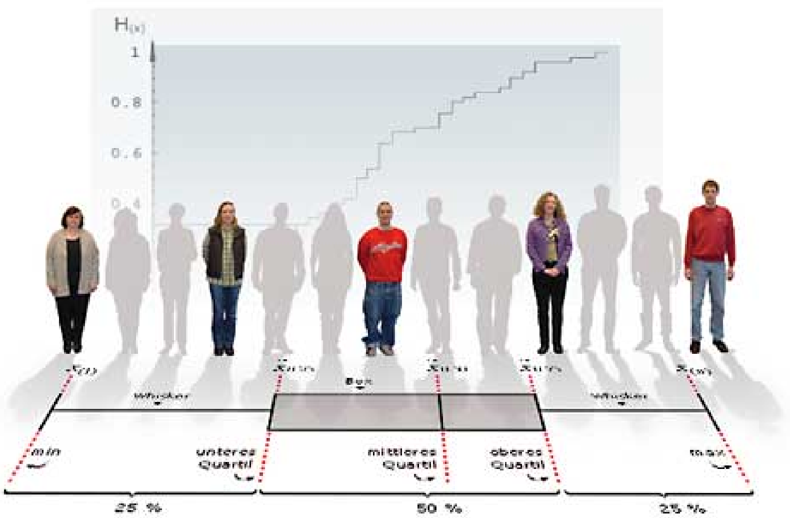
QBX - Quantile und Boxplot



Überblick und Lernziele

In dieser Lerneinheit beginnen wir mit der Betrachtung von Lagemaßen. Lagemaße sind wichtige Kennzahlen zur Charakterisierung univariater Daten.

Unter anderem werden Sie den Katzenschnurrbart als universelles Hilfsmittel der Statistik schätzen lernen.

Lernziele

In dieser Lerneinheit betrachten wir die Quantile und die daraus abgeleiteten Boxplots.

Nach dem Durcharbeiten dieser Lerneinheit sollen Sie

- Quantile bestimmen und interpretieren können,
- Quantile im Schaubild der empirischen Verteilung ablesen können,
- Quantile für klassierte Daten bestimmen und interpretieren können,
- die Fünf-Punkte-Zusammenfassung ermitteln können,
- den Boxplot bestimmen und zeichnen können.

Gliederung

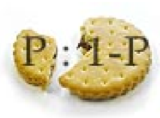
Zeitbedarf und Umfang

Für die Durcharbeitung dieser Lerneinheit benötigen Sie ca. 120 Minuten und für die Übungen mit der Statistiksoftware **R** ca. 120 Minuten.

1 Einführung

Die Bestimmung von Quantilen (*lat.*: quantus - wieviel, wie groß) spielt in der Statistik eine große Rolle. In Lerneinheit GST „Grundbegriffe der Statistik" haben Sie bereits den Median, das 50 % Quantil kennengelernt.
Jetzt wollen wir uns systematisch mit Quantilen beschäftigen.

Spezielle Quantile x_p teilen eine Verteilung in Abschnitte gleicher Häufigkeit ein, z. B.:



- 3 Quartile teilen die geordneten Werte in 4 Abschnitte mit den relativen Häufigkeiten $h_1 = h_2 = h_3 = h_4 = 0{,}25$
  (das 2. Quartil ist gerade der Median);
- 4 Quintile in 5 Abschnitte mit $h_1 = ... = h_5 = 0{,}2$;
- 9 Dezile in 10 Abschnitte mit $h_1 = ... = h_{10} = 0{,}1$;
- 99 Perzentile in 100 Abschnitte mit $h_1 = ... = h_{100} = 0{,}01$
  etc.

Der Median (gleich 2.Quartil) teilt die Daten in der Mitte.

Benennung der Quantile	Anzahl der Intervalle	Anteil
Terzile	3	1/3
Quartile	4	1/4
Quintile	5	1/5
Dezile	10	1/10
Vigintile	20	1/20
Perzentile (Zentile)	100	1/100

Tab.: Benennung der Quantile und ihre Intervalle

Anmerkung

Die Bezeichnungen Terzil, Quartil, Quintil usw. sind wegen der Aufteilung in drei, vier, fünf, etc. Abschnitte eigentlich selbsterklärend. Der Anteil jedes Abschnitts beträgt entsprechend 1/3, 1/4, 1/5 usw. Jedes Quantil bezeichnet man als x_p mit dem Anteil p, der links von ihm liegt, z. B. $x_{1/3}$ und $x_{2/3}$ für die beiden Terzile, $x_{0,25}$, $x_{0,5}$ und $x_{0,75}$ für die drei Quartile usw.

2 Das Quantil (Definition)

Wie man empirische Quantile aus Daten bestimmt, legt die folgende Definition fest:

Definition

Quantile

Das p-Quantil x_p teilt eine Stichprobe $x_1 \ldots, x_n$ so ein, dass mindestens ein Anteil p der Werte kleiner oder gleich x_p ist und gleichzeitig mindestens ein Anteil $1-p$ der Werte größer oder gleich x_p ist.

Synonyme: p-Fraktil (*lat.*: fractio - Bruch), empirisches (oder deskriptives) p-Quantil.

Anmerkungen

Die p-Quantile x_p teilen die Daten in zwei Gruppen auf. Unterhalb ($<$) von x_p befinden sich höchstens 100 · p Prozent der Merkmalswerte und oberhalb ($>$) von x_p befinden sich höchstens 100 · (1-p) Prozent der Merkmalswerte.

Die Elemente einer geordneten Datenliste bezeichnen wir mit $x_{(i)}$.
Es gilt: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ für eine Liste der Länge n.

Das p-Quantil x_p mit $(0 < p < 1)$ wird folgendermaßen berechnet:

$$x_p = \begin{cases} x_{(\lfloor np+1 \rfloor)}, & \text{np nicht ganzzahlig} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & \text{np ganzzahlig} \end{cases}$$

$\lfloor np+1 \rfloor$ bedeutet, dass auf die nächste ganze Zahl abgerundet wird (Gaußklammer).

Exkurs

Exkurs: Gaußklammer (Siehe Anhang)

Eigentlich würden wir gerne die Forderung $F_n(x_p) = p$ erfüllen.
Da F_n eine Treppenfunktion ist, wird die folgende Fallunterscheidung erforderlich:

- Wenn $n \cdot p$ nicht ganzzahlig ist, dann sind Quantile realisierte Merkmalswerte.
- Ist $n \cdot p$ ganzzahlig, dann wird zur Interpolation die Intervallmitte der beiden benachbarten Werte $x_{(np)}$ und $x_{(np+1)}$ verwendet.

