Kapitel

8 Eine metrische Variable

Wir sprechen von METRISCHEN Variablen, wenn Beobachtungen nach Festlegen der Maßeinheit sinnvoll durch Zahlen repräsentiert und umgekehrt diese Zahlen klar interpretiert werden können. Differenzen von Variablenwerten haben eine Bedeutung (mindestens Intervallskala).

DISKRETE metrische Variablen liegen vor, wenn die Messung in nicht mehr weiter unterteilbaren Einheiten erfolgt. Meist sind es Zählvariablen mit ganzen Zahlen als Werten. Typische Beipiele sind etwa Anzahl Geschwister oder Dauer eines Krankenstands (in Tagen gemessen).

STETIGE metrische Variablen werden in beliebig teilbaren Einheiten gemessen und können zumindest in bestimmten Bereichen der reellen Zahlenachse im Prinzip jeden Wert annehmen. Viele Beispiele, bei denen die Messung mit physikalischen Messgeräten erfolgt, fallen darunter, etwa Körpergewicht und -größe, Wartezeit vor einem Bankschalter.

Fast alle der hier vorgestellten Methoden und Verfahren benötigen keine genaue Unterscheidung zwischen diskret und stetig. Einzige Ausnahme sind diskrete Variablen, die nur wenig Werte annehmen können (etwa Geschwisterzahl, Alter In Jahren bei Volksschulkindern etc.).

Lernziele:

Nach Durcharbeiten dieses Kapitels haben Sie Folgendes erreicht:

- Sie verstehen, was mit unterschiedlichen Maßzahlen beschrieben wird, und können diese Maßzahlen mit R berechnen. Sie kennen die wichtigsten Verteilungsformen.
- Sie können die Verteilung einer metrischen Variablen mit Tabellen, Histogrammen und Boxplots beschreiben.
- Sie können überprüfen, ob eine bestimmte Verteilung – speziell eine Normalverteilung – vorliegt.
- Sie sind in der Lage, den Mittelwert über ein Konfidenzintervall zu schätzen und gegen einen vorgegebenen Wert zu testen.

8.1 Wie kann man die Verteilung einer metrischen Variablen beschreiben?

Unter dem Begriff VERTEILUNG fassen wir mehrere Aspekte zusammen:

- In welchem Bereich liegen die Daten?
- Wo innerhalb dieses Bereiches sind die Daten stärker, wo schwächer vertreten?
- Gibt es ein Zentrum der Daten oder mehrere Zentren oder gar keines?
- Variieren die Daten stark oder nur wenig?
- Liegen die Daten symmetrisch um einen Wert?

Fallbeispiel 12: Verfahrensdauer am Verwaltungsgerichtshof

Datenfile: `vwgh.csv`

Gegen Abgabenbescheide von Behörden kann Berufung eingelegt werden. In Österreich ist die Berufungsbehörde 2. Instanz der Verwaltungsgerichtshof (VwGH). In einer Studie (Hornik et al, 2008) wurden alle Entscheidungen des VwGH zwischen 2000 und 2004 in Abgabensachen untersucht.
Ein Gegenstand der Untersuchung war die Zeit, die zwischen Einbringung der Beschwerde bis zur Entscheidung im VwHG vergeht. Insgesamt wurden 3827 Entscheidungen untersucht.

Fragestellung:
Wie kann die Verteilung der Verfahrensdauern in der Stichprobe beschrieben werden?

8.1.1 Klassifizieren, Tabellen und Histogramme

Histogramme und Klassifizieren

Im Datenfile sind die Verfahrensdauern vor dem VwGH in der Variablen `dauer3` enthalten. In einigen Fällen konnte das Datum der Einbringung der Beschwerde beim VwGH nicht erhoben werden. In diesen Fällen hat die Variable den Wert -9999.

Nach dem Einlesen der Daten schließen wir die Fälle aus, zu denen keine Dauer vor dem VwGH bekannt ist, und benennen die Variable neu mit `vwghdauer`. Anschließend lassen wir für diese Variable ein Histogramm erstellen.

R

```
> vwgh <- read.csv2("vwgh.csv", header = TRUE)
> attach(vwgh)
> vwghdauer <- dauer3[dauer3 != -9999]
> detach(vwgh)
> hist(vwghdauer)
```

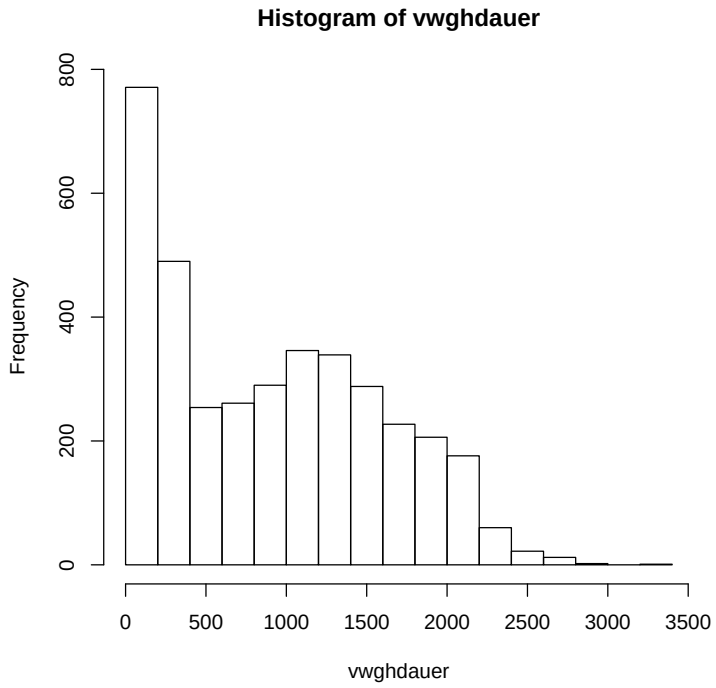


Abbildung 8.1: Histogramm mit konstanten Klassenbreiten

Das Erscheinungsbild ist auf den ersten Blick ähnlich dem von Balkendiagrammen (▶ Abschnitt 6.2.2), die Höhe der Balken korrespondiert mit den Häufigkeiten für die jeweiligen Klassen.

Ein Unterschied ist, dass die Balken ohne Zwischenraum nebeneinander stehen. Grund dafür ist, dass eine metrische Variable zur Definition der x-Achse dient und nicht wie beim Balkendiagramm eine kategoriale. Die dem Histogramm zugrunde

liegende Klassifizierung der Variablenwerte hat zu direkt benachbarten Klassen geführt.

Hier kommen auf 1000 Einheiten auf der x-Achse jeweils fünf Klassen, also umfasst jede 200 Tage. Somit steht die erste Klasse für Verfahrensdauern bis zu 200 Tagen, die zweite Klasse für Verfahrensdauern von mehr als 200 aber höchstens 400 Tagen, usw.

Die Beschriftung der y-Achse bedeutet, dass die Höhe der Balken absolute Häufigkeiten der einzelnen Klassen anzeigen. In die Klasse mit höchstens 200 Tagen Verfahrensdauer fallen also etwas weniger als 800 Beobachtungen.

Die Auswahl der Klassenanzahl sowie der Klassengrenzen kann man entweder R überlassen oder selbst treffen. So sind in diesem Beispiel in den ersten zwei Klassen sehr viele Beobachtungen, in den letzten nur sehr wenige. Dort wo viele Beobachtungen liegen, wäre eine feinere Klasseneinteilung wünschenswert. Für die langen Verfahrensdauern wäre eine gröbere Einteilung ausreichend.

In R kann dies durch einen Vektor, in dem die Klassengrenzen enthalten sind, erfolgen. Die Klassengrenzen werden hier im Vektor `grenzen` gespeichert und im Histogrammaufruf mit `breaks=grenzen` als Klassengrenzen festgelegt. Zusätzlich wird ein eigener Titel für das Histogramm angegeben und die x-Achse nicht mit dem Variablennamen beschriftet.

R

```
> grenzen <- c(0, 100, 200, 300, 400, 600, 800, 1000, 1200,
+     1400, 1600, 1800, 2000, 2200, 2500, 2800, 3300)
> hist(vwghdauer, breaks = grenzen, main = "VwGH-Verfahrensdauer",
+     xlab = "Dauer in Tagen")
```

Der Effekt der geänderten Klassengrenzen ist einerseits zu Beginn (bis 400 Tage) eine feinere und nach 2200 Tagen eine gröbere Einteilung der Klassen; andererseits ist die y-Achse jetzt anders beschriftet und die Skaleneinteilungen zeigen sehr kleine Werte an. Dieser Effekt tritt auf, sobald die Klassenbreiten nicht konstant. Anstatt absoluter Klassenhäufigkeiten anzuzeigen, sind die Balkenhöhen jetzt so ausgelegt, dass die Fläche (und nicht die Höhe) des Balkens der relativen Häufigkeit einer Klasse entspricht. Die Gesamtfläche aller Klassen ergibt 1.

Bei der Wahl der Klassenanzahl ist mit Vorsicht vorzugehen. Faustregeln dafür, in wie viele Klassen die Einteilung erfolgen soll, gibt es zuhauf. Allen liegt die Idee zugrunde, einerseits wenig Klassen zu bilden, um eine kompakte Darstellung der Daten zu erhalten, und andererseits doch so viele Klassen, um möglichst wenig Informationsverlust zu erleiden. Einige Vorschläge für die Klassenanzahl k bei n Beobachtungen sind:

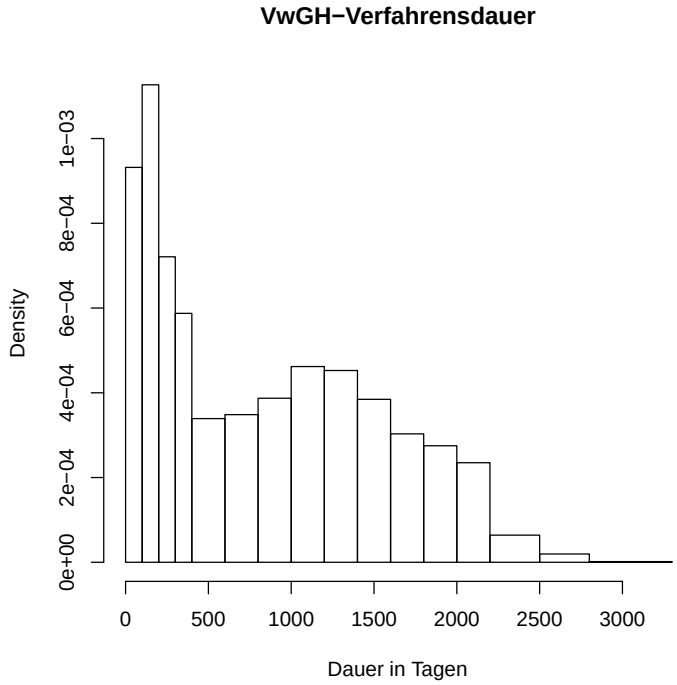


Abbildung 8.2: Histogramm mit variablen Klassenbreiten

- $5 \leq k \leq 20$:
  Im VwGH-Beispiel ($n = 3745$ relativ groß) würde man eher in Richtung Obergrenze gehen.
- $k \approx \sqrt{n}$:
  Dieser Vorschlag würde zu mehr als 60 ($\sqrt{3745} = 61.2$) Klassen führen, das ist eindeutig zu viel. Diese Faustregel ist nur für einen moderaten Stichprobenumfang ($n \leq 200$) brauchbar.
- k so, dass $2^k \approx n$
  Also etwa 12 ($2^{12} = 4096$) Klassen.

Tabellen

Liegen nur wenig Beobachtungen vor oder kann die Variable, die beschrieben werden soll, nur wenige Werte annehmen, ist es denkbar, wie bei einer kategorialen Variablen eine einfache Auszählung (▶ Abschnitt 6.2.1) durchzuführen.

Für dieses Beispiel ist es sinnlos, da mehrere Hundert unterschiedliche Werte vorliegen und uns die entstehende Auflistung kaum einen Überblick über die

Verteilung bietet. Wir können aber die Werte in KLASSEN (Bereiche, Intervalle) zusammenfassen und die Auszählung für die Klassen erstellen lassen.

Dies kann mit dem Befehl `cut()` leicht durchgeführt werden. Wenn dieselben Klassengrenzen wie für das Histogramm gelten sollen, können wir den Vektor `grenzen` verwenden. Die neu gebildete Variable `vwghdauerkat` anstatt der eigentlichen Verfahrensdauer die Kategorie, in die die jeweilige Verfahrensaduer fällt. Da automatisch auch Labels für Klasseneinteilung produziert werden und diese hier recht lang werden, weichen wir mit `dig.lab=4` von der Voreinstellung ab. Die Häufigkeitstabelle für diese neue Variable hat wegen der vielen Kategorien keine sehr schöne Form.

