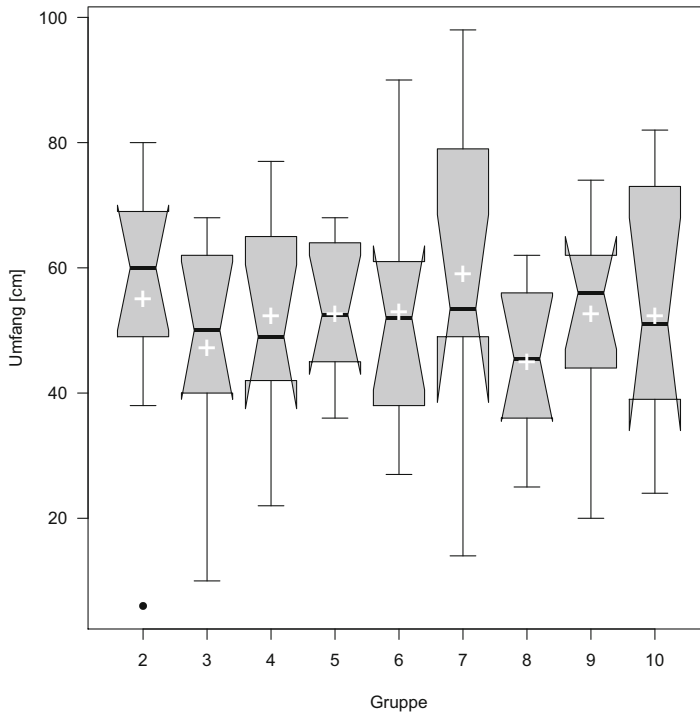
**Abb. 1.4** Boxplot der 100 Eschenumfänge. Während die *horizontale Linie* der Box den Median angibt, zeigt das *Kreuz* den Mittelwert an. Bei symmetrisch verteilten Daten (wie hier) sind diese beiden Zentralitätsmaße sehr ähnlich

den Mittelwert angeben werden sollten. Da in unserem Fall ein Maßband kaum genauer als 0.5 cm messen kann, halte ich eine Kommastelle für den Mittelwert für verzichtbar. Genauigkeit sollte nicht geheuchelt sein.

Eine weitere grafische Darstellungsform für eine Stichprobe ist der Boxplot (genauer der *box-and-whiskers-plot*; Abb. 1.4). Im Gegensatz zum Histogramm, das allein die Rohdaten tabuliert und aufrägt, basiert der Boxplot auf Stichprobenstatistiken, die wir gerade kennengelernt haben. Leider unterscheiden sich Statistikprogramme darin, was in einem Boxplot aufgetragen wird. Grundsätzlich gibt der Boxplot ein Zentralitätsmaß (üblicherweise den Median) als horizontalen, dicken Strich. Da herum wird eine Box gezeichnet, die (üblicherweise) den Bereich von der 1. zur 3. Quartile umfasst, also 50% der Datenpunkte. Schließlich zeigen die Fehlerbalken (Schnurrbarthaare = *whiskers*) die extremen Werte, aber nur, wenn sie innerhalb von 1.5 Mal der Boxlänge (= 1.5 IQR) liegen. Sonst werden die Extremwerte weggelassen (üblich) oder durch Punkte abgebildet (unüblich).

Schließlich gibt es noch eine Variante in der die Box um den Median Kerben (*notches*) erhält. Diese Kerben sind das nicht-parametrische Analogon zum Standardfehler des Mittelwertes und berechnen sich als  $\pm 1.58 \cdot \text{IQR} \cdot \sqrt{n}$ . Nach Chambers et al. (1983, S. 62) zeigen überlappende Kerben an, dass sich zwei Datensätze nicht signifikant unterscheiden.

Der Boxplot fasst die Informationen deutlich stärker zusammen als das Histogramm und ist entsprechend weniger informativ. Andererseits können hiermit mehrere Datensätze unmittelbar verglichen werden. Abb. 1.5 zeigt als Boxplot die gleichen Daten wie Abb. 1.2, nur auf weniger Raum.



**Abb. 1.5** Boxplot der Eschenumfänge von Gruppe 2-9. Der Extremwert der Gruppe 2 liegt außerhalb des 1.5-fachen IQR und ist deshalb zusätzlich abgebildet. Bei den anderen Gruppen liegen die Extremwerte zwar auch außerhalb der *Box*, aber eben nicht außerhalb des 1.5-fachen IQR. Die *Kerben*, als visuelles Maß für Unterschiedlichkeit, ragen hier manchmal über die Boxen hinaus, was dann wie Arme/Beine aussieht. Die *weißen Kreuze* geben den Mittelwert der Gruppe an

Hier sehen wir, dass die Kerben größer sein können als die Box, was zu männchenartigen Abbildungen führt. Hübsch oder nicht, wir sehen, dass alle Stichproben sich ähnlich sind.