Es gibt auch andere Berechnungsvorschläge. In diesem Kurs verwenden wir den angegebenen Vorschlag. Praktisch sind die Unterschiede nicht relevant.

Klassierte Daten

Die gesamte Fläche des Histogramms hat den Wert Eins. Für klassierte Daten wird x_p durch Interpolation so bestimmt, dass die Histogrammfläche links von x_p den Wert p hat.

Ist k die Klasse, in die das p-Quantil x_p fällt, x_{k-1} die untere Grenze dieser Klasse und b_k die Klassenbreite, dann ist x_p näherungsweise gegeben durch:

$$x_p = x_{k-1} + \frac{p - F_{k-1}}{h_k} \cdot b_k$$

Dabei ist k so zu wählen, dass $F_{k-1} \leq p \leq F_k$. Durch diese Vorschrift wird x_p gerade so gewählt, dass die Histogrammfläche links von x_p den Wert p hat.

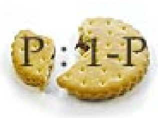
Natürlich ist die Bestimmung von Quantilen aus den Originaldaten immer vorzuziehen. Die Formel für klassierte Daten wird nur verwendet, wenn die Originaldaten nicht zugänglich sind.

2.1 Spezielle Quantile

Die Bedeutung eines p-Quantils für jedes p zwischen 0 und 1 ist klar und eindeutig definiert. In der Praxis haben sich in verschiedenen Bereichen unterschiedliche Vorlieben für spezielle Werte von p entwickelt.

Anmerkung

Die folgende Tabelle gibt Ihnen eine Übersicht. **(Bitte nicht auswendig lernen!)**



Die am meisten verwendeten Quantile sind die Perzentile, Quartile und natürlich der Median. Deren genaue Bedeutung sollte Ihnen jederzeit präsent sein.

Quantile und ihre Bennenung

P	Benennung d. Quantils	alternative Benennungen			
		x. Vigintil	**x. Dezil**	**x. Quintil**	**x. Quartil**
0,01	1. Perzentil				
0,02	2. Perzentil				
0,05	5. Perzentil	1. Vigintil			
0,10	10. Perzentil	2. Vigintil	1. Dezil, unteres Dezil		
0,20	20. Perzentil	4. Vigintil	2. Dezil	1. Quintil, unteres Quintil	
0,25	25. Perzentil	5. Vigintil			1. Quartil, unteres Quartil
0,33	1. Terzil				
0,50	50. Perzentil	10. Vigintil	5. Dezil	**Median**	2. Quartil, mittleres Quartil
0,66	2. Terzil				
0,75	75. Perzentil	15. Vigintil			3. Quartil, oberes Quartil
0,80	80. Perzentil	16. Vigintil	8. Dezil	4. Quintil, oberes Quintil	
0,90	90. Perzentil	18. Vigintil	9. Dezil, oberes Dezil		
0,99	99. Perzentil				

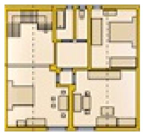
Tab.: Quantile und ihre Bennenung

2.2 Berechnung von Quantilen

Die Quantile der Mieten können sehr hilfreich sein, um herauszufinden welche Chancen wir hätten, eine bezahlbare Wohnung zu finden. Im folgenden Beispiel sehen Sie, wie Quantile bestimmt werden können.

Beispiel

Quantile aus den Daten



Der geordnete Datensatz der Mieten (in €) von 99 freien Einzimmerwohnungen in Berlin sieht wie folgt aus:

Geordnete Daten: Angebote der Einzimmerwohnungen mit der Angabe der Mieten

231 247 251 254 256 256 263 277 288 **301** 302 306 322 333 333 333 340 348 352 354 359 363 364 368 368 369 373 373 374 375 378 380 382 384 384 384 384 388 388 395 395 395 395 405 405 409 413 417 420 423 431 432 433 444 446 448 454 461 473 474 475 477 481 500 500 501 506 506 507 511 512 519 535 538 546 546 552 553 564 581 589 609 609 628 666 681 691 699 727 742 832 916 947 1008 1151 1175 1190 1253 1429

Wir wollen das 0,1-Quantil (Dezil) der Mieten von 99 Einzimmerwohnungen in Berlin bestimmen.

Da $n \cdot p = 99 \cdot 0{,}1 = 9{,}9$ nicht ganzzahlig ist, erhalten wir $x_{(\lfloor np+1 \rfloor)} = x_{(10)}$ als 0,1-Quantil. Bei unseren Mieten gilt $x_{(10)} = 301$.

Interpretation

90 % der angebotenen Einzimmerwohnungen kosten mehr als 301 € Miete, nur 10 % der Wohnungen sind für höchstens 301 € / Monat zu haben.

Analog lassen sich andere p-Quantile berechnen und interpretieren.

2.3 Übungen zu Quartilen (1/2)

Wir verwenden noch einmal die Mietpreise von Einzimmerwohnungen.

Geordnete Daten

Angebote der Einzimmerwohnungen mit der Angabe der Mieten

231 247 251 254 256 256 263 277 288 301 302 306 322 333 333 333 340 348 352 354 359 363 364 368 368 369 373 373 374 375 378 380 382 384 384 384 384 388 388 395 395 395 395 405 405 409 413 417 420 **423** 431 432 433 444 446 448 454 461 473 474 475 477 481 500 500 501 506 506 507 511 512 519 535 538 546 546 552 553 564 581 589 609 609 628 666 681 691 699 727 742 832 916 947 1008 1151 1175 1190 1253 1429

„423“ ist der 50. Wert.

Berechnen

Übung QBX-01

Mietpreise und Quantilberechnungen

Bestimmen Sie die Quartile Q_1 (unteres Quartil) und Q_3 (oberes Quartil) sowie die Quartilsmitte.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Berechnen

Übung QBX-02

Median und Quartilsmitte

Unterscheiden sich Median und Quartilsmitte?