R

```
> vwghdauerkat <- cut(vwghdauer, breaks = grenzen, dig.lab = 4)
> tdauer <- table(vwghdauerkat)
> tdauer
```

```
vwghdauerkat
    (0,100]   (100,200]   (200,300]   (300,400]   (400,600]
        349         422         270         220         254
  (600,800] (800,1000] (1000,1200] (1200,1400] (1400,1600]
        261         290         346         339         288
(1600,1800] (1800,2000] (2000,2200] (2200,2500] (2500,2800]
        227         206         176          72          22
(2800,3300]
          3
```

Eine umfagreichere, aber vor allem übersichtlichere Tabelle erhalten wir, indem die Häufigkeitstabelle als Spaltenvektor ausgegeben wird. Zusätzlich werden relative (Prozente) und kumulierte relative Häufigkeiten bestimmt. Der sinnlosen Ausgabe vieler Nachkommastellen wird mit `options(digits=2)` ein Riegel vorgeschoben.

R

```
> options(digits = 2)
> n <- sum(tdauer)
> prozent <- tdauer * 100/n
> kumproz <- cumsum(prozent)
> cbind(absolut = tdauer, Prozent = prozent, kumuliert = kumproz)
```

```
            absolut Prozent kumuliert
(0,100]         349    9.32       9.3
(100,200]       422   11.27      20.6
(200,300]       270    7.21      27.8
(300,400]       220    5.87      33.7
```

(400,600]	254	6.78	40.5
(600,800]	261	6.97	47.4
(800,1000]	290	7.74	55.2
(1000,1200]	346	9.24	64.4
(1200,1400]	339	9.05	73.5
(1400,1600]	288	7.69	81.1
(1600,1800]	227	6.06	87.2
(1800,2000]	206	5.50	92.7
(2000,2200]	176	4.70	97.4
(2200,2500]	72	1.92	99.3
(2500,2800]	22	0.59	99.9
(2800,3300]	3	0.08	100.0

Fallbeispiel 12: Interpretation des Histogramms und der Tabelle

Wir haben den Bereich den Verfahrensdauern (▶ Abbildung 8.2) in 16 Klassen eingeteilt, zu Beginn vier Klassen mit einer Berite von 100 Tagen, dann relativ viele Klassen mit einer Breite von 200 Tagen, zum Schluss drei breitere Klassen.
Gleich zu Beginn stehen vier hohe Balken, die für Entscheidungen stehen, die innerhalb der ersten 400 Tage (also ca. 13 Monate) erfolgt sind. Aus der Tabelle kann in der Spalte mit kumulierten Häufigkeiten abgelesen werden, dass ziemlich genau ein Drittel der Beschwerden innerhalb dieses Zeitraumes bearbeitet wurde.
Danach gibt es noch einmal eine Häufung zwischen 1000 und 1400 Tagen (also ca. 3 und 4 Jahren). Anschließend nehmen die beobachteten Verfahrensdauern langsam ab. Über 2800 Tagen ist kein Balken erkennbar. Das bedeutet nicht, dass in diesen Bereich keine Beobachtungen fallen.
Aus der Häufigkeitstabelle ist ersichtlich, dass es tatsächlich genau drei Beschwerden sind, bei denen es im VwGH mehr als 2800 Tage bis zu einer Entscheidung brauchte.

Von juristischer Seite werden die kurzen Verfahren hauptsächlich auf Formalerledigungen zurückgeführt (etwa Zurückweisung wegen Formalfehlern). Für die sehr langen Verfahrensdauern gibt es keine inhaltliche Erklärung.

8.1.2 Maßzahlen zur Beschreibung der Verteilung

Mit Histogrammen kann die Verteilung einer metrischen Variablen grafisch gut beschrieben werden. Oft besteht aber auch der Wunsch, mit wenigen Zahlen wesentliche Angaben über die Verteilung zu treffen.

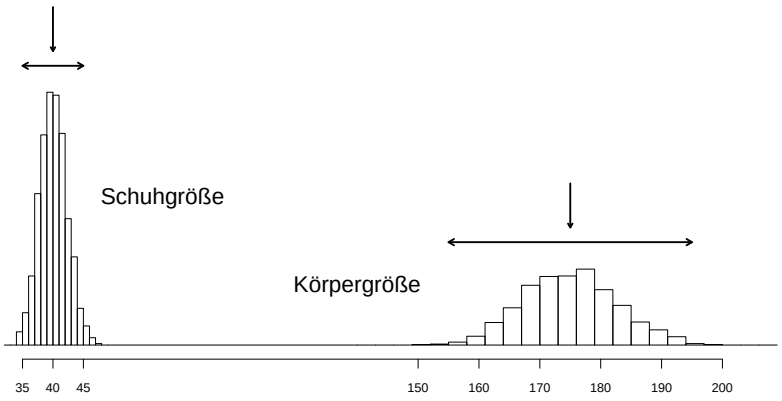


Abbildung 8.3: Verteilung von Schuhgröße und Körpergröße

In der Abbildung 8.3 sind zwei Histogramme, je eines für die Schuhgröße und die Körpergröße, auf einer gemeinsamen x-Achse aufgebaut. Maßzahlen, die beschreiben, wo das Zentrum der Daten ist, nennt man LAGEMASSE. Für die Schuhgröße sollen das Werte um 40, für die Körpergröße um 175 sein. Maßzahlen, die angeben, wie stark die Daten variieren, nennt man STREUUNGSMASSE. Die Schuhgrößen variieren weit weniger stark als die Körpergrößen, die Streuungsmaße sollen daher für die Schuhgröße kleinere Werte als für die Körpergröße ergeben.

Für die Erläuterung und die Berechnung der verschiedenen Lage- und Streuungsmaße soll der folgende kleine hypothetische Datensatz mit nur 10 Werten dienen:

7 10 16 9 12 13 9 8 10 9

Lagemaße

Die drei wichtigste Lagemaße sind:

- MITTELWERT $\bar{x}$
  Das wohl bekannteste Lagemaß, das durch Aufsummieren der Werte x_i und anschließendes Dividieren durch die Anzahl n der Werte gewonnen wird

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

Für den kleinen Beispielsdatensatz bedeutet es also:

$$(7 + 10 + 16 + 9 + 12 + 13 + 9 + 8 + 10 + 9)/10 = 103/10 = 10.3$$

- **Median $\tilde{x}$**

  Nach dem Sortieren der Daten wird der Wert in der Mitte bestimmt. Bei einer ungeraden Anzahl von Werten ist der Wert eindeutig, bei einer geraden Anzahl von Werten mittelt man die beiden Werte, die der Mitte am nächsten sind.

  Sortieren führt zu:

  7 8 9 9 9 10 10 12 13 16

  Die Werte 9 und 10 sind der Mitte am nächsten, also ist:

  $$\tilde{x} = (9 + 10)/2 = 9.5$$

- **Modus (Modalwert)**

  Der am häufigsten auftretende Wert im Datensatz ist der Modus. Als einziges der vorgestellten Lagemaße ist er auch für kategoriale Variablen einsetzbar (und wird auch dort fast ausschließlich für diese eingesetzt).

  Im Beispieldatensatz ist der Modus 9 (kommt dreimal vor).

Eigenschaften der vorgestellten Lagemaße:

- Für die sinnvolle Anwendung des Medians genügen ordinal skalierte Variablen.
- Der Mittelwert erfordert metrische Daten.
- Der Mittelwert kann weit stärker als der Median durch einzelne extreme Werte (Ausreißer) beeinflusst werden als der Median (▶ Abschnitt 8.1.4).
- Bei einer Version des Mittelwerts wird ein gewisser Prozentsatz (etwa 5%) der kleinsten und größten Werte weggelassen und aus den Restdaten der Mittelwert berechnet. Man spricht vom getrimmten Mittel. Damit reduziert man den Einfluss von Ausreißern wesentlich.

Exkurs 8.1: Quartile, Quantile und Perzentile

Zur Beschreibung von Verteilungen dient auch die Angabe von bestimmten Positionen in der Verteilung. So gibt etwa das Medianeinkommen jenes Einkommen an, das die Hälfte der Bevölkerung höchstens und die andere Hälfte der Bevölkerung mindestens erreicht.

Eine erste Verallgemeinerung führt zu Quartilen, wenn man von den zwei Hälften des Datensatzes zu den vier Vierteln übergeht. Das 1. Quartil (auch unteres Quartil genannt und mit Q_1 bezeichnet) ist jener Wert, der das Viertel der kleinen Werte von den oberen drei Vierteln trennt. Analog ist das 3. Quartil (obere Quartil, Q_3) jener Wert, der das Viertel der großen Werte von den unteren drei Vierteln trennt. In diesem Sinn kann der Median auch als 2. Quartil aufgefasst werden.

Die Bestimmung der Quartile ist im Prinzip einfach: Q_1 (bzw. Q_3) ist der Median der unteren (bzw. oberen) Hälfte. So einfach die Idee, so uneinheitlich die Ausführungen (etwa bei ungeradem Stichprobenumfang, wenn nicht eindeutig klar ist, was untere bzw. obere Hälfte des Datensatzes ist).

Geht man von Vierteln zu beliebigen Aufteilungen über, spricht man von Quantilen. Das α-Quantil (α ist ein Wert zwischen 0 und 1) ist jener Wert so, dass der Anteil der Beobachtungen, die **höchstens** so groß sind, gleich α ist. Q_1 (bzw. Q_3) ist also das 0.25-Quantil (bzw. 0.75-Quantil) und der Median das 0.5-Quantil.

Verwendet man statt Anteilen zwischen 0 und 1 für α Prozentangaben zwischen 0 und 100, spricht man auch von Perzentilen. Q_1 ist also das 25-Perzentil.

Ist man also an der Grenze (nach unten) interessiert, ab der die 10% der am schlechtesten Verdienenden beginnen, geht es um das 0.1-Quantil (10-Perzentil) des Einkommens. Geht es auf der anderen Seite um das 1% der Topverdiener, kommt das 0.99-Quantil (99-Perzentil) ins Spiel, weil 99% höchstens bis zu dieser Grenze kommen.

Die bis jetzt besprochenen Quartile, Quantile und Perzentile beziehen sich auf eine gegebene Stichprobe. Analoge Fragestellungen für theoretische Verteilungen treten oft in Zusammenhang mit statistischen Tests auf. Implizit sind sie uns schon in den vorigen zwei Kapiteln mit der χ^2-Verteilung begegnet. Es ging um die Bewertung, ob ein aus der Stichprobe berechneter X^2-Wert so groß ist, dass eine Nullhypothese verworfen werden muss. Anstelle eines Vergleichs des p-Werts mit dem Signifikanzniveau α, etwa $\alpha = 0.05$, wäre es auch möglich, den errechneten X^2-Wert mit einem passenden Quantil $1 - \alpha$, also meist 0.95, einer χ^2-Verteilung zu vergleichen (▶ Abschnitt 6.9).

Streuungsmaße

An Streuungsmaßen besprechen wir:

- Varianz s^2
  Man berechnet die Abweichungen der Beobachtungen x_i vom Mittelwert $\bar{x}$, quadriert diese und berechnet davon den Mittelwert. Aus technischen Gründen ist es günstiger, den Mittelwert nicht durch Division durch n, sondern durch $n-1$ zu bilden:

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

  $$(7-10.3)^2 + (10-10.3)^2 + \cdots + (9-10.3)^2)/(10-1) = 64.1/9 = 7.1222$$

- Standardabweichung s
  Dies ist die Wurzel der Varianz

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

  $$s = \sqrt{7.1222} = 2.6687$$

- Spannweite
  Differenz zwischen größtem und kleinstem Wert

  $$16 - 7 = 9$$

- Quartilsabstand (Interquartilbereich, QD)
  Differenz zwischen drittem und erstem Quartil

  $$QD = Q_3 - Q_1$$

  Berechnet man Q_1 (bzw. Q_3) als Median der unteren (bzw. oberen) Datenhälfte, erhält man:

  $$QD = 12 - 9 = 3$$

  R würde in der Standardeinstellung einen etwas anderen Wert für das 3.Quartil (nämlich 11.5) und in der Folge für den Quartilsabstand 2.5 ausgeben. Insgesamt stehen neun Methoden der Quantilsberechnung zur Auswahl, die über einen passendes Argument im Aufruf der `quantile()`-Funktion ausgewählt werden können. Genaueres kann über die Hilfefunktion `?quantile` in Erfahrung gebracht werden.

Hinweis

Eigenschaften der vorgestellten Streuungsmaße:

- Die Berechnung von Streuungsmaßen ist nur bei metrischen Daten sinnvoll.
- Streuungsmaße können nicht negativ werden.
- Varianz, Standardabweichung und Spannweite sind nur dann 0, wenn alle Werte ident sind.
- Varianz, Standardabweichung und Spannweite können weit stärker von einzelnen Werten (Ausreißern) beeinflusst werden als der Quartilsabstand (▶ Abschnitt 8.1.4).
- Im Allgemeinen werden die Werte nicht direkt interpretiert, sondern nur die entsprechenden Werte zwischen Gruppen verglichen (etwa: die Streuung in zwei Gruppen unterscheidet sich, weil ein bestimmtes Streuungsmaß deutlich unterschiedliche Werte in den Gruppen annimmt).