Beim Betrachten eines Boxplots achten wir vor allem auf zwei Dinge: (1) Liegt der Median etwa in der Mitte der Box? Wenn das der Fall ist (etwa für Gruppe 5 und 8), dann sind die Stichprobendaten etwa symmetrisch verteilt. Wenn dem so ist, dann ist der Mittelwert ähnlich dem Median und die Chance, dass die Daten normalverteilt sind, ist gegeben. (2) Sind die Fehlerbalken, die *whiskers*, etwa gleich lang? Die Interpretation ist die gleiche wie bei (1).

Im vorliegenden Fall sind also die Daten von Gruppe 8 ansprechend symmetrisch, die von Gruppe 3 deutlich schief. Im direkten Vergleich mit Abb. 1.2 zeigen sich die Boxplots als weniger aufschlussreich und schwieriger zu interpretieren. Andererseits ist die Datenlage mit 10 Punkte je Gruppe auch sehr gering.

Nur der Vollständigkeit halber sei darauf hingewiesen, dass es eine Reihe Mischformen aus Histogramm und Boxplot gibt, die die Verteilung der Daten mit in die Form der Box aufnehmen. Beispielhaft sei der Violinplot genannt, der z. B. von Dormann et al. (2010, dort Abb. 2) benutzt wurde. Einen guten Überblick über die Wahrnehmung solcher Visualisierungen geben Ibrekk und Morgan (1987).

### 1.1.3 Stichprobenstatistiken am Beispiel

Unser Beispieldatensatz sind Beobachtungen von roten Mauerbienen beim Bevorraten ihrer Brutzellen. Dazu fliegen die Bienen weg, sammeln Pollen und lagern diesen dann in eine Brutzelle ein. Wenn diese voll ist, wird ein Ei hineingelegt und die Zelle mit Lehm zugemauert. Unsere Daten geben die Dauer der Pollensammelflüge in einer Obstplantage an (in Minuten). Insgesamt haben wir 101 Werte dazu bestimmt.

```
1.79 2.79 1.65 2.15 2.98 1.88 1.93 1.86 2.00 2.18 2.71 1.80 1.71 2.05 1.71
2.06 1.88 1.37 2.22 2.01 2.53 2.56 2.84 2.44 2.49 2.00 2.81 2.86 1.86 1.79
2.26 2.64 3.30 2.70 2.85 2.56 1.73 1.42 1.49 2.06 2.89 1.80 2.09 3.09 1.93
2.37 1.77 1.93 1.83 2.09 2.84 2.20 2.60 1.88 2.07 1.76 2.46 2.07 2.09 2.22
1.69 2.51 1.89 2.34 1.82 1.98 1.39 1.99 1.64 2.00 2.03 2.02 2.27 2.13 2.30
2.05 2.57 2.17 2.20 1.89 2.34 1.49 2.57 2.11 2.42 1.84 3.41 1.93 2.09 1.91
2.55 1.71 2.37 2.53 2.58 2.29 1.98 1.90 2.04 2.09 1.42
```

Das Wichtigste ist zunächst das Histogramm der Datenpunkte (Abb. 1.6). Wir kombinieren dies hier zum direkten Vergleich mit einem horizontalen Boxplot. Wir sehen, dass die meisten Flüge so um die 2 Minuten dauern, selten weniger als 1.5 oder mehr als 3 Minuten. Der Boxplot zeigt eine klare Abweichung von der Symmetrie, sowohl in der *box*, als auch in den *whiskers*.

Jetzt berechnen wir die Maße aus Tab. 1.2. Die meisten Werte sind redundant, ungebrauchlich oder wenig intuitiv. Wir schauen uns den Zoo der Zentralitäts- und Streuungsmaße hier nur einmal an, um diese Statistiken gesehen zu haben. Sie sind (etwa) nach Wichtigkeit geordnet. Zunächst die Zentralitätsmaße Mittelwert, Median, Modus und Hubers robusten Mittelwert:

$$\text{Mittelwert} = \bar{x} = 2.16$$

$$\text{Median} = 2.07$$

$$\text{Mode} = 2.09$$

$$\text{Hubers Mittelwert} = 2.14$$

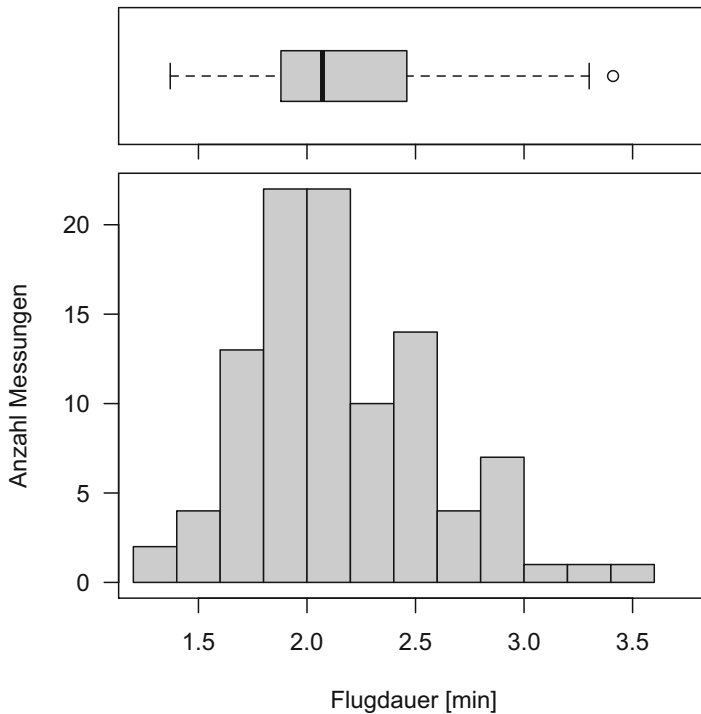
$$\text{Standardfehler des Mittelwerts } se = 0.042$$

$$95\% \text{ Konfidenzintervall} = [2.079, 2.245]$$

Als Nächstes zu den Streuungsmaßen:

$$\text{Standardabweichung } s = 0.422$$

$$\text{Hubers Standardabweichung} = 0.400$$



**Abb. 1.6** Dauer der Provisionierungsflüge der Roten Mauerbiene (*Osmia bicornis*) in einer Apfelplantage. Boxplot horizontal über einem Histogramm

Varianz  $s^2 = 0.177$

Varianzkoeffizient  $CV = 0.20^{12}$

95 % Quantilen = [1.420, 3.035]

Interquartilabstand  $IQR = 0.58$

mittlere absolute Abweichung  $MAD = 0.400$

Intervall = [1.37, 3.41] mit Spannweite 2.04

Schiefte = 0.569

Kurtosis = 0.026.

Viele Zahlen, doch was sagen sie uns?

Zunächst können wir den Unterschied zwischen Mittelwert und Median betrachten (2.16 gegenüber 2.07). Beide liegen relativ eng beieinander, der Mittelwert ist etwas höher, da es mehr längere als kürzere Flugdauern gibt (das Histogramm ist etwas rechtsschief). Der Mode ist ziemlich aussageelos, da selbst der häufigste Wert laut Histogramm nur 5 Mal vorkam. Hubers Mittelwert liegt nahe dem arithmetischen Mittelwert, also ist letz-

<sup>12</sup> Häufig wird dieser Wert mit 100 multipliziert und in % angegeben, hier also 20 %.

terer nicht durch Ausreißer verzerrt. Schließlich gibt uns der Standardfehler an, dass bei erneutem Messen der neue Mittelwert nicht weit von  $\bar{x}$  entfernt ist (da  $se$  einen kleinen Wert hat). Genauer gesagt liegen 95 % aller Mittelwerte solcher Wiederholungen im Intervall  $[\bar{x} - 2se, \bar{x} + 2se] = [2.078, 2.246]$ . Diese Werte sind nahezu identisch zum 95 % Konfidenzintervall. Letzteres unterstellt für seine Berechnung, dass die Daten normalverteilt sind. Dieser Bereich umfasst den Median gerade nicht mehr.

Die Streuungsmaße quantifizieren wie stark die Werte sich unterscheiden. Nur wenige lassen sich ohne weitere Annahmen direkt interpretieren. Ein Varianzkoeffizient von 0.2 (oder 20 %) deutet an, dass die Werte nur moderat relativ zum Mittelwert variieren. In anderen Worten, die Standardabweichung von  $s = 0.422$  ist nicht besonders groß (oder besonders klein) für einen Mittelwert von  $\bar{x} = 2.16$ . CV-Werte unter 0.05 zeichnen sehr hohe Präzision aus, solche über 0.2 geringe. In der Ökologie liegen sie gelegentlich sogar über 1, da wir meistens nur wenige Datenpunkte haben und ein hochvariables System beproben.