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Berechnen

Übung QBX-03

Mietpreise und Quantilberechnungen (noch einmal)

Verfahren Sie entsprechend, wenn sich der Datensatz nur auf (die ersten) 50 Mieten erstreckt.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

2.4 Übungen zu Quartilen (2/2)

Die Inhaberin des Weinfachgeschäftes Maestro möchte sich einen eigenen Weinberg im Veneto kaufen. Auf eine Anzeige in einschlägigen Zeitschriften bekam sie 10 Grundstücksangebote. Die Grundstücke haben folgende Preise umgerechnet in Tausend Euro (Merkmal X):

10,5	5	4,3	12	14	10,5	6,2	5	7,3	10,5

Formulieren

Übung QBX-04

Grundstückspreise: Verteilung

Geben Sie die empirische Verteilungsfunktion grafisch an.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Formulieren

Übung QBX-05

Grundstückspreise: Verteilung 2

Bestimmen und interpretieren Sie $F_n(13)$.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Zeichnen / Entwerfen

Übung QBX-06

Grundstückspreise: Median

Bestimmen Sie den Median der Angebotspreise anhand der empirischen Verteilung (Funktionsgraf). Interpretieren Sie das Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Zeichnen / Entwerfen

Übung QBX-07

Grundstückspreise: Quantil

Bestimmen Sie grafisch das 20 %-Quantil und interpretieren Sie das Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

2.5 Quantile und Häufigkeiten

Wie Sie schon gesehen haben, erhalten wir aus den kumulierten relativen Häufigkeiten den Anteil der Werte, die kleiner oder gleich einer vorgegebenen Grenze sind.

Umgekehrt gefragt: Bei welchem x erreicht die kumulierte Häufigkeit einen vorgegebenen Anteil p? Diese x-Werte sind die empirischen Quantile. Am folgenden Beispiel können Sie sich den Zusammenhang nochmals verdeutlichen.

Beispiel

Körpergrößen von Studierenden der Statistik

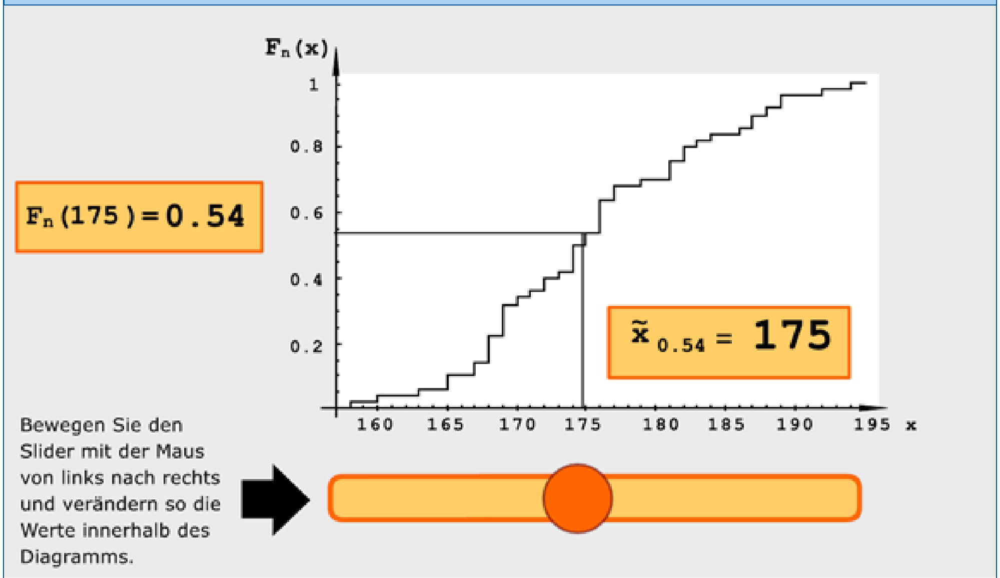
Hier greifen wir auf einen Datensatz zurück, der aus einer von Studierenden der Statistik durchgeführten Erhebung der Körpergrößen von Studierenden an der Berliner Hochschule für Technik im WS 2021 stammt:

194 179 169 175 181 188 187 171 176
168 176 175 168 169 174 186 169 177
177 170 172 172 174 163 181 187 189
189 176 165 173 176 181 182 165 174
168 158 192 176 167 169 160 182 168

Den Zusammenhang zwischen Verteilung und Quantilen sollen die folgenden Informationen veranschaulichen:

Interaktion

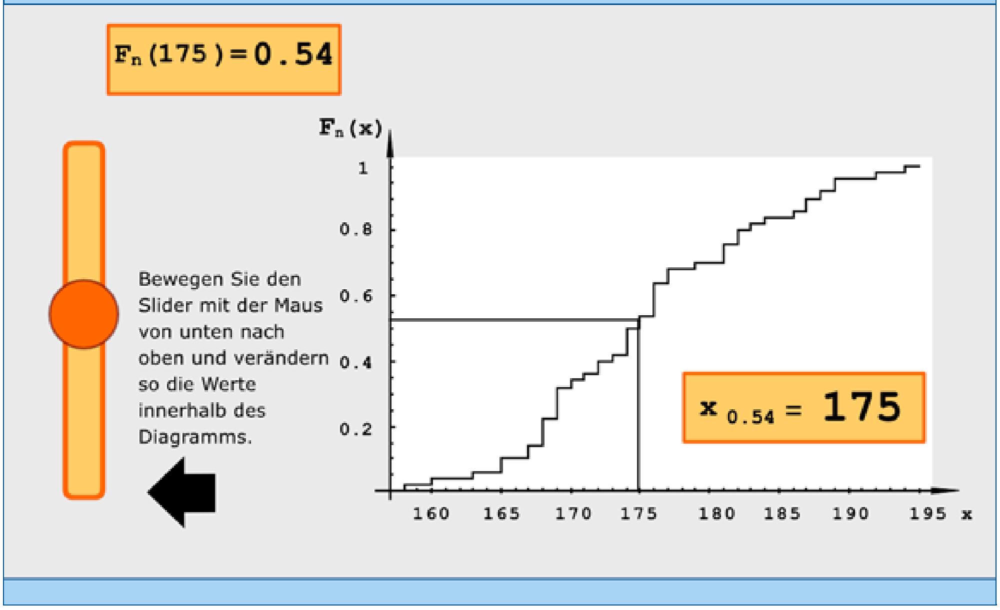
Empirische Verteilung: $F_n(x)$ ablesen



In der Animation „Empirische Verteilung: $F_n(x)$ ablesen“ kann im Intervall von $[157, 194]$ mit der Scrollleiste jede ganze Zahl eingestellt werden. Zu der jeweiligen Zahl kann der Wert der Verteilungsfunktion $F_n(x_i) = h(X \leq x_i)$ abgelesen werden.