Natürlich sind die wichtigsten Maßzahlen in R leicht zu berechnen. Zunächst bestimmen wir Lagemaße und geben sie in einem Block aus:

```
> mittelwert <- mean(vwghdauer)
> median <- median(vwghdauer)
> getrimmter_mw <- mean(vwghdauer, trim = 0.05)
> rbind(mittelwert, median, getrimmter_mw)
```

```
                  [,1]
mittelwert    914.8529
median        868.0000
getrimmter_mw 887.6808
```

Analog gehen wir mit Quantilen vor:

```
> minimum <- min(vwghdauer)
> quartil_1 <- quantile(vwghdauer, 0.25)
> quartil_3 <- quantile(vwghdauer, 0.75)
> maximum <- max(vwghdauer)
> rbind(minimum, quartil_1, quartil_3, maximum)
```

```
           25%
minimum      2
quartil_1  258
quartil_3 1443
maximum   3262
```

Die seltsame Beschriftung dieser Spalte (25%) rührt daher, dass die `quantile`-Funktion neben dem Wert auch eine Beschriftung mitgibt. Von den in dieser Spalte enthaltenen Werten haben die Quartile eine solche, die erste dieser Beschriftungen wird angezeigt.

Die Streuungsmaße sind ebenfalls leicht abrufbar:

```
> varianz <- var(vwghdauer)
> standardabweichung <- sd(vwghdauer)
> quartilsabstand <- IQR(vwghdauer)
> rbind(varianz, standardabweichung, quartilsabstand)
```

```
                          [,1]
varianz           462931.2030
standardabweichung    680.3905
quartilsabstand      1185.0000
```

Einzig der Modus ist nicht direkt implementiert und muss als jene Stelle oder – wenn nicht eindeutig – als jene Stellen berechnet werden, an der oder denen die Häufigkeitstabelle ihr Maximum annimmt.

```
> tabdauer <- table(vwghdauer)
> modus <- which(tabdauer == max(tabdauer))
> modus
```

```
119
107
```

Die zwei Angaben im Output enthalten zuerst den Modus (119) und zusätzlich die Angabe, wo dieser Wert in einer Häufigkeitstabelle der eigentlichen Werte zu finden wäre. Von allen unterschiedlichen Verfahrensdauern ist 119 also der 107-kleinste Wert.

Die Werte für das Minimum, erstes Quartil, Median, drittes Quartil und Maximum werden auch als Fünf-Punkt-Zusammenfassung bezeichnet. In R stehen dafür im Prinzip zwei Funktionen bereit. Mit `fivenum()` werden die fünf Werte ermittelt:

```
> fivenum(vwghdauer)
```

```
[1]    2  258  868 1443 3262
```

Mit `summary()` wird zusätzlich der Mittelwert bestimmt:

```
> summary(vwghdauer)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0   258.0   868.0   914.9  1443.0  3262.0
```

Überdies kann es zu kleinen Unterschieden in den ausgegebenen Quartlien kommen, wenn durch die Voreinstellungen unterschiedliche Methoden für die Berechnung der Quartile festgelegt sind.

Fallbeispiel 12: Interpretation der Maßzahlen

Die berechneten Maßzahlen sind eine geballte Ladung an Information.
Der Mittelwert von ca. 915 Tagen bedeutet, dass es im Durchschnitt ziemlich genau 2.5 (=915/365) Jahre gedauert hat, bis am VwGH über eine Beschwerde entschieden wurde. Dass das getrimmte Mittel und der Median kleiner sind, liegt daran, dass einige Verfahren sehr lange gedauert haben und diese den Mittelwert im Vergleich dazu nach oben ziehen.
Die kürzeste Verfahrensdauer beträgt zwei Tage, die längste 3263 Tage (also fast neun Jahre). Ein Viertel der Verfahren war nach 258 Tagen abgeschlossen, ein Viertel dauerte länger als 1443 Tage (fast vier Jahre).

Der Wert für die Varianz ist deshalb sehr hoch, weil die Verfahrensdauern nicht nur stark variieren, sondern auch einen großen Wertebereich abdecken (von 2 bis 3262). Hätte man die Verfahrensdauern nicht in Tagen, sondern in Wochen erhoben, wäre der Wert für die Varianz nur $462931.2/(7^2) = 9447.58$.

Weitere Kennzeichen einer Verteilung

Lage- und Streuungsmaße sind die wichtigsten Maßzahlen für die Verteilung einer Stichprobe. Es gibt noch weitere Aspekte bei der Beschreibung der Form einer Verteilung beeinflussen.

- Schiefe
  Mit dem Schiefekoeffizienten wird versucht, die Abweichung der Häufigkeitsverteilung von einer symmetrischen Verteilung (im Idealfall auch in einem symmetrischen Histogramm ersichtlich) zu messen.
  Man unterscheidet rechtsschiefe und linksschiefe Verteilungen. Bei rechtsschiefen Verteilungen ist der Median (deutlich) kleiner als der Mittelwert, bei linksschiefen Verteilungen sind die Verhältnisse gerade umgekehrt. In so gut wie jedem Land ist die Einkommensverteilung rechtsschief. Ebenso sind dies auch die Verfahrensdauern im VwGH-Beispiel (▶ Abbildung 8.6).
  Eine Verteilung, die weder links- noch rechtsschief ist, nennt man symmetrisch.
- Wölbung (Kurtosis, Exzess)
  Im Vergleich zur Normalverteilung (▶ Exkurs 8.2) wird untersucht, ob mehr oder weniger Gewicht auf den Enden der Verteilung liegt. Sinnvoll sind sog.

Wölbungskoeffizienten nur bei (in etwa) symmetrischen Verteilungen interpretierbar.

- Unimodalität und Multimodalität

  Haben die Daten ein Zentrum, um das herum die Daten verteilt liegen, spricht man von einer eingipfeligen (unimodalen) Verteilung. Gibt es mehrere Zentren, so ist die Verteilung mehrgipfelig (multimodal). Für Ein- bzw. Mehrgipfeligkeit gibt es keine Maßzahlen, sie wird am besten an einem Histogramm überprüft.

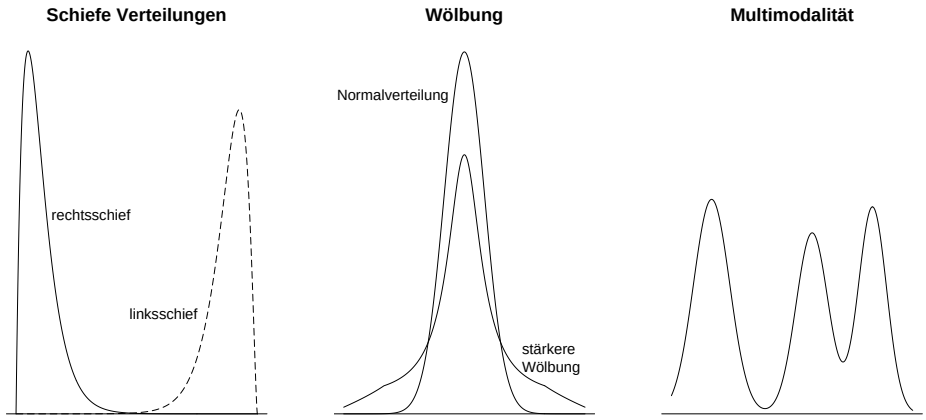
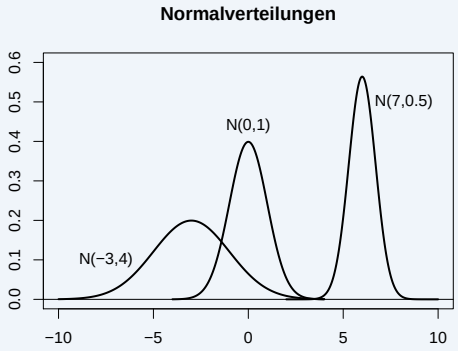


Abbildung 8.4: Verteilungsformen

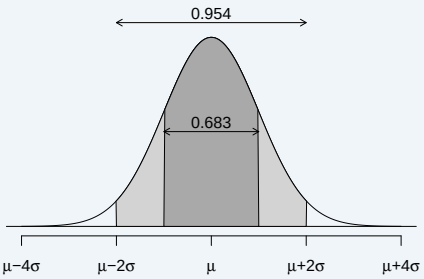
In Abbildung 8.4 sind mehrere Verteilungsformen dargestellt. In der linken Grafik sind schiefe Verteilungen abgebildet. Die mittlere Grafik enthält symmetrische Verteilungen, eine Normalverteilung und dazu eine Verteilung mit stärkerer Wölbung. Die Verteilung rechts ist nicht ein-, sondern mehrgipfelig (multimodal).

Exkurs 8.2: Normalverteilung

Die Normalverteilung ist eine theoretische Verteilung, deren Form durch die berühmte Glockenkurve bestimmt ist.
Normalverteilung steht nicht für eine einzelne Verteilung, sondern für eine Familie von Verteilungen, die durch zwei Parameter (Kenngrößen) bestimmt sind, die mit μ und σ^2 bezeichnet werden. μ ist dabei der Erwartungswert (Mittelwert), σ^2 die Varianz (somit σ die Standardabweichung) der theoretischen Verteilung. Die Kurzschreibweise ist: $N(\mu, \sigma^2)$.



Normalverteilungen sind symmetrisch und eingipfelig. Im zentralen Bereich sind die Daten am stärksten konzentriert. Im Bereich $\mu \pm \sigma$ (also vom Mittelwert eine Standardabweichung nach links und nach rechts) liegen 68.3% (also etwas mehr als zwei Drittel) und im Bereich $\mu \pm 2\sigma$ sind es 95.4% der Daten.



Mit der Wahl $\mu \pm 1.96\sigma$ überdeckt man genau 95% der Daten. Dieser Wert von 1.96 ist schon bei Konfidenzintervallen für Anteile (▶ Abschnitt 6.5) in der Formel 6.3 ohne große Erklärung aufgetaucht. Ersetzt man 1.96 durch 2.58, sind 99% der Daten im zentralen Bereich.

Diese Anteile gelten für normalverteilte Variablen. Wenn die Verteilung symmetrisch und eingipfelig ist, sind die Abweichungen von diesen Werten aber nicht sehr groß, wenn keine Normalverteilung vorliegt.

8.1.3 Boxplot

Minimum, 1. Quartil, Median, 3. Quartil und Maximum werden oft als 5-Punkt-Zusammenfassung für eine Variable angegeben. Darin enthalten sind mit dem Median ein Lagemaß und implizit auch Spannweite und Quartilsabstand, also zwei Streuungsmaße.

Auf diesen fünf Punkten ist auch der Boxplot aufgebaut. Die Grenzen der Box sind durch die Quartile bestimmt, an der Stelle des Medians ist die Box unterteilt. Linien (Whiskers) zum Minimum und Maximum vervollständigen den Boxplot.

```
> boxplot(vwghdauer, ylab = "Dauer in Tagen")
```

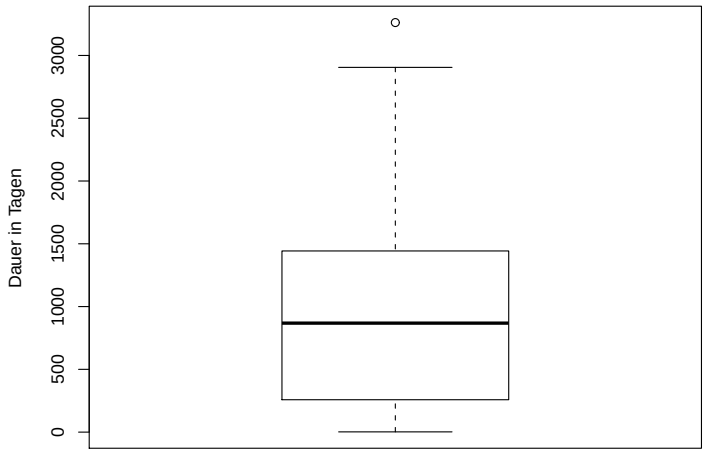


Abbildung 8.5: Boxplot der VwGH-Verfahrensdauern

Statistikpakete wie R hängen meist eine Ausreißersuche an, die das Erscheinungsbild leicht abändern kann. Ausreißer werden mit als einzelne Punkte markiert und die Linien werden nicht bis zu den Ausreißern gezogen.

Boxplots können platzsparend auch liegend dargestellt werden (▶ Abbildung 8.6). Beim folgenden liegenden Boxplot (`horizontal=TRUE`) für die Verfahrensdauern ist überdies die Ausreißersuche abgestellt (`range=0`), der Mittelwert markiert (`abline()`), die Skalenbeschriftung verfeinert (`seq(0,3500,250)` und senkrecht zur Achse (`las=2`) angebracht.