Der Unterschied zwischen 95 % Konfidenzintervall und den 95 % Quantilen kann man an diesen Statistiken gut sehen. Das Konfidenzintervall beschreibt, ähnlich wie der Standardfehler, die Genauigkeit des Mittelwertsschätzers  $\bar{x}$ . Die Quantilen hingegen beschreiben den Wertebereich der Daten: 95 % liegen zwischen 1.42 und 3.04.

Der Interquartilabstand ist mit 0.58 größer als die Standardabweichung (0.4). Das deutet darauf hin, dass die Daten nicht symmetrisch sind, weil dadurch der IQR in einer Richtung größer wird, während die Standardabweichung nur wenig zunimmt. Und, in der Tat, sehen wir dies nicht nur im Histogramm, sondern auch in der Schiefe, die mit 0.57 deutlich von 0 (symmetrisch) abweicht. Positive Werte indizieren rechtsschiefe (= linkssteile) Daten, während negative linksschiefe (= rechtssteile) anzeigen.

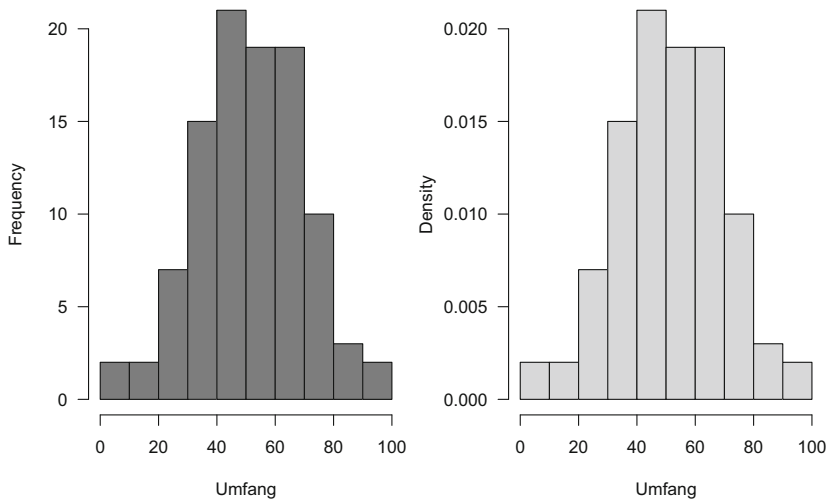
Zusammenfassend können wir also sagen, dass die Mauerbienen im Mittel  $2.16 \pm 0.042$  (Mittelwert  $\pm$  Standardfehler) Minuten für einen Pollensammelflug brauchen. Wenn wir jetzt einen erneuten Flug messen würden, dann wäre anzunehmen, dass er eben jene 2.16 Minuten dauern würde, aber sehr wahrscheinlich im Intervall  $[1.42, 3.04]$  liegt.

---

## 1.2 Häufigkeit, Dichte und Verteilung

Im übernächsten Kap. 3 werden wir uns intensiver mit Verteilungen auseinandersetzen. Hier wollen wir einen Übergang versuchen, von unseren Stichproben (z. B. als Histogramm), zu eben jenen Verteilungen.

Bislang hatte unser Histogramm auf der y-Achse die Anzahl der Beobachtungen in der jeweiligen Größenklasse (siehe etwa Abb. 1.3). In anderen Worten, das Histogramm beruht auf der *Häufigkeit* der Daten. Mit Häufigkeit meinen wir hier schlicht die Anzahl pro Klasse (in Anlehnung ans englische Wort *frequency* gelegentlich auch als Frequenz bezeichnet). Alternativ können wir auch die *Dichte* auftragen. Das ist die Häufigkeit geteilt durch die Anzahl der Messungen und die Klassenbreite. Die Dichte über alle Klassen



**Abb. 1.7** Messungen aller 100 Eschenumfänge, als Häufigkeit- und als Dichtehistogramm

summiert sich zu 1. Das Einzige, was sich in dieser Abbildung ändert, ist die Skalierung der y-Achse (siehe Abb. 1.7).