Interaktion

Empirische Verteilung: Quantile ablesen



In der Animation „Empirische Verteilung: Quantile ablesen“ kann der Wert der Verteilungsfunktion eingestellt werden. Zu jedem Wert $F_n(x_i)$ kann das entsprechende Quantil abgelesen werden.

In der Animation wird der Zusammenhang zwischen den kumulierten relativen Häufigkeiten, den Quantilen und ihren Werten anschaulich dargestellt werden.

2.6 Quartile für klassierte Daten (Beispiel)

Wie berechnen wir Quantile, wenn wir nicht auf die Originaldaten zugreifen können? Die Berechnungsvorschrift hatten wir schon zu Beginn des Kapitels erläutert.

Wir wollen jetzt das obere Quartil Q_3 bestimmen. Hierzu greifen wir auf die bereits aus der Lerneinheit DHV bekannten Daten zurück, die wir Ihnen hier noch einmal darstellen:

Häufigkeitstabelle: Klassierte Mieten

Index k	**Klasse** K_k $x_{k-1} < X \leq x_k$	**absolute Klassenhäufigkeit** $n_k = n(x_{k-1} < X \leq x_k)$	**kumulierte relative Klassenhäufigkeit** F_k
1	K_1 = (0; 200]	0	0
2	K_2 = (200; 400]	43	0,43
3	K_3 = (400; 600]	38	0,82
4	K_4 = (600; 800]	9	0,91
5	K_5 = (800; 1000]	3	0,94
6	K_6 = (1000; 1200]	4	0,98
7	K_7 = (1200; 1400]	1	0,99
8	K_8 = (1400; 1600]	1	1

Tab.: Häufigkeitstabelle: Klassierte Mieten

Achtung

Berechnungen auf der Grundlage von Originaldaten sind grundsätzlich denen auf der Basis klassierter Werte vorzuziehen.

Die Anwendung von Methoden für klassierte Daten hilft uns, wenn wir keine andere Berechnungsgrundlage haben.

Beispiel

Quantile klassierter Daten

$p = 0{,}75.$

Wegen $F_2 = 0{,}43 \leq p \leq F_3 = 0{,}82$ ergibt sich, dass Q_3 in der Klasse K_3 liegt. Untere Grenze: 400, obere Grenze 600.

Wir lesen die benötigten Werte aus der Tabelle ab und erhalten als

Oberes Quartil (p = 0,75, k = 3):

$$x_p = x_{k-1} + \frac{p - F_{k-1}}{h_k} \cdot b_k$$

$$\approx 400 + \frac{0{,}75 - 0{,}43}{0{,}38} \cdot 200 \approx 568$$

d. h. 75 % der Mieten der 99 freien Einzimmerwohnungen in Berlin sind gleich oder niedriger als 568 € bzw. ein Viertel der Mieten der Einzimmerwohnungen ist gleich oder höher als 568 €.

2.7 Übung zu Quartilen

Der Inhaber des Weinfachgeschäftes Maestro hat eine neue Proseccosorte in sein Sortiment aufgenommen und interessiert sich für die Anzahl der pro Tag verkauften Flaschen dieser Proseccosorte. Vier Wochen lang hat er täglich die Anzahl der verkauften Flaschen notiert und in der folgenden Tabelle zusammengefasst.

Tab.: Beispiel für verkaufte Flaschen

Verkaufte Flaschen	Anzahl der Tage
0	3
2	5
3	7
5	4
7	3
8	2

Berechnen

Übung QBX-08

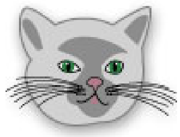
Verkaufszahlen: Quartile

Geben Sie die Quartile der Verteilung an und stellen Sie die Ergebnisse grafisch dar. Interpretieren Sie bitte kurz das Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

3 Boxplot



Was macht die Katze mit dem Bart? Wenn eine Katze durch eine enge Öffnung kriecht, muss sie keine Angst haben, stecken zu bleiben. Woran liegt das? Die Schnurrbarthaare am Katzenkopf sind empfindlich gegenüber Berührung und zeigen ihr, wieviel Raum ihr Körper beansprucht, um durch eine Öffnung zu gelangen.

Eine ähnliche Datenanzeige ist auch in der Statistik hilfreich.

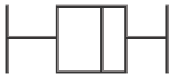
Ergänzen wir die Quartile und den Median noch um den kleinsten Wert $x_{min} = x_{(1)}$ und den größten Wert $x_{max} = x_{(n)}$, wissen wir genauso viel wie die Katze.

Die Quartile (Q_1, Q_3), der Median ($\tilde{x}_{0,50}$) sowie das Minimum (x_{min}) und Maximum (x_{max}) teilen den Datensatz in vier Teile. Jeder dieser Teile enthält in etwa ein Viertel der Beobachtungswerte.

Die Angabe dieser fünf Werte wird auch als 5-Punkte-Zusammenfassung bezeichnet. Sie enthält wesentliche Informationen über die Verteilung der Beobachtungen.

Die 5-Punkte-Zusammenfassung besteht aus:

Minimum, Q_1, Median, Q_3, Maximum



Die grafische Darstellung der 5-Punkte-Zusammenfassung heißt box and whisker plot (engl.: box - Schachtel, whisker - Schnurr- oder Barthaar, plot - Plan). Die Benennung geht auf den amerikanischen Statistiker John W. Tukey zurück.

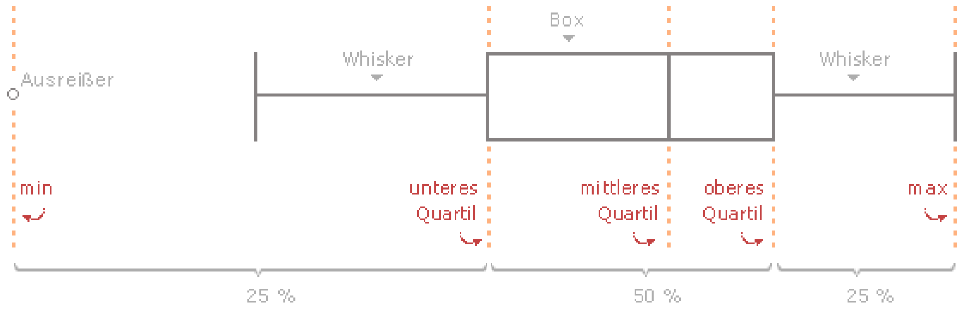
Abb.: John W. Tukey

Man erhält mit dem Boxplot eine komprimierte grafische Darstellung der Daten, die sehr gut zum Vergleich verschiedener Datensätze geeignet ist. Unterschiede in der Varianz, bezüglich des Medians oder bzgl. Minimum und Maximum lassen sich leicht erkennen. Darüber hinaus gewinnt man schnell einen Eindruck darüber, ob die Beobachtungen in einem Datensatz annähernd symmetrisch verteilt sind oder ob es sich um eine schiefe Verteilung handelt. Auftretende „verdächtige“ Ausreißer sind sofort ersichtlich.