R

```
> boxplot(vwghdauer, horizontal = TRUE, axes = FALSE,
+     range = 0, cex.main = 1.5, main = "VwGH-Verfahrensdauern",
+     xlab = "Dauer in Tagen")
> axis(1, seq(0, 3500, 250), las = 2)
> abline(v = mean(vwghdauer))
```

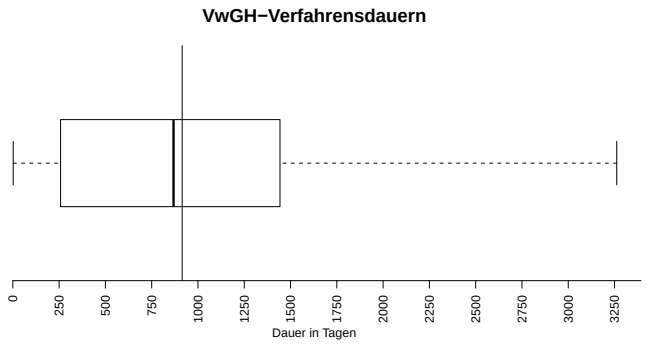


Abbildung 8.6: Boxplot der VwGH-Verfahrensdauern

Fallbeispiel 12: Interpretation des Boxplots

Wenn wir zur Interpretation den zweiten Boxplot (▶ Abbildung 8.6) mit der genaueren Skalenbeschriftung heranziehen, können wir einiges über die Verfahrensdauern vor dem VwGH ableiten. Diese Skala erlaubt nämlich das ungefähre Ablesen wichtiger Werte.

- Q_1 als untere Begrenzung der Box ist etwas größer als 250, Q_3 als obere Begrenzung ist etwas kleiner als 1500.
- Der Median als Unterteilung der Box liegt zwischen 750 und 1000.
- Der Quartilsabstand als Differenz $Q_3 - Q_1$ ist etwas größer als 1000 und kann aus der Länge der Box abgeschätzt werden.
- Das Minimum ist ca. 0 und das Maximum liegt etwas über 3250.

Das erste Viertel der Daten ist somit auf den Bereich zwischen 0 und ca. 250 konzentriert, das nächste Viertel geht bis knapp unter 900, ist also breiter. Das dritte Viertel reicht nicht ganz bis 1500, ist also in etwa gleich breit wie das vorige Viertel. Das letzte Viertel ist allein mindestens so breit wie die vorigen drei Viertel.
Daraus können wir auf eine schiefe Verteilung schließen; da die Daten zu Beginn stärker konzentriert sind, ist die Verteilung rechtsschief.

Zur Beschreibung der Verteilung sind Histogramme besser als Boxplots geeignet. So konnten wir aus dem Histogramm (▶ Abbildung 8.2) ersehen, dass eine zweigipfelige Verteilung vorliegt. Der Boxplot bietet diese Einsicht nicht.

Der Vorteil von Boxplots liegt im grafischen Vergleich von Verteilungen einer Variablen in mehreren Gruppen und wird uns im Kapitel 10 wieder begegnen.

8.1.4 Ausreißer

Im Boxplot (▶ Abbildung 8.5) ist eine Beobachtung als Ausreißer markiert worden. Ausreisser sind allgemein Beobachtungen, die nicht zur selben Grundgesamtheit gehören oder zu gehören scheinen wie die (meisten) übrigen Elemente der untersuchten Stichprobe.

Liegt nur eine metrische Variable pro Beobachtung vor, fallen Ausreißer durch extreme Werte in dieser Variablen auf. Bei mehreren Variablen können Extremwerte in einer Variablen zur Entdeckung von Ausreißern führen, in komplizierteren Fällen sind sie besser versteckt und können nur durch die simultane Untersuchung mehrerer Variablen ausspioniert werden.

Ursachen von Ausreißern

Ausreißer können aus mehreren Gründen in Datensätzen auftauchen:

- Fehler bei der Datenaufnahme, etwa durch ein defektes Messgerät
- Codierfehler; unklare Maßeinheiten, etwa Körpergröße wird in Metern statt Zentimetern angegeben.
- Schreib- oder Tippfehler
- Ausreißer weicht zwar deutlich von den meisten anderen Werten ab, ist aber durchaus denkbar (reliable Ausreißer).

Umgang mit Ausreißern

Ausreißer können Ergebnisse statistischer Auswertungen stark beeinflussen, in schlechten Fällen verfälschen. Was macht man mit Beobachtungen, die man als Ausreißer entdeckt hat?

Natürlich wird man bei eindeutigen Datenfehlern versuchen, den Fehler zu korrigieren. Ist eine Korrektur nicht möglich, muss diese Beobachtung weggelassen werden bzw. der fehlerhafte Variablenwert auf fehlender Wert (missing) gesetzt werden.

Schwieriger ist der Umgang mit reliablen Ausreißern. Eine Möglichkeit ist, Analysen ohne diese Ausreißer durchzuführen, aber dieses Weglassen auch zu dokumentieren. Eine andere Möglichkeit ist, die Resultate der Analyse mit und ohne Ausreißer zu berichten.

Für manche Fragestellungen gibt es statistische Verfahren, die gegenüber vorhandenen Ausreißern wenig empfindlich sind. Solche Verfahren werden als ROBUST bezeichnet und sie stellen eine weitere Möglichkeit dar, möglichen Ausreißern in den Daten zu begegnen.

Median und auch das getrimmte Mittel sind robuste Lagemaße, der Quartilsabstand ist ein robustes Streuungsmaß.

Fallbeispiel 12: Interpretation des Ausreißers

Eine Überprüfung der Beobachtung mit dem Wert 3262 für die Verfahrensdauer konnte Schreib- und Tippfehler ausschließen, es ist also ein reliabler Ausreißer. Ursachenforschung für die Länge des Verfahrens ist nicht Aufgabe dieses Buchs. Was sind die Auswirkungen auf die Maßzahlen?
Der Ausreißer im Datensatz ist kein Grund zu großer Sorge. Einerseits ist er nur als moderat eingestuft worden, andererseits ist bei einer so großen Stichprobe (n=3745) eine etwas abweichende Beobachtung nicht sehr einflussreich. Das zeigt auch die folgende Auflistung einiger Maßzahlen:

	mit Ausreißer	ohne Ausreißer
$\bar{x}$	914.9	914.2
$\tilde{x}$	868.0	868.0
s	680.4	679.4
QD	1185.0	1184.5

Mittelwert und Standardabweichung haben eine geringe Änderung durch das Weglassen des Ausreißers erfahren, die robusten Maße haben überhaupt nicht ($\tilde{x}$) oder nur minimal (QD) reagiert.

8.1.5 Weitere grafische Beschreibungsmethoden

Histogramme und Boxplots sind nicht die einzigen grafischen Beschreibungsverfahren für metrische Variablen. Einige weitere Verfahren werden hier vorgestellt; ihr Einsatz ist aber weniger häufig und nur bei moderatem Stichprobenumfang sinnvoll. Wir wechseln zu einem Datensatz aus dem Golfsport.

Fallbeispiel 13: Golf: US-Masters in Augusta 2009

Datenfile: `augusta2009.csv`

Eines der traditionsreichen Turniere im Golf ist das US-Masters in Augusta, das auf dem sehr berühmten Platz des Augusta National Golf Club gespielt wird.
Dieser 18-Loch-Platz hat Bahnen unterschiedlichen Schwierigkeitsgrades; für manche Bahnen werden von einem guten Spieler drei Schläge, für die meisten vier Schläge und für einige schwierig zu spielende Bahnen fünf Schläge bis zum Einlochen erwartet. Der Platz ist so angelegt, dass der Sollwert für eine Runde (also alle 18 Bahnen) bei 72 Schlägen liegt.

Nach zwei Spielrunden scheiden die schlechter platzierten Spieler aus, die anderen spielen weitere zwei Runden. Im Datenfile sind die Ergebnisse des Masters aus dem Jahr 2009 enthalten, nämlich für jede der vier Runden. Sieger wurde Angel Cabrera, der nach vier Runden mit 276 Schlägen wie Chad Campbell und Kenny Perry 12 unter Par war und das notwendige Playoff gewinnen konnte.

Fragestellung:
Gibt es weitere Verfahren zur grafischen Beschreibung?

Stem-and-Leaf-Plot

Stem-and-Leaf-Plots boten die Möglichkeit einer grafischen Darstellung einer Verteilung schon zu Zeiten, als für Drucker kaum mehr als der Zeichensatz einer Schreibmaschine verfügbar war. Bei kleinen Datensätzen bieten sie nicht nur – ähnlich wie Histogramme – eine Übersicht nicht nur über die Verteilung, sondern sogar über zeigen die einzelnen Werte zumindest gerundet an.

Am Beispiel der Golfdaten sei dies demonstriert. Wir beschränken uns auf die nach den ersten zwei Runden besten 50 Spieler, die den Cut geschafft haben. In `Strokes` sind die insgesamt benötigten Schläge für alle absolvierten Runden enthalten. Da in dieser Variablen allerdings auch die benötigten Schläge der Spieler enthalten sind, die den Cut nicht geschafft haben, müssen wir eine Auswahl auf jene Spieler treffen, die auch die dritte – und somit auch die vierte – Runde gespielt haben.

R

```
> golf <- read.csv2("augusta2009.csv", header = TRUE)
> attach(golf)
> Gesamt <- Strokes[!is.na(R3)]
> detach(golf)
> stem(Gesamt, scale = 0.5)
```

```
The decimal point is 1 digit(s) to the right of the |

27 | 66689
28 | 0000111223344
28 | 56666666666777778889999
29 | 0133444
29 | 88
```

Abbildung 8.7: Stem-and-Leaf-Plot der Golfdaten

Welche Informationen sind im Plot (▶ Abbildung 8.7) enthalten?

- Die einzelnen Zahlen des Datensatzes werden in einen Stamm (Stem) und ein Blatt (Leaf) aufgeteilt. Die Angaben zum Stamm sind links, diejenigen zum Blatt einer Zahl sind rechts vom |-Zeichen zu finden. Die Angabe `The decimal point is 1 digit(s) to the right of the |` besagt, dass die Angaben zum Stamm mit zehn zu multiplizieren sind. Es liegen also Stämme mit den Größen 270, 280 und 290 vor. Die Stämme 280 und 290 treten zweimal auf. Je einmal für die niedrigen Werte 280–284 (bzw. 290–294) und einmal für die hohen Werte 285–289 (bzw. 295–299).
- Die einzelnen Werte des Datensatzes könnte man rekonstruieren, indem man Stamm und Blatt der einzelnen Beobachtungen zusammenführt. Wenn wir also mit der niedrigen 280er Klasse beginnen: es gibt viermal eine 0, daher also insgesamt viermal den Wert 280. Analog interpretierend kann man ableiten, dass dreimal 281 und je zweimal 282, 283 und 284 auftreten.
- Die Werte liegen somit auch in sortierter Reihenfolge vor. Es ist leicht, etwa den viertkleinsten Wert (278) zu bestimmen.
- Der Stem-and-Leaf-Plot ist ein um 90 Grad gedrehtes Histogramm (mit konstanten Klssenbreiten).

Da die Stichprobenwerte mit wenigen Ziffern angezeigt werden müssen, ist oft eine Rundung notwendig. In diesem Fall ist eine exakte Rekonstruktion der ursprünglichen Daten nicht mehr möglich.

Punkt- und Stabdiagramme

Den Versuch, die Einzeldaten als Punkte anzuzeigen, unternehmen PUNKTDIAGRAMME . In R steht dazu die Funktion `stripchart` zur Verfügung. Den grafisch ansprechenderen Output kann man mit der Funktion `DOTplot` aus dem **UsingR**-Package erstellen.

```
> library("UsingR")
> DOTplot(Gesamt, main = "US-Masters 2009")
```

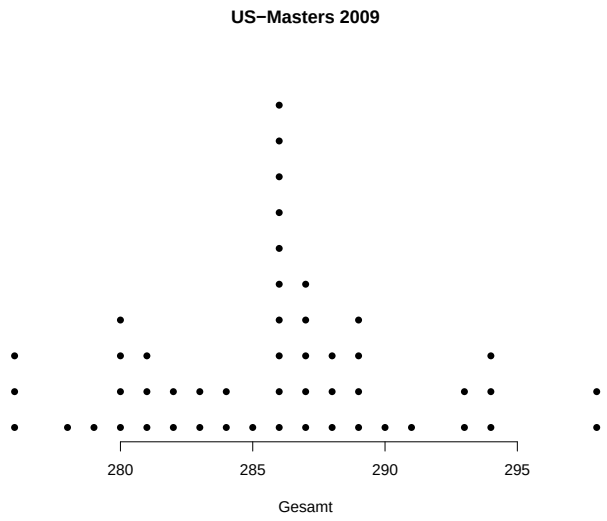


Abbildung 8.8: Punktdiagramm der Golfdaten

Eine ähnliche Idee wie mit Punktdiagrammen verfolgt man mit STABDIAGRAMMEN. Statt Punkte werden Striche zur Markierung der Beobachtungen verwendet. In R muss man etwas tricksen, um zu einer guten Achsenbeschriftung zu gelangen.

```
> plot(table(Gesamt), main = "US-Masters 2009", xlab = "Gesamt",
+     ylab = "Anzahl", cex.main = 1.4, axes = FALSE)
> axis(1)
> axis(2)
```

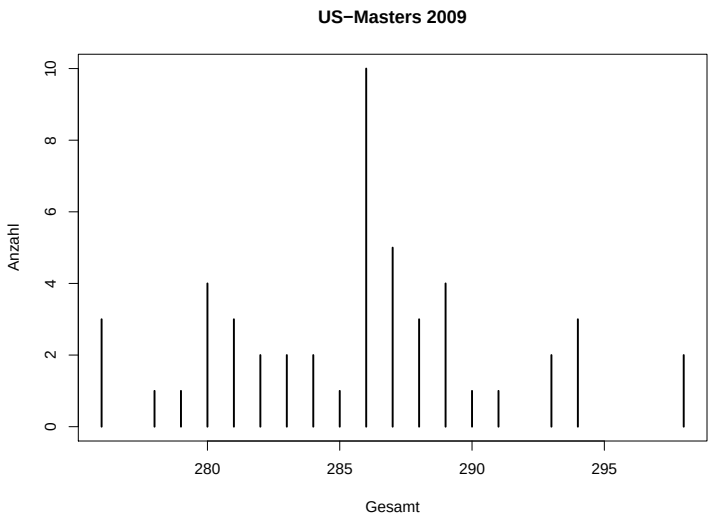


Abbildung 8.9: Stabdiagramm der Golfdaten

Beide Diagramme (▶ Abbildung 8.8 und ▶ Abbildung 8.9) vermitteln denselben Eindruck über die Lage der Daten. Zuerst kommen die drei Spieler mit dem Minimum an Schlägen, die das Playoff bestritten. Dann folgt der Hauptteil der Spieler mit zwischen 280 und 290 Schlägen. Ein paar Spieler sind deutlich abgefallen (mindestens 293 Schläge).