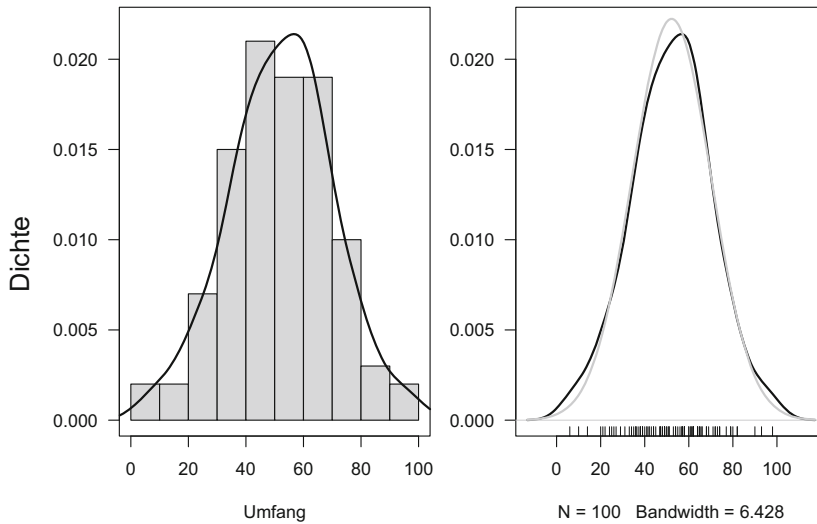
Nun ist ja die Einteilung in Klassen willkürlich. Es gibt zwar Algorithmen, die besonders schöne Einteilungen vornehmen, aber es findet immer eine Kategorisierung in Klassen statt. Aber die zugrundeliegenden Daten sind ja kontinuierlich; eine Esche kann jeden Umfang innerhalb des gemessenen Intervalls annehmen. Deshalb wäre es doch logisch, wenn man auch die Stichprobe kontinuierlich darstellen würde.

Nein, leider ist das nicht ganz so logisch! Die Stichprobe ist ja nur eine Realisierung der zugrundeliegenden Verteilung. Entsprechend kann man nicht einfach die Stichprobe zur Verteilung machen. Erst wenn wir wirklich *viele* Datenpunkte haben sollte diese Stichprobe der wahren Verteilung ziemlich ähnlich sehen.

Unter diesem Vorbehalt sehen wir uns mal an, wie das Dichtehistogramm als Dichteverteilung aussehen würde (Abb. 1.8 links). Diese Kurve zeigt, wie eine kontinuierliche Dichte der Daten aussehen würde. Diese Dichtekurve wird durch einen gleitenden Kernel berechnet.<sup>13</sup> Die Informationen unter der x-Achse geben die Anzahl der Datenpunkte an ( $N = 100$ ), sowie die Breite des Kernels (hier 6.4 cm). Wir nehmen hier einfach einmal an, dass diese Methode schon weiß, was sie tut.

Als Ergebnis haben wir jetzt nicht mehr eine Darstellung unserer Stichprobendaten, sondern eine kontinuierliche Verteilung auf Basis unserer Stichprobendaten. Wir hoffen, dass diese Verteilung der wahren Verteilung der Grundgesamtheit aller Eschenumfänge sehr ähnlich ist.

<sup>13</sup> Die Mathematik dahinter ist recht kompliziert. Die Idee ist eine gleitende Berechnung der Dichte, wobei nur die Punkte im Einzugsbereich zur Berechnung beitragen.



**Abb. 1.8** Messungen aller 100 Eschenumfänge als empirische Dichtekurve über das Dichtehistogramm gelegt (*links*) und mit unterlegter Normalverteilung in grau (*rechts*). Die *Strichlein* (*rechts*, im Englischen als *tick* und in ihrer Gesamtheit als *rug* bezeichnet) geben die Lage der Datenpunkte an

Zur Illustrierung können wir noch eine Normalverteilung dazulegen (Abb. 1.8, rechts). Diese hat zwei Parameter (siehe Abschn. 3.4.1), nämlich den Mittelwert  $\mu$  und die Standardabweichung  $\sigma$ . Wir berechnen also den Mittelwert unserer Stichprobe  $\bar{x}$  ( $= 52.2$ ) und ihre Standardabweichung  $s$  ( $= 17.94$ ) und nehmen an, dass sie gute Schätzer für  $\mu$  und  $\sigma$  sind. Wie wir sehen liegen die zwei Kurven sehr eng beieinander. Wir könnten jetzt mit dem Brustton der Überzeugung sagen, dass die Eschenumfänge normalverteilt sind.

Wieso aber gerade die Normalverteilung? Gibt es noch andere Verteilungen? Können wir irgendwie statistisch prüfen, ob eine Verteilung passt? Diese Fragen werden uns im Kap. 3 beschäftigen.



Parametrische Statistik

Verteilungen, maximum likelihood und GLM in R

Dormann, C.F.

2017, XXIII, 363 S. 132 Abb., 31 Abb. in Farbe.,

Softcover

ISBN: 978-3-662-54683-3