Achtung

Ausreißer/Outlier

Als Ausreißer werden extreme Werte betrachtet, die weit von den anderen Daten entfernt liegen. Ihre Ursache kann die zufällige Variabilität der Daten sein. Oft sind sie aber auf Messfehler, Fehler bei der Datenübertragung oder besondere Ereignisse wie z. B. Ausfälle im System, Wetterphänomene, Feiertage etc. zurück zu führen. Um dann die Ergebnisse der Datenanalyse nicht zu verfälschen, schließt man diese Ausreißer von der weiteren Analyse aus. Was als Ausreißer gilt, ist eine Frage der Festlegung.



Es hat sich bewährt, Datenpunkte als Ausreißer zu betrachten, die mehr als die 1,5 fache Breite der Box von der Box entfernt liegen; daher definiert man die sogenannten Zäune als unteres Quartil minus 1,5 Boxbreiten und oberes Quartil plus 1,5 Boxbreiten. Im professionellen Boxplot werden die Whisker dann nicht bis zum Minimum und Maximum gezeichnet, sondern nur bis zu den extremsten Datenwerten, die innerhalb dieser Zäune liegen. Ausreißer werden dann als einzelne Punkte jenseits der Whisker dargestellt.

Wenn Sie in R die Funktion `boxplot` nutzen, ist die oben beschriebene Festlegung für Whisker und Ausreißer voreingestellt (der Faktor 1.5 kann mit Argument `range` angepasst werden). Die Ausreißer erscheinen als Punkte im Plot.

In der weiterführenden Statistik können Sie alternative Methoden zur Ausreißerbestimmung kennenlernen.

3.1 Boxplot (Beispiele)

Die 5-Punkte-Zusammenfassung und ihre grafische Darstellung als Boxplot sind so wichtig für die Anwendung, dass Sie diese in aller Tiefe verstanden haben sollten. Man kann diese Darstellung quasi universell einsetzen, vorausgesetzt die Daten lassen sich der Größe nach ordnen (mindestens ordinalskaliert). Betrachten wir die folgenden zwei Beispiele.

Beispiel Körpergröße

Am Beispiel einiger Mitglieder des Berliner VFH-Projektteams werden wir die 5-Punkte-Zusammenfassung demonstrieren. Es geht um die Körpergröße. Am Ende können Sie die jeweiligen Größen der Teammitglieder ausmessen.

1. Schritt:
Betrachten Sie zunächst die unterschiedlichen Körpergrößen der dargestellten Personen des VFH-Teams.

Abb.: VFH-TEAM ungeordnet

2. Schritt:
Wir sortieren die hier abgebildeten Personen der Größe nach.

Abb.: VFH-TEAM geordnet

3. Schritt:
Überlegen Sie bitte, welche Plätze in der nach Größe geordneten Reihenfolge den fünf charakteristischen Werten: Min, Q_1, Q_2 (Median), Q_3,und Max entsprechen.

Klar ist, dass bei n = 13 Werten (Personen) der 1. Wert = Min und der 13. Wert = Max ist.

Das Minimum liegt an erster Stelle, für Q_1(unteres Quartil) ergibt sich der 4. Wert, für den Median der 7. Wert, das obere Quartil Q_3 entspricht dem 10. Wert und das Maximum liegt am 13. Platz.

Hinweis

Verwechseln Sie bitte nicht Wert und Platz!

Die Quantile entsprechen dem Wert und nicht dem Platz.

Abb.: VFH-TEAM hervorgehoben

In der Abbildung sind die charakteristischen Plätze hervorgehoben.

Rolloverbild

Lesen Sie bitte die Werte für die 5-Punkte-Zusammenfassung am Rolloverbild ab.

Geordnete Stichprobe

$$X_{(1)} \le X_{(2)} \le X_{(3)} \le X_{(4)} \le X_{(5)} \le X_{(6)} \le X_{(7)} \le X_{(8)} \le X_{(9)} \le X_{(10)} \le X_{(11)} \le X_{(12)} \le X_{(13)}$$

Warum gibt es hier keine Ausreißer? Die Boxbreite ist 9, Zäune sind bei 165 und 194 und es gibt keinen Datenwert außerhalb der Zäune.

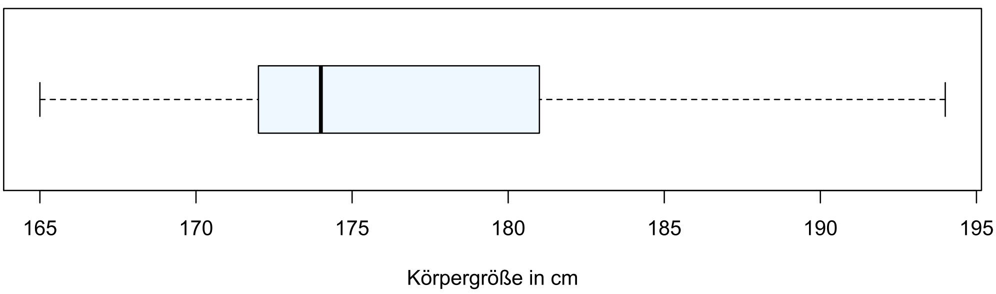


Abb.: Boxplot Körpergröße des VFH-Teams

Wäre die größte Person 195 groß, würde der rechte Whisker bis 191 reichen und es gäbe einen Ausreißer bei 195. Im R-Code haben wir den Wert 194 mal schnell geändert und erhalten den unteren Boxplot.