Empirische Verteilungsfunktion

Die EMPIRISCHE VERTEILUNGSFUNKTION gibt für jeden Wert aus dem Wertebereich der Stichprobe den Anteil an Beobachtungen an, die diesen Wert nicht übersteigen. Inhaltlich entspricht sie kumulierten relativen Häufigkeiten und wird lieber grafisch als tabellarisch ausgegeben.

R

```
> plot(ecdf(Gesamt), main = "US-Masters 2009", xlab = "Gesamt")
```

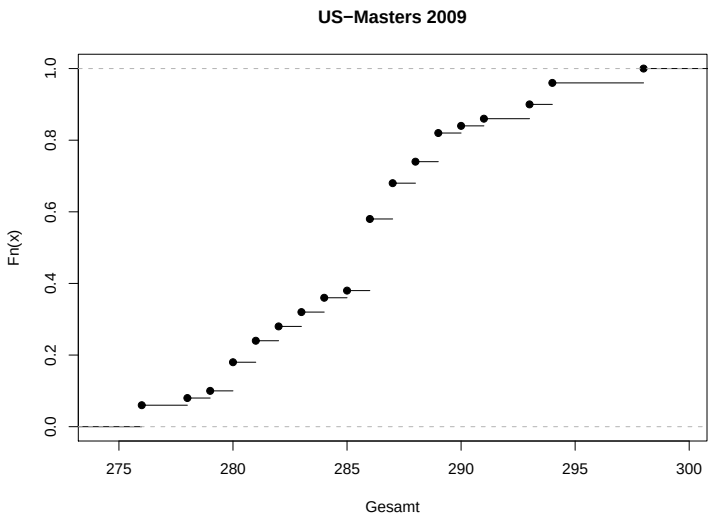


Abbildung 8.10: Empirische Verteilungsfunktion der Golfdaten

Das Ergebnis ist eine Treppenfunktion. Die erste Stufe ist an der Stelle 276, dem Minimum der Daten, die letzte Stufe ist bei 298, dem Maximum der Daten. Die Stufen sind unterschiedlich hoch, je nach Häufigkeit der einzelnen Werte; die höchste Stufe ist bei 286, dem häufigsten Wert in der Stichprobe. Hier macht die empirische Verteilungsfunktion einen Sprung von ungefähr 0.4 auf 0.6. Ungefähr 60 Prozent der Beobachtungen haben einen Wert von höchstens 286.

8.2 Ist ein Mittelwert in der Grundgesamtheit anders als eine bestimmte Vorgabe?

Fallbeispiel 14: Dauern die Verfahren am VwGH länger?

Datenfile: `vwgh.csv`

Wir haben im vorigen Abschnitt die Verfahrensdauern in Abgabensachen am Verwaltungsgerichtsgerichtshof untersucht. Eine frühere Untersuchung über die VwGH-Entscheidungen der Jahre 1979 bis 1985 hatte ergeben, dass in diesen Jahren die durchschnittliche Verfahrensdauer 1 Jahr und 3 Monate betragen hatte.

Fragestellung:
Dauern die Verfahren durchschnittlich länger als in den Jahren 1979 bis 1985?

In der Untersuchung der VwGH-Entscheidungen der Jahre 1979 bis 1985 wurden die Verfahrensdauern in Monaten, in unseren Daten über die Jahre 2000 bis 2004 in Tagen erhoben. Für den Vergleich mit unseren Daten müssen wir den Wert von einem Jahr und drei Monaten umrechnen, wir erhalten 456 ($= 365 \cdot 1.25$) Tage.

R

```
> boxplot(vwghdauer, main = "VwGH-Verfahrensdauern",
+     cex.main = 1.5, xlab = "Dauer in Tagen", cex.lab = 1.3,
+     horizontal = TRUE)
> abline(v = 456)
```

In den Boxplot (▶ Abbildung 8.11) ist der Vergleichswert 456 als Referenzlinie eingezeichnet.

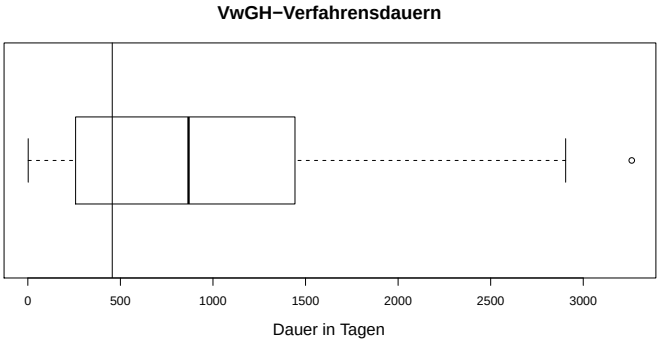


Abbildung 8.11: Boxplot der VwGH-Verfahrensdauern mit Vergleichswert

Exkurs 8.3: Zentraler Grenzwertsatz

In der Wahrscheinlichkeitsrechnung ist es eine einfache Übung nachzuweisen, dass Mittelwerte aus einer normalverteilten Grundgesamtheit ebenfalls normalverteilt sind. Gemeint ist dabei Folgendes (ähnlich ▶ Abschnitt 6.9):

- Man zieht wiederholt Stichproben eines festgelegten Stichprobenumfangs.
- In jeder Stichprobe wird der Mittelwert bestimmt.
- Wenn der Vorgang oft (etwa 10 000-mal) wiederholt wird, erhalten wir viele Mittelwerte.
- Die Verteilung dieser Werte kann z.B. in einem Histogramm dargestellt werden. Die Form des Histogramms wird sehr ähnlich der Glockenform der Normalverteilung sein.

In der grafischen Beschreibung der Verfahrensdauern haben wir aber eine schiefe und zweigipfelige Verteilung festgestellt. Zwar trifft das nur auf die Stichprobe zu, die Abweichungen von einer Normalverteilung sind aber so groß, dass wir auch für die Grundgesamtheit annehmen können, dass keine Normalverteilung vorliegt (für Tests ▶ Abschnitt 8.4). Was gilt für Mittelwerte aus solchen Grundgesamtheiten?

Hier hilft der ZENTRALE GRENZWERTSATZ:

> Für große Zufallsstichproben sind die Mittelwerte approximativ normalverteilt.

Was bedeutet das?

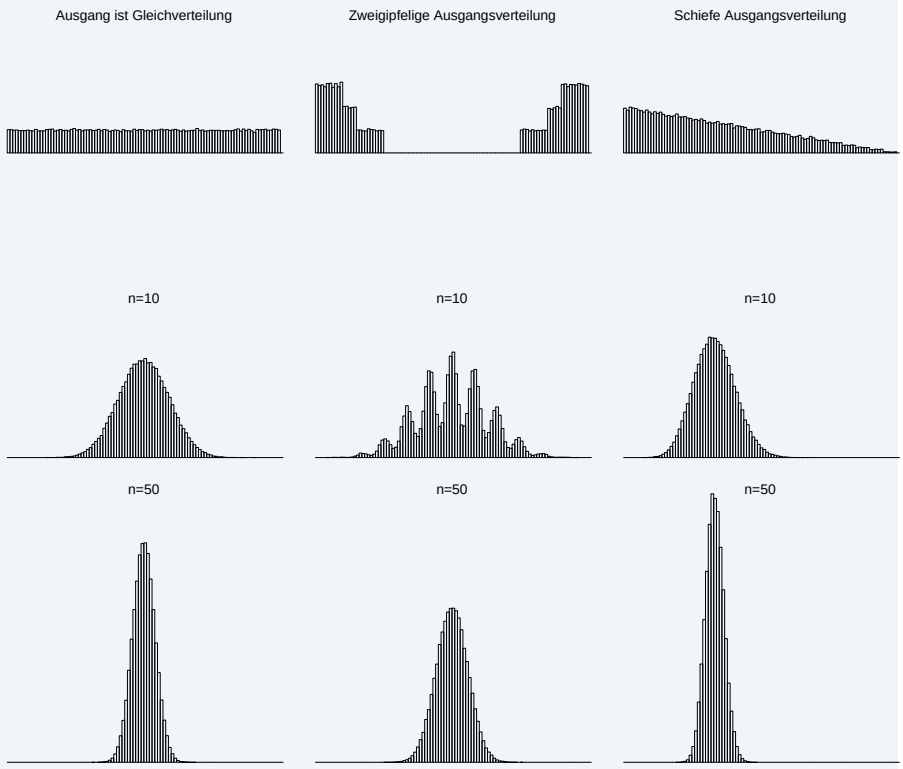
- Die Aussage gilt für großes n, in mathematischer Formulierung noch abschreckender für $n \to \infty$. Simulationen zeigen, dass schon für moderates n ($n \geq 30$) die Abweichungen von der Normalverteilung nur mehr gering sind.
- Die approximative Normalverteilung wird auch bei extremen Ausgangsverteilungen erreicht. Also auch schiefe, mehrgipfelige oder auch diskrete Verteilungen führen bei ausreichend großem n zu ungefährer Normalverteilung des Mittelwerts. In die Bestimmung des Konfidenzintervalls für Anteile (▶ Abschnitt 6.5) ist auch der zentrale Grenzwertsatz eingeflossen.
- Der Gipfel in der Mittelwertsverteilung ist beim Mittelwert der Ausgangsverteilung.
- Die Varianz der Mittelwerte fällt mit dem Stichprobenumfang gemäß $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Mit dem folgenden Plot sollen die eben besprochenen Punkte an Ausgangsverteilungen, die klar von der Normalverteilung abweichen, veranschaulicht werden. Die Ausgangsverteilungen (obere Zeile) sind:

- Gleichverteilung: eine zwar symmetrische Verteilung aber ohne klaren Gipfel
- U-förmige Verteilung: symmetrisch, aber zwei Gipfel, noch dazu an den Enden der Verteilung
- Schiefe Verteilung: nicht symmetrisch, Gipfel am linken Ende der Verteilung

Für die Simulation wurden jeweils 100 000 Mittelwerte berechnet und diese in Histogrammen zusammengefasst.

Exkurs 8.3: (Fortsetzung)



In der zweiten bzw. dritten Zeile sind die Verteilungen der Mittelwerte von jeweils 10 bzw. 50 Beobachtungen geplottet. Spätestens bei Mittelwerten aus 50 Beobachtungen liegt ungefähre Normalverteilung vor. Man erkennt auch, dass die Varianz der Mittelwerte mit zunehmendem Stichprobenumfang kleiner wird.

Das Histogramm (▶ Abbildung 8.2) der VwGH-Verfahrensdauern zeigt eine schiefe und mehrgipfelige Verteilung an, also eine starke Abweichung von der Normalverteilung. Aufgrund des sehr hohen Stichprobenumfanges kann man über den zentralen Grenzwertsatz argumentieren, dass für den Mittelwert dennoch eine Normalverteilung vorliegt.

Damit kann der Ein-Stichproben-t-Test zur Untersuchung der Fragestellung eingesetzt werden. Wie beim Anteilstest (▶ Abschnitt 6.4) können zwei- oder einseitige Alternativhypothesen überprüft werden.

Ein-Stichproben-t-Test eines Mittelwerts

Nullhypothese H_0: $\mu = \mu_0$

Alternativhypothese H_A: $\mu \neq \mu_0$ oder $\mu > \mu_0$ oder $\mu < \mu_0$

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} \tag{8.1}$$

- μ ... der unbekannte Mittelwert der Grundgesamtheit (hier aller VwGH-Entscheidungen)
- μ_0 ... der Wert, den wir kennen (oder den wir festlegen) und gegen den wir prüfen wollen, in unserem Beispiel ist er 456.
- $\bar{x}$ ... Mittelwert der Stichprobe
- n ... Stichprobenumfang
- s ... Standardabweichung in der Stichprobe
- Die Teststatistik t folgt unter H_0 annähernd einer t-Verteilung mit $n - 1$ Freiheitsgraden.
- Die Fragestellung, ob die Verfahren länger dauern, entspricht der Alternativhypothese $H_A : \mu > 456$.

R

```
> t.test(vwghdauer, mu = 456, alternative = "greater")
```

```
        One Sample t-test

data:  vwghdauer
t = 41.2706, df = 3744, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 456
95 percent confidence interval:
 896.5606      Inf
sample estimates:
mean of x
 914.8529
```

Abbildung 8.12: t-Test für den Mittelwert der VwGH-Verfahrensdauern

Fallbeispiel 14: Interpretation des Mittelwerttests

Der Titel des Testergebnisses (▶ Abbildung 8.12) besagt, dass ein Ein-Stichproben-t-Test berechnet wurde. Nach der Angabe, für welche Daten der Test berechnet wurde, werden die eigentlichen Testergebnisse ptäsentiert:

- Der Wert der Teststatistik beträgt: 41.2706.
- Unter H_0 folgt die Teststatistik einer t-Verteilung mit 3744 Freiheitsgraden (es liegen $n = 3745$ Verfahrensdauern vor).
- Der p-Wert ist fast 0.
- Es wurde der Test mit der einseitigen Alternativhypothese $H_A : \mu > 456$ gerechnet.

Danach folgt noch die Ausgabe eines einseitigen Konfidenzintervalls (▶ Abschnitt 8.3) und des aus der Stichprobe berechneten Mittelwerts für die Verfahrensdauern (914.8529).

Aufgrund des sehr kleinen p-Werts wird die Nullhypothese verworfen und die Alternativhypothese angenommen. Inhaltlich bestätigt das Testergebnis die Vermutung, die man aufgrund der Stichprobe schon gewonnen hat: Der Mittelwert ist ziemlich genau das Doppelte des Werts für den Zeitraum 1979 bis 1985. Die Verfahren dauern durchschnittlich länger.
Juristen führen die längeren Verfahrensdauern zum Teil auf die größere Anzahl von Beschwerden, die vor den VwGH gebracht werden, und die dadurch verursachte Überlastung des VwGH zurück.

8.3 In welchem Bereich kann man den Mittelwert in einer Grundgesamtheit erwarten?

Fallbeispiel 15: Wie lange dauern die Verfahren am VwGH?

Datenfile: `vwgh.csv`

Wir haben die Verteilung der Verfahrensdauern durch Histogramme und Boxplots grafisch und durch Maßzahlen numerisch beschrieben.
Das hat die Stichprobe betroffen. Aber weitere Beobachtungen würden zu geänderten Grafiken und Maßzahlen führen.

Fragestellung:
Wie groß ist der Mittelwert aller Verfahrensdauern?

Wir haben eine sehr große Stichprobe als Basis unserer Maßzahlen, dennoch würden weitere Beobachtungen bewirken, dass sich die Werte – vermutlich nur leicht, aber dennoch – verändern. Eine völlig andere Stichprobe würde auch kaum genau die Werte unserer Stichprobe reproduzieren. Den Mittelwert der Stichprobe als Wahrheit, also als den Mittelwert der Grundgesamtheit auszugeben, wäre also entweder naiv oder überheblich.

Die Idee, die zur Anwendung kommt, ist analog der bei Konfidenzintervallen für Anteile (▶ Abschnitt 6.5). Man ersetzt den einzelnen Wert (hier den Mittelwert) der Stichprobe durch ein Intervall, in dem vermutlich der unbekannte Mittelwert der Grundgesamtheit liegt. Das Resultat ist ein Konfidenzintervall für den Mittelwert, das durch die Angabe der Unter- und Obergrenze des Intervalls festgelegt ist.

Sind die Daten normalverteilt oder kann, weil der Stichprobenumfang ausreichend groß ist, aufgrund des zentralen Grenzwertsatzes auf eine Normalverteilung des Stichprobenmittels geschlossen werden, können Konfidenzintervalle über die Normalverteilung berechnet werden.

In R werden Konfidenzintervalle für den Mittelwert automatisch auch bei jedem Aufruf eines Einstichproben-t-Tests berechnet (▶ Abschnitt 8.2). Allerdings werden bei einseitigen Alternativhypothesen nur die weniger üblichen einseitigen Konfidenzintervalle erstellt. Für die gewohnteren zweiseitigen Konfidenzintervalle genügt in R die Berechnung eines zweiseitigen t-Tests mit beliebigem Wert in der Nullhypothese. Ist man nur am Konfidenzintervall aber nicht an den Ausgabewerten zum t-Test interessiert, genügt:

R

```
> t.test(vwghdauer)$conf.int
```

```
[1] 893.0547 936.6511
attr(,"conf.level")
[1] 0.95
```

Fallbeispiel 15: Interpretation des Konfidenzintervalls

Trotz schiefer Ausgangsverteilung können wir wegen des sehr großen Stichprobenumfanges auf die Wirkung des zentralen Grenzwertsatzes auf die Verteilung des Stichprobenmittels hoffen und gelangen zu den Grenzen des Konfidenzintervalls: 893.05 und 936.65.
In diesem Bereich liegt vermutlich der Mittelwert der Verfahrensdauern **aller** Beschwerden an den VwGH.

Will man das Konfidenzniveau (Sicherheitsniveau) von den standardmäßig eingestellten 95% abändern (etwa auf 99%), kann im Aufruf von `t.test()` das Argument `conf.level=0.99` angegeben werden.

8.4 Folgt eine metrische Variable einer bestimmten Verteilung?

Fallbeispiel 16: Normalverteilung beim US-Masters 2009?

Datenfile: `augusta2009.csv`

Ein Rundenergebnis bei einem Golfturnier setzt sich aus 18 Teilergebnissen auf den einzelnen Bahnen zusammen. Solche Summen aus Einzelergebnissen lassen sich oft durch Normalverteilungen beschreiben. So auch die Ergebnisse früherer US-Masters.

Fragestellung:
Sind die benötigten Schläge für die vier Runden normalverteilt?

Wir werden uns auf den Turnierendstand beschränken, ein Übungsbeispiel befasst sich mit Stand nach zwei Runden.

Eine Möglichkeit einer grafischen Überprüfung bietet ein Histogramm. Zwar sind nur ganze Zahlen denkbar (die Rundenergebnisse sind diskret), aber ein Histogramm gibt einen besseren Eindruck einer Verteilung einer metrischen Variablen als ein Balkendiagramm, das Platz zwischen den Balken lässt.

In dieses Histogramm (▶ Abbildung 8.13) zeichnen wir die Dichtefunktion (Glockenkurve) jener Normalverteilung ein, bei der Mittelwert und Varianz mit der Stichprobe übereinstimmen. Dazu wird an vielen Stellen `xx` die Dichtefunktion der Normalverteilung berechnet (`dnorm()`). Mit `lines()` werden die so berechneten Punkte durch Strecken verbunden.

R

```
> hist(Gesamt, freq = FALSE, main = "US-Masters 2009")
> xx <- seq(min(Gesamt), max(Gesamt), 0.01)
> lines(xx, dnorm(xx, mean = mean(Gesamt), sd = sd(Gesamt)))
```

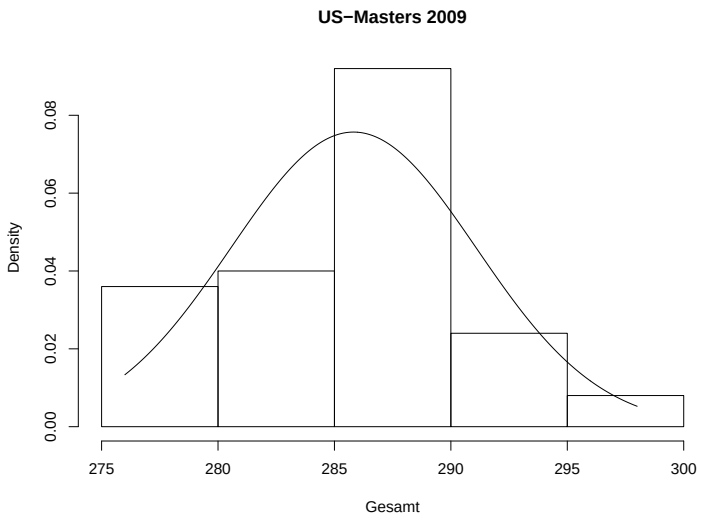


Abbildung 8.13: Histogramm der Golfdaten mit Normalverteilungskurve

Die Unterschiede zwischen Balkenhöhen und den Werten der Dichtefunktion (▶ Abbildung 8.13) sind nicht sehr groß. Das Histogramm ist nicht ganz symmetrisch, Abweichungen wie hier können in Stichproben auftreten und sprechen noch nicht gegen eine Normalverteilung in der Grundgesamtheit.

Methoden, die über den Vergleich von einem Histogramm mit einer Normalverteilungskurve hinausgehen, haben mehrere Begründungen:

- Nicht immer ist die Stichprobe so groß, dass ein Histogramm sinnvoll erstellt werden kann.
- Man will sich nicht nur mit einem grafischen Überblick zufrieden geben, man will auch einen Test, der zu einer Entscheidung führt, anwenden.
- Am häufigsten wird die Frage nach einer Normalverteilung gestellt. Vergleiche gegen andere Verteilungen kommen aber auch vor.

8.4.1 Q-Q-Plot

Auch bei geringem Stichprobenumfang und für viele Verteilungen kann in R ein Q-Q-Plot (Quantil-Quantil-Plot) erstellt werden.

Elemente des Q-Q-Plots

Testverteilung: Die zu testende Verteilung kann konkret mit allen Parameterwerten spezifiziert sein. Der häufigere Fall ist aber der, dass nicht alle Parameterwerte spezifiziert sind, sondern aus der Stichprobe geschätzt werden. Im Beispiel der Golfdaten kann etwa nur Normalverteilung überprüft werden, für μ und σ werden Mittelwert und Standardabweichung der Stichprobe verwendet.

Beobachtete Quantile: Die Stichprobenwerte (hier von 50 Teilnehmern die Anzahl der Schläge in den vier Runden) sind die beobachteten Quantile (hier die Quantile für 1%, 3%, 5%, ..., 99%).

Erwartete Quantile: Die erwarteten Quantile werden aus der zu testenden Verteilung berechnet.

Q-Q-Plot: Der Plot selbst ist ein Streudiagramm. Pro Beobachtung bestimmen erwartetes und beobachtetes Quantil einen Punkt im Diagramm.

Idealbild: Für die zu testende Verteilung spricht im Idealfall, wenn beobachtete und erwartete Quantile genau übereinstimmen. Natürlich ist das in einer konkreten Stichprobe nie anzutreffen, leichte Abweichungen davon werden kaum Argwohn erwecken. Systematische Abweichungen fallen auf und bieten Hinweise auf Abweichungen von der Testverteilung.

Die notwendigen Operationen stellen in R keine hohe Hürde dar:

R

```
> n <- length(Gesamt)
> xx <- (1:n - 0.5)/n
> quantil_beob <- sort(Gesamt)
> quantil_erw <- qnorm(xx, mean = mean(Gesamt),
+     sd = sd(Gesamt))
> plot(quantil_erw, quantil_beob, xlab = "Theoretische Quantile",
+     ylab = "Beobachtete Quantile")
> abline(0, 1)
```

Für einen Normalverteilungs-Q-Q-Plot kann eine eigene R-Funktion (`qqnorm()`) verwendet werden, eine kleine Änderung ist bei den angezeigten erwarteten Quantilen zu beobachten. Einen Plot wie soeben kann ebenfalls mit mit einer R-Funktion (`qqplot()`) erstellt werden, hier geben wir den Lageparameter mit 288 (vier Runden zu je 72 Schlägen) vor und schätzen ihn nicht aus der Stichprobe.