VFH-TeamBoxplot.R (1 KB)

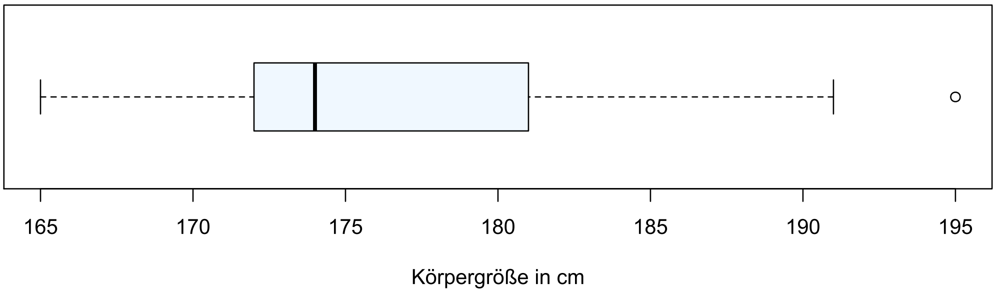


Abb.: Boxplot Körpergröße des VFH-Teams mit Ausreißer

Alles klar?! Gut dann kommen wir zum zweiten Beispiel.

Beispiel

Beispiel Leistungsbilanz

Die Zahlungsbilanz fasst systematisch Transaktionen zwischen Gebietsansässigen und Gebietsfremden verschiedener Staaten zusammen. Die Leistungsbilanz ist ein wichtiger Teil der Zahlungsbilanz und zeigt die Differenz zwischen der Summe der Exporte und Einkommen und der Summe der Einfuhren und Verbindlichkeiten. Dabei beziehen sich die Ausfuhren und Einfuhren sowohl auf Güter als auch auf Dienstleistungen, während das Einkommen sich sowohl auf Primär- als auch auf Sekundäreinkommen bezieht. Der Wert des Leistungsbilanzsaldos entspricht der Differenz zwischen Spareinlagen und Investitionen für die Wirtschaft.

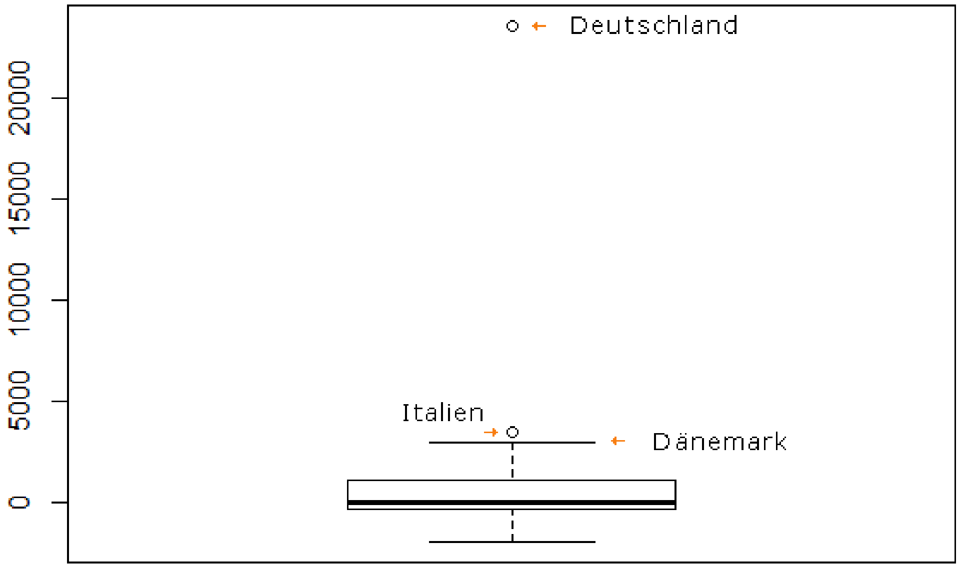
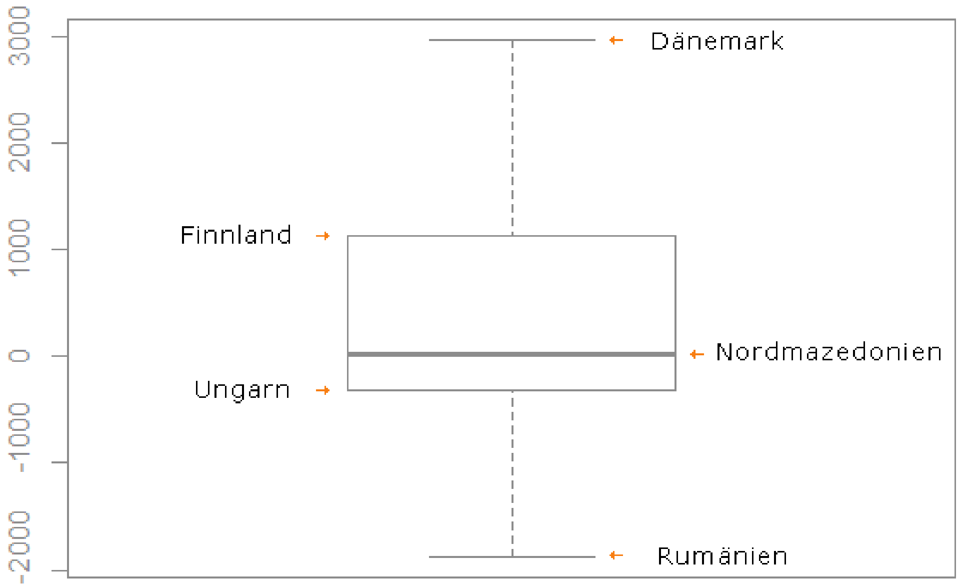


Abb.: Leistungsbilanz mit Ausreißern

Dargestellt wird die Leistungsbilanz der EU Staaten mit den Ausreißer Deutschland und Italien (Stand 09.2021) Die Angaben in der Grafik verstehen sich in Millionen Euro. Die Daten stammen von EUROSTAT (https://ec.europa.eu/), dem statistischen Amt der Europäischen Union. Der Boxplot erfasst die Leistungsbilanz der 27 EU Länder als auch der offiziellen Beitrittskandidaten Nordmazedonien und Serbien. Wobei Irland, Zypern, die Niederlande, Österreich und das ehemalige Mitglied Vereinigtes Königreich keine Angaben gemacht haben.

Deutschland und Italien werden im oberen Boxplot als Ausreißer betrachtet da sie mehr als das 1,5 fache der Breite der Box vom oberen Rand der Box bzw. vom oberen Quartil entfernt liegen. Deutschlands Leistungsbilanz ist fast achtmal so groß wie die von Dänemark. Dänemark hat abgesehen von den Ausreißern die höchste Leistungsbilanz der EU Länder, die Angaben gemacht haben und stellt daher im Boxplot Ende des oberen Whiskers dar. Da Deutschland so weit vom Rest des Datensatzes weg liegt, muss die y-Achse entsprechend weit reichen und die Darstellung wird unleserlich.