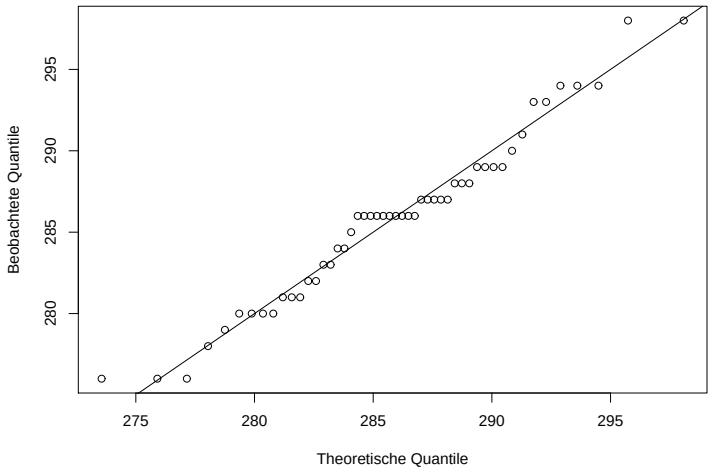


Abbildung 8.14: Q-Q-Plot der Golfdaten

R

```
> par(mfrow = c(1, 2))
> qqnorm(Gesamt, main = "N(.,.)")
> qqline(Gesamt)
> n <- length(Gesamt)
> xx <- (1:n - 0.5)/n
> quantil_erw <- qnorm(xx, mean = 288, sd = sd(Gesamt))
> qqplot(quantil_erw, Gesamt, main = "N(288,.)")
> abline(0, 1)
> par(mfrow = c(1, 1))
```

Anmerkungen zu Q-Q-Plots:

- Liegen Punkte genau auf der eingezeichneten Gerade im ersten Q-Q-Plot (▶ Abbildung 8.14), so sind erwratetes und beobachtetes Quantil ident. Diese Gerade zeigt also das Idealbild, dass die Stichprobenverteilung exakt der unterstellten Verteilung entspricht.
- Im zweiten Q-Q-Plot (▶ Abbildung 8.15 links) sind auf der x-Achse nicht die Quantile der unterstellten Normalverteilung, sondern die der Standardnormalverteilung aufgetragen. Die eingezeichnete Gerade geht durch die Punkte

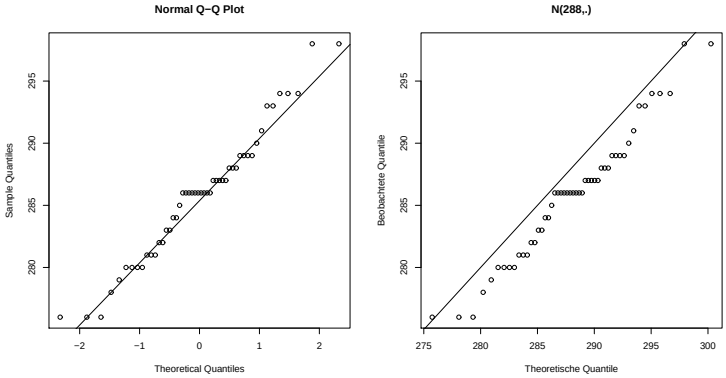


Abbildung 8.15: Weitere Q-Q-Plots der Golfdaten

der ersten bzw. dritten Quartile und stellt ebenfalls einen Anhaltspunkt für die Interpretation dar.

- Üblicherweise werden die Parameterwerte für die Verteilungen aus den Daten geschätzt. Man kann aber auch die Parameter der Verteilungen fixieren, wie im letzten Beispiel (▶ Abbildung 8.15 rechts), wo der Mittelwert mit 288 (`mean=288`) festgelegt wurde.

- Q-Q-Plots mit anderen Verteilungen werden durch die entsprechende Berechnung der erwarteten Quantile realisiert; also statt `qnorm()` durch `qunif()`, wenn statt der Normalverteilung eine Gleichverteilung zur Anwendung kommen soll.

Fallbeispiel 16: Interpretation der Q-Q-Plots

Der Q-Q-Plot (▶ Abbildung 8.14) hat als x-Achse die erwarteten Quantile aus der Normalverteilung, als y-Achse die beobachteten Quantile (die benötigten Schläge).
Als Orientierung ist eine 45°-Gerade eingezeichnet. Bei Punkten oberhalb dieser Geraden sind die beobachteten Quantile größer, bei Punkten unterhalb kleiner als die erwarteten Quantile.

Die Unterschiede zwischen beobachteten und erwarteten Quantilen sind nicht systematisch. Es gibt keine großen Bereiche, wo Punkte nur unter oder nur über der 45°-Gerade liegen.
Über den zweiten Q-Q-Plot (▶ Abbildung 8.15 links) kommen wir zum selben Schluss.

Der dritte Q-Q-Plot (▶ Abbildung 8.15 rechts) unterstellt eine Normalverteilung mit Mittelwert 288 ($= 72 \cdot 4$, durchschnittlich Par 0 in den vier Runden). Der Großteil der Punkte liegt unter der Bezugsgeraden, die beobachteten Quantile sind also fast durchwegs kleiner als die erwarteten.
Es gibt also kaum Grund, an der Normalverteilung der benötigten Schläge für die vier Runden in Augusta 2009 zu zweifeln. Hingegen ist eine Normalverteilung mit Mittelwert 288 nicht passend.

8.4.2 Kolmogorov-Smirnov-Test und Shapiro-Wilk-Test

Mit Q-Q-Plots gewinnen wir grafisch einen Eindruck, ob die Stichprobe einer vermuteten Verteilung folgt. Möglicherweise kann aus dem Plot die Art der Abweichung interpretiert werden. Ausreißer fallen auf Q-Q-Plots ebenso auf wie größeres Gewicht auf den Enden einer Verteilung. Es besteht aber auch der Wunsch, Verteilungsannahmen mit dem Arsenal der Testtheorie zu überprüfen. Zwei Beispiele dafür werden hier vorgestellt.

Kolmogorov-Smirnov-Test

Mit dem Kolmogorov-Smirnov-Test kann die Nullhypothese, dass eine bestimmte Verteilung oder eine bestimmte Verteilungsfamilie vorliegt, überprüft werden. Der Test basiert rechnerisch auf dem Vergleich der empirischen mit der theoretischen Verteilungsfunktion. Nach dem Aufruf von `ecdf()` steht mit `Fn()` eine Funktion zur Berechnung der empirischen Verteilungsfunktion zur Verfügung. Mit der Plotfunktion wird sie im Diagramm als Treppenfunktion eingezeichnet. Zum Vergleich wird an vielen Stellen `xx` der Wert der Verteilungsfunktion der Normalverteilung berechnet (mit `pnorm()`) und diese mit `lines()` in das Diagramm eingezeichnet.

Die Differenzen (`diff`) zwischen den beiden werden berechnet und die Stelle bestimmt (`xmax`), wo diese Differenz am größten ist. Diese Differenz wird im Diagramm (▶ Abbildung 8.16) eingezeichnet (mit `lines()`), die x-Koordinaten sind jeweils `xmax`, die y-Koordinaten werden durch die empirische Verteilungsfunktion `Fn` und die theoretische Verteilungsfunktion `pnorm` gegeben. Zur Markierung wird diese Strecke etwas breiter gezeichnet (`lwd=3`). Die Beschriftung erfolgt mit `text()`.

R

```
> Fn <- ecdf(Gesamt)
> plot(Fn, main = "Kolmogorov-Smirnov-Test")
> xx <- seq(min(Gesamt), max(Gesamt), 0.1)
> mg <- mean(Gesamt)
> sdg <- sd(Gesamt)
> lines(xx, pnorm(xx, mean = mg, sd = sdg))
> diff <- Fn(xx) - pnorm(xx, mean = mg, sd = sdg)
> maxdiff <- which(abs(diff) == max(abs(diff)))
> xmax <- xx[maxdiff]
> lines(c(xmax, xmax), c(Fn(xmax), pnorm(xmax, mean = mg,
+     sd = sdg)), lwd = 3)
> text(xmax, Fn(xmax) - diff[maxdiff]/2, "Kolmogorov-Smirnov-D",
+     pos = 4)
> ks.test(Gesamt, "pnorm", mean = mg, sd = sdg)
```

Wie bei Q-Q-Plots können die Parameter einer Verteilung aus den Daten geschätzt werden, es können aber auch spezielle Parameterwerte vorgegeben werden. In diesem Beispiel wurden mit `mean=mean(Gesamt)` und `sd=sd(Gesamt)` die Werte aus der Stichprobe ermittelt.

Andere Verteilungen können im Aufruf durch den jeweiligen Namen der Verteilungsfunktion (etwa `punif` für die Gleichverteilung) ausgewählt werden.

Shapiro-Wilk-Test auf Normalverteilung

Am häufigsten werden Verteilungstest auf Normalverteilung durchgeführt. Zwar ist der Kolmogorov-Smirnov-Test dafür einsetzbar, mächtiger ist in solchen Situationen allerdings der Shapiro-Wilk-Test. Bei diesem werden die Differenzen zwischen größtem und kleinstem Wert, zwischen zweitgrößtem und zweitkleinsten Wert der Stichprobe, etc. mit Differenzen aus der Normalverteilung verglichen und bewertet. Die Spezifikation einer bestimmten Normalverteilung durch Angabe konkreter Parameterwerte für μ und σ ist nicht möglich.

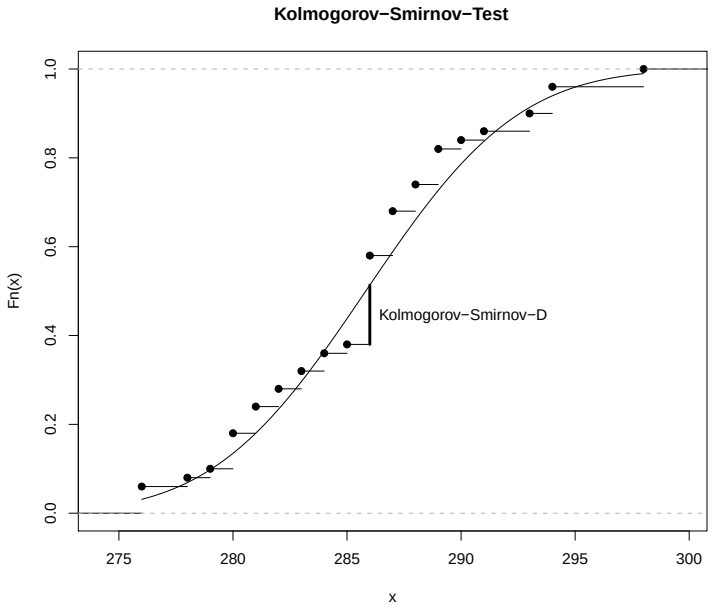


Abbildung 8.16: Kolmogorov-Smirnov-Teststatistik

```
        One-sample Kolmogorov-Smirnov test

data:  Gesamt
D = 0.1336, p-value = 0.3339
alternative hypothesis: two-sided
```

Abbildung 8.17: Kolmogorov-Smirnov-Test der Golfdaten

R

```
> shapiro.test(Gesamt)
```

```
        Shapiro-Wilk normality test

data:  Gesamt
W = 0.9694, p-value = 0.2185
```

Abbildung 8.18: Shapiro-Wilk-Test der Golfdaten

Fallbeispiel 16: Interpretation der beiden Verteilungstests

Für den Kolmogorov-Smirnov-Test (▶ Abbildung 8.17) wurden Stichprobenmittelwert und -varianz als Parameter der Normalverteilung gewählt. Sowohl der Kolmogorov-Smirnov-Test als auch der Shapiro-Wilk-Test (▶ Abbildung 8.18) zeigen mit p-Werten von 0.3339 bzw. 0.2185 keine signifikanten Abweichungen von einer Normalverteilung.

Wie schon der Q-Q-Plot vermuten ließ, gibt es keinen ausreichenden Grund, die Normalverteilung als Verteilung für die benötigten Schläge für beim US-Masters auszuschließen.

8.4.3 Anpassungstest mit der χ^2-Verteilung

Fallbeispiel 17: Überfälle auf Trafiken

Datenfile: `trafik.csv`

Trafiken sind eine speziell österreichische Institution: Kleine Geschäfte, die hauptsächlich als Verkaufsstellen für Tabakwaren, Zeitungen und Magazine dienen und oft auch Schreibwaren, Fahrscheine für öffentliche Verkehrsmittel etc. anbieten.
Sie sind auch immer wieder Ziele von Überfällen, die in der Boulevardpresse je nach Saison Drogenabhängigen, ausländischen Banden (Kriminaltourismus) oder anderen Konzentraten medialen Zorns zugeschrieben werden.
Nun sind solche Überfälle zum Glück seltene Ereignisse und die Häufigkeiten seltener Ereignisse kann oft durch eine Poissonverteilung gut beschrieben werden. Im Datenfile sind für die 53 Kalenderwochen des Jahres 2009 die Anzahl Überfälle auf Trafiken in dieser Woche angegeben.

Fragestellung:
Ist die Anzahl von Überfällen auf Trafiken pro Woche poissonverteilt?

Neben dem Kolmogorov-Smirnov-Test ist der χ^2-Test auf eine vorgegebene Verteilung (▶ Abschnitt 6.3.2) eine weitere Möglichkeit, eine Stichprobe gegen eine bestimmte Verteilung oder Verteilungsform zu testen. Dieser Test ist eigentlich zur Anwendung auf eine kategoriale Variable bestimmt. Um eine metrische Variable zu testen, ist folgendes Schema zu bearbeiten:

1. Klassifizierung der metrischen Variablen in k Klassen.

2. Berechnen der Wahrscheinlichkeiten für die einzelnen Klassen unter der Annahme, dass die zu testende Verteilung vorliegt.
3. Durchführung des χ^2-Tests für vorgegebene Wahrscheinlichkeiten. Werden keine Parameter aus der Stichprobe geschätzt, gilt für die Freiheitsgrade die Beziehung $df = k - 1$. Werden jedoch aus Stichprobe p Parameter geschätzt, muss eine Korrektur bei den Freiheitsgraden erfolgen: $df = k - 1 - p$.