Abb.: Leistungsbilanz ohne Ausreißer



Leistungsbilanz der EU Staaten ohne die Ausreißer Deutschland und Italien. Quelle: eurostat Dänemark hat die höchste Leistungsbilanz (2960,8 Mio), Rumänien die niedrigste (-1872 Mio). Nordmazedonien stellt genau den Median dar(15,8 Mio). Von den zentralen 50 % der Länder hat Ungarn die niedrigste Leistungsbilanz (unteres Quartil, -310,7 Mio) und Finnland die höchste (oberes Quartil, 1129 Mio). Die Leistungsbilanz aller anderen Staaten beweget sich zwischen -1872 Millionen und +2960,8 Millionen.

.

Ausreißer können verzerren, sowohl Grafiken, wie Sie sehr schön an Hand der Boxplots oben sehen können, aber wichtiger noch – auch statistische Kennwerte und Analysen.

Betrachten wir nun die Mittelwerte der Leistungsbilanzen, einmal mit den Ausreißern Italien und Deutschland, $\overline{x} = 1.212$ Millionen und einmal ohne die Ausreißer $\overline{x} = 144{,}52$ Millionen wird klar, dass der Mittelwert durch Ausreißer deutlich beeinflusst und ggf. verzerrt wird. Der Median hingegen ist stabil in Bezug auf Ausreißer.

Wie ein Berliner Professor sagte: *„Ausreißer können Datenmüll sein oder man gewinnt den Nobelpreis damit“*. In der Regel liegt die Wahrheit irgendwo dazwischen. In jedem Fall sollten Sie Ausreißer nie unbeachtet lassen.

Die Daten und den R-Code zum Erzeugen der Boxplots können Sie sich hier herunterladen.

Leistungsbilanz.csv (3 KB)
Leistungsbilanz.R (1 KB)

Beispiel

Beispiel Frauenquote

Den Datensatz in dem es um den Frauenanteil in Aufsichtsräten privatwirtschaftlicher Unternehmen geht, kennen Sie schon aus der Lerneinheit *„DHV - Daten, Häufigkeiten und Verteilung“*. Seit 2016 gilt in Deutschland für Unternehmen die börsennotiert und mitbestimmt sind eine feste Frauenquote von mindestens 30 % für Neubesetzungen im Aufsichtsrat. Hat ein Unternehmen mit verbindlicher Frauenquote weniger als 30 % Frauen in seinem Aufsichtsrat, muss es frei werdende Posten so lange an Frauen vergeben, bis die 30-Prozent-Marke erreicht ist. In diesem Beispiel wollen wir nun die Boxplots der Unternehmen mit (81 Firmen) und ohne (946 Firmen) verbindliche Frauenquote vergleichen.

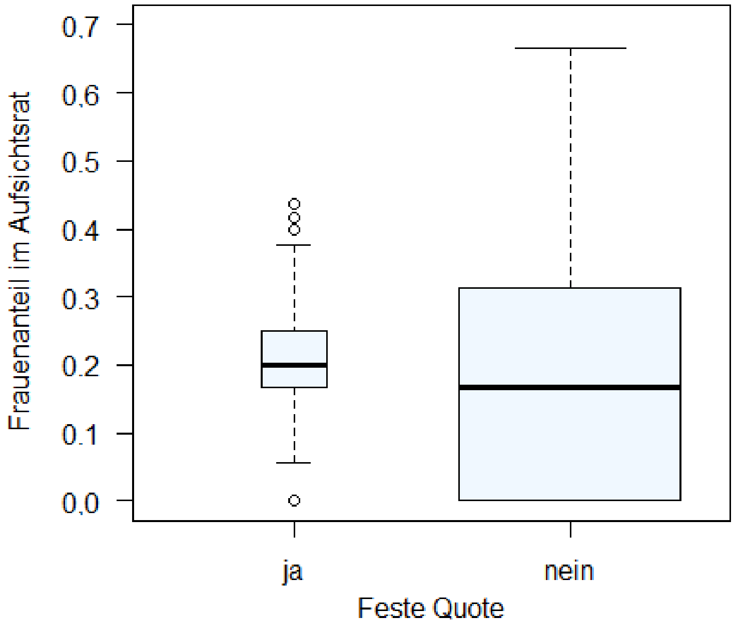


Abb.: Unternehmen mit und ohne verbindlicher Frauenquote

Umfang Ja (n=81) nein (n=946)

Was bei der Betrachtung sofort auffällt, ist dass die Unternehmen ohne verbindliche Quote mehr streuen, die Werte liegen zwischen 0 und ca. 0,7 während bei den Unternehmen mit verbindlicher Quote nur Werte zwischen 0 und etwas über 0,4 vorkommen. Das dürfte dem sehr unterschiedlichen Datenumfang geschuldet sein. Während wir nur 81 Unternehmen mit fester Quote haben, sind es bei denen ohne mehr als zehnmal so viele, nämlich 946. Der Median liegt bei den Unternehmen mit fester Quote tatsächlich etwas höher als bei denen ohne, etwa bei 0,2.

Da die Regelung erst seit 2016 gilt, Stand heute also sechs Jahre, wird es wohl noch ein wenig dauern bis die 30 % erreicht sind. Deutlich überlegen sind die Firmen mit fester Quote wenn es um Aufsichtsräte geht, die komplett männlich besetzt sind. Es gibt hier nur eine Firma ohne eine Frau im Aufsichtsrat, die zudem als Ausreißer gekennzeichnet wird.

Bei den Unternehmen ohne feste Quote liegen Firmen mit Aufsichtsräten komplett ohne Frauen auf dem unteren Quartil, d. h. mindestens 25 % der Unternehmen ohne feste Quote haben keine einzige Frau im Aufsichtsrat.

Interessant ist sicher auch wer die Ausreißer bei den Firmen mit fester Quote sind. Ein Blick in die Daten und Berechnungen verrät uns, dass es sich bei dem Unternehmen ohne Frauen im Aufsichtsrat um „Fresenius SE & Co. KGaA“ aus Bad Homburg handelt. Die beiden Ausreißer nach oben, mit einem vergleichsweise hohen Frauenanteil, sind die „Deutsche Börse Aktiengesellschaft“ und „Henkel AG & Co. KGaA“.