Für das Beispiel und die Bearbeitung in R bedeutet es:

1. Einteilung der Anzahl Überfälle in Klassen. Die Klasseneinteilung muss aufgrund der unterstellten Verteilung, nicht der beobachteten Verteilung, erfolgen. Zu beachten ist dabei, dass die Klassen ausreichend hohe erwartete Häufigkeiten aufweisen. Dies kann üblicherweise nicht ohne Blick auf die Daten erfolgen.

```
> trafik <- read.csv2("trafik.csv", header = TRUE)
> attach(trafik)
> tabueberfall <- table(ueberfall)
> tabueberfall
```

```
ueberfall
 0  1  2  3  4  5  6
 6 10 13 14  4  5  1
```

Vermutlich ist es notwendig, die die beiden letzten Klassen zusammzuleg

```
> wochen <- length(ueberfall)
> tab6 <- c(tabueberfall[1:5], wochen - sum(tabueberfall[1:5]))
> names(tab6)[6] <- "5+"
> tab6
```

```
 0  1  2  3  4 5+
 6 10 13 14  4  6
```

2. Berechnen der Wahrscheinlichkeiten für die einzelnen Klassen (`pois6`) unter der Annahme, dass eine Poissonverteilung vorliegt. Dazu wird der Parameter λ der Poissonverteilung aus den Daten als mittlere Anzahl pro Woche geschätzt (`lambda`) und mit diesem Wert die Wahrscheinlichkeiten für die sechs Klassen berechnet (`pois6`).

```
> total <- sum(ueberfall)
> lambda <- total/wochen
> relHpois <- dpois(0:4, lambda = lambda)
> pois6 <- c(relHpois[1:5], 1 - sum(relHpois))
> cbind(absolut = tab6, relativ = tab6/wochen, poisson = pois6,
+     erwartet = pois6 * wochen)
```

```
   absolut   relativ    poisson  erwartet
0        6 0.1132075 0.09456285  5.011831
1       10 0.1886792 0.22302559 11.820356
2       13 0.2452830 0.26300188 13.939100
3       14 0.2641509 0.20676248 10.958412
4        4 0.0754717 0.12191184  6.461328
5+       6 0.1132075 0.09073535  4.808973
```

Wir sehen, dass nur für die letzte Klasse die erwartete Häufigkeit etwas kleiner als 5 ist, somit der Anpassungstest mit dieser Klasseneinteilung gut durchführbar ist.

3. Durchführung des χ^2-Tests für die Variable mit den Klassenzugehörigkeiten (`tab6`) und den berechneten Wahrscheinlichkeiten `pois6` als Vorgabe. Allerdings müssen wir berücksichtigen, dass wir λ aus den den Daten bestimmt haben. Somit können wir die Funktion `chisq.test()` verwenden, um den X^2-Wert zu berechnen. Der dabei berechnete p-Wert ist aber über eine χ^2-Verteilung mit 5 Freiheitsgraden ermittelt worden, relevant sind aber 4 Freiheitsgrade. Wir berechnen den p-Wert eigenständig unter Verwendung der Verteilungsfunktion der χ^2-Verteilungen `pchisq()`. R

```
> ct <- chisq.test(tab6, p = pois6)
> p.wert <- 1 - pchisq(ct$statistic, df = 4)
> rbind(ct$statistic, p.wert)
          X-squared
          2.6152343
p.wert    0.6241268
```

Fallbeispiel 17: Interpretation χ^2-Tests

Die Tabelle mit den Häufigkeiten zeigt für Wochen mit keinem, einem oder zwei Überfällen keine großen Unterschiede zu den Erwartungen aus einer Poissonverteilung. Die größten Unterschiede sind bei drei oder vier Überfällen pro Woche festzustellen. Wochen mit drei Überfällen sind über-, Wochen mit vier Überfällen unterrepräsentiert.

Dass diese Unterschiede nicht überbewertet werden dürfen, sagt dass Ergebnis des Anpassungstests. Wegen des hohen p-Werts (0.6241) gibt es keinen Grund, an der Nullhypothese, dass die Überfallshäufigkeiten einer Poissonverteilung folgen, zu zweifeln.

Im Package **vcd** bietet die Funktion `goodfit()` die Möglichkeit, die Verteilung von Zählvariablen auf Poisson-, Binomial- oder Negativbinomialverteilung zu testen. Im obigen Beispiel ging es uns aber in erster Linie darum, die prinzipielle Vorgangsweise bei solchen Fragestellungen zu demonstrieren.

8.5 R Befehle im Überblick

`abline(a,b,v,h)` erlaubt das Einzeichnen einer Geraden in einen Plot. Die Gerade kann durch die Angabe der Werte für Konstante `a` und Anstieg `b` oder durch die Angabe des Werts, der eine vertikale `v=` oder horizontale `h=` Gerade bestimmt, definiert werden.

`axis(side)` ermöglicht das Einzeichnen einer Achse in einen Plot. Für `side` kann der Wert 1 (Achse unten), 2 (Achse links), 3 (Achse oben) und 4 (Achse rechts) angegeben werden.

`boxplot(x)` erstellt einen Boxplot für `x`. Neben vielen Optionen kann mit `horizontal = TRUE` der Boxplot liegend und damit platzsparend dargestellt werden.

`cut(x, breaks)` teilt den Bereich von `x` in Intervalle, die durch `breaks` definiert werden und vergibt Namen entsprechend den Intervallen, in die die Beobachtungen fallen. In diesem Abschnitt wurde dieser Befehl zur Umkodierung einer Variablen eingesetzt.

`dnorm(x, mean, sd)` berechnet an den Stellen x den Wert der Dichtefunktion einer Normalverteilung mit Mittelwert `mu` und Standardabweichung `sd`.

`DOTplot(x)` erzeugt ein Punktdiagramm.

`dpois(x, lambda)` berechnet die Wahrscheinlichkeiten, dass eine poissonverteilte Zufallsvariable mit Parameter `lambda` die Werte von `x` annimmt.

`ecdf(x)` erzeugt eine Funktion, über die die empirische Verteilungsfunktion einer Variablen `x` berechnet und geplottet werden kann.

`fivenum(x)` gibt die Fünf-Punkt-Zusammenfassung der Variablen `x` aus.

`hist(x, breaks)` erstellt ein Histogramm für die Werte in der Variablen `x`. Selbst gewünschte Intervallgrenzen können mit `breaks=` angegeben werden.

`IQR(x)` berechnet den Quartilsabstand von `x`.

`ks.test(x, y)` berechnet einen Kolmogorov-Smirnov-Test für einen Datensatz `x` gegen eine Verteilung, deren Verteilungsname in `y` oder deren Verteilungsfunktion in `y` angegeben ist. Parameter der Verteilung können zusätzlich fixiert werden.

`max(x)` berechnet das Maximum von `x`.

`mean(x, trim)` berechnet den Mittelwert von `x`. Ist etwa `trim=0.05` gesetzt, werden 5% der kleinsten und 5% der größten Beobachtungen bei der Berechnung weggelassen.

`median(x)` berechnet den Median von `x`.

`min(x)` berechnet das Minimum von `x`.

`plot(x,y)` für zwei gleich lange Vektoren `x` und `y` werden Punkte mit diesen x- und y-Koordinaten in ein x-y-Diagramm eingezeichnet.

`pchisq(x, df)` berechnet an den Stellen x den Wert der Verteilungsfunktion einer χ^2-Verteilung mit `df` Freiheitsgraden.

`pnorm(x, mean, sd)` berechnet an den Stellen x den Wert der Verteilungsfunktion einer Normalverteilung mit Mittelwert `mu` und Standardabweichung `sd`.

`qnorm(p, mean, sd)` berechnet die Quantile für gewünschte Werte `p` einer Normalverteilung mit Mittelwert `mu` und Standardabweichung `sd`.

`qqline(y)` zeichnet in einen Q-Q-Plot eine Referenzgerade durch die Punkte der ersten und dritten Quartile ein.

`qqnorm(y)` erstellt für `y` einen Q-Q-Normalplot.

`qqplot(x,y)` berechnet Quantile für zwei Datensätze `x` und `y` und plottet sie gegeneinander auf.

`quantile(x, probs)` berechnet Quantile von `x`. Wird für `probs` nichts angegeben, werden alle Quartile (inklusive Minimum und Maximum) berechnet. Wird etwa `probs=seq(0,1,0.1)` spezifiert, werden die Quantile für 0, 10, .., 90 und 100 Prozent berechnet.

`sd(x)` berechnet die Standardabweichung von `x`.

`shapiro.test(x)` berechnet einen Shapiro-Wilk-Test für einen Datensatz `x`.

`sort(x)` sortiert einen Vektor `x` in aufsteigender, mit der Option `decreasing=TRUE` in absteigender Reihenfolge.

`stem(x)` erstellt einen Stem-and-Leaf-Plot. Mit der Option `scale` kann die Länge des Plots kontrolliert werden.

`summary(x)` gibt die Fünf-Punkt-Zusammenfassung und den Mittelwert der Variablen `x` aus.

`t.test(x, mu, alternative)` berechnet einen t-Test. In diesem Kapitel für die Berechnung des Ein-Stichproben-t-Tests eingesetzt. Mit `mu` kann der Wert der Nullhypothese angegeben, mit `alternative` die Alternativhypothese formuliert werden (`"two.sided"`, `"less"`, `"greater"`).

`var(x)` führt zur Berechnung der Varianz von `x`.

`which(x, arr.ind)` gibt jene Indizes des Vektors `x` an, für die der logische Ausdruck `arr.ind` den Wert `TRUE` annimmt.

8.6 Zusammenfassung der Konzepte

Um die Verteilung einer metrischen Variablen grafisch zu beschreiben, werden hauptsächlich Histogramme eingesetzt. Die numerische Beschreibung in Tabellen wird leicht unübersichtlich, eine Zusammenfassung auf einzelne Werte führt zu Lage- und Streuungsmaßen. Im Boxplot sind mehrere solcher Maße enthalten.

Schlüsse auf die Grundgesamtheit führen zu Konfidenzintervallen und Tests für den Mittelwert.

Eine grafische Überprüfung, ob eine bestimmte Verteilung vorliegt, stellen Q-Q-Plots dar. Tests auf allgemeine Verteilungen sind der Kolmogorov-Smirnov-Test und der χ^2-Anpassungstest, der Shapiro-Wilk-Test ist ein Test auf Normalverteilung.

- Verteilung: Angaben dazu, wie stark die Daten in verschiedenen Bereichen vertreten sind.
- Histogramm: grafische Darstellung der Verteilung einer metrischen Variablen
- Boxplot: grafische Darstellung der größenmäßigen Einteilung in Viertel
- Lagemaße: Kennzahlen, die das Zentrum der Daten angeben sollen.
- Streuungsmaße: Kennzahlen, die angeben sollen, wie stark die Daten variieren.
- Ausreißer: Beobachtungen, die sich stark von den meisten anderen unterscheiden.
- Konfidenzintervall für den Mittelwert: Angabe eines Intervalls, in dem der Mittelwert der Grundgesamtheit vermutlich liegt.
- Ein-Stichproben t-Test: Test, ob der Mittelwert in einer Stichprobe sehr stark von einem vorgegebenen Wert abweicht.
- Q-Q-Plot: grafische Überprüfung, ob eine Stichprobe einer bestimmten Verteilung folgt
- Kolmogorov-Smirnov-Test: Test auf eine bestimmte Verteilung
- Shapiro-Wilk-Test: Test auf eine Normalverteilung

8.7 Übungen

1. **Alter bei Amtsantritt der US-Präsidenten**

   Im Datenfile `us-president.csv` ist das Alter der US-Präsidenten bei Amtsantritt angegeben.

   - Beschreiben Sie den Datensatz mit Histogramm, Boxplot und Maßzahlen!

2. **VwGH-Daten: Verfahrensdauer in der ersten Instanz**

   Gegen Abgabenbescheide von Behörden kann Berufung eingelegt werden. In Österreich ist die Berufungsbehörde 2. Instanz der Verwaltungsgerichtshof (VwGH). In einer Studie wurden alle Entscheidungen des VwGH zwischen 2000 und 2004 in Abgabensachen untersucht. Ein Untersuchungsgegenstand waren die Verfahrensdauern.
   Im Datenfile `vwgh.csv` ist auch die Länge des Verfahrens in der zweiten Berufungsinstanz angegeben (`dauer3`).

   - Dauern die Verfahren in der 2. Instanz länger als vor 20 Jahren, als sie im Schnitt 1 Jahr und 3 Monate dauerten?

- Ist das Ergebnis nur deshalb signifikant, weil nicht wenige Ausreißerwerte vorliegen?

3. **US-Masters 2009: Runden 1 und 2**

- Beschreiben Sie die Verteilung der für zwei Runden im US-Masters in Augusta 2009 insgesamt benötigten Schläge (Variablen `R1` und `R2` im Datenfile `augusta2009.csv` enthalten die zwei ersten Rundenergebnisse).
- In welchem Intervall würde man aufgrund dieser Stichprobe die notwendigen Schläge für die zwei ersten Runden annehmen?
- Ist die Anzahl Schläge für die ersten zwei Runden normalverteilt? Welche Ergebnisse zeigen der Shapiro-Wilk-Test und der Kolmogorov-Smirnov-Test? Weist der Q-Q-Plot Auffälligkeiten auf?

Die Datendateien sowie ausgewählte Lösungen finden Sie auf der Webseite des Verlags.

CWS