Wenn Sie die Boxplots einmal selber erstellen wollen, haben wir hier den Datensatz und den R-Code dazu verlinkt:

FrauenquoteBoxplots.R (2 KB)

Aufsichtsraete.csv (3 KB)

3.2 Boxplot (Definition)

Definition

Boxplot

Die grafische Darstellung der fünf Fünf-Punkte-Zusammenfassung heißt Box-and-Whisker-Plot oder kurz Boxplot.

1. kleinste Merkmalsausprägung x_{min}
2. unteres Quartil Q_1 ($x_{0,25}$)
3. mittleres Quartil Median ($\tilde{x}_{0,50}$)
4. oberes Quartil Q_3 ($x_{0,75}$)
5. größte Merkmalsausprägung x_{max}

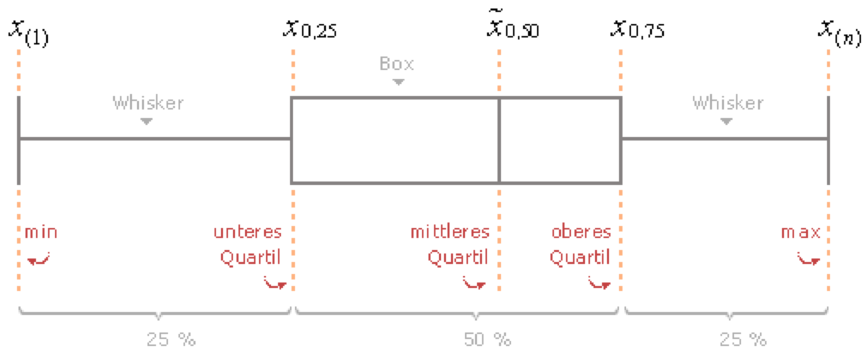


Abb.: Boxplot

Anmerkungen

Ein Boxplot liefert einen sehr guten visuellen Eindruck über essenziell wichtige Kennwerte und Eigenschaften eines Datensatzes:

- Lage des Median
- Streuungsmaße:
  - Spannweite - Ausdehnung eines Boxplot (Differenz x_{max} - x_{min}),
  - Quartilsabstand - Ausdehnung der Box (Differenz Q_3 - Q_1),
- Minimum und Maximum
- Symmetrie: sind die Daten symmetrisch verteilt oder ist die Verteilung asymmetrisch (schief)?

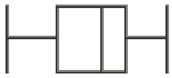
Zwischen dem unteren Quartil Q_1 und dem oberen Quartil Q_3 liegen 50 % der Einzelwerte. Zwischen dem Minimum x_{min} und dem unteren Quartil Q_1 bzw. zwischen dem oberen Quartil Q_3 und dem größten Merkmalswert x_{max} liegt jeweils ein Viertel der Einzelwerte.

Ist ein Boxplot symmetrisch bzw. asymmetrisch, so ist dies auch die zugehörige Häufigkeitsverteilung.

Mit Hilfe von Boxplots lassen sich verschiedene Beobachtungsreihen gut vergleichen.

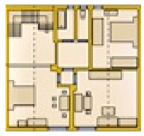
3.3 Boxplot (Übung 1/2)

In der folgenden Übung können Sie sich von der Wirksamkeit des Boxplot als „Datenanzeige“ überzeugen.



Geordnete Daten

Mietangebote Einzimmerwohnungen (in €)



231 247 251 254 256 256 263 277 288 301 302 306 322 333 333 333 340 348 352 354 359 363 364 368 368 369 373 373 374 375 378 380 382 384 384 384 384 388 388 395 395 395 395 405 405 409 413 417 420 423 431 432 433 444 446 448 454 461 473 474 475 477 481 500 500 501 506 506 507 511 512 519 535 538 546 546 552 553 564 581 589 609 609 628 666 681 691 699 727 742 832 916 947 1008 1151 1175 1190 1253 1429

Zeichnen / Berechnen

Übung QBX-09

Mietpreise · · · · · 5 Punkte

Stellen Sie die Mietpreise mit Hilfe eines Boxplots dar.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

3.4 Boxplot (Übung 2/2)

Bei einem Online-Getränkehändler wurden in den Monaten Januar 2020 bis Dezember 2021 folgende Anzahlen von Reklamationen registriert. Beobachtungen eines Verlaufs werden als Zeitreihen bezeichnet.

Tab.: Beispiel für registrierte Reklamationen

Monat	J	F	M	A	M	J	J	A	S	O	N	D
2020	0	4	1	4	2	8	2	4	3	2	1	6
2021	7	1	6	1	9	7	8	8	4	3	7	0

Formulieren

Übung QBX-10

Reklamationen: Daten

Erläutern Sie, warum es nur einen Merkmalsträger gibt. Wie heißt der Merkmalsträger?

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Formulieren

Übung QBX-11

Reklamationen: Skalierung

Wie ist das Erhebungsmerkmal skaliert?

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-12

Reklamationen: Quartile

Bestimmen Sie die Quartile.

Lösung (Siehe Anhang)

Bearbeitungszeit: 15 Minuten

Zeichnen

Übung QBX-13

Reklamationen: Boxplot

Zeichnen Sie den Boxplot der Erhebung.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Zusammenfassung

- Sie haben in dieser Lerneinheit gelernt Quantile zu berechnen.
- Spezielle Quantile teilen eine Verteilung in Abschnitte gleicher Häufigkeit ein.
- Empirische Quantile werden bei Bedarf durch Interpolation bestimmt.
- Quantile x_p können am Schaubild der empirischen Verteilungsfunktion abgelesen werden.
- Die Fünf-Punkte-Zusammenfassung einer Verteilung ist die Grundlage für die Charakterisierung von Daten mit Hilfe des Box-and-Whisker-Plots.
- Der Boxplot vermittelt einen schnellen Überblick über die Daten und ist fast universell einsetzbar.

Sie sind am Ende dieser Lerneinheit angelangt. Auf den folgenden Seiten finden Sie noch die Übungen zur Wissensüberprüfung und die Übungen für die Statistiksoftware R.

Wissensüberprüfung

Nutzen Sie die folgenden Übungen um Ihr Wissen zu testen. Versuchen Sie erst die Aufgaben selbst zu lösen, bevor Sie sich die Lösung ansehen.

Multiple Choice

Übung QBX-14

Markieren Sie, ob die Aussagen richtig oder falsch sind.

	Richtig	Falsch	Auswertung
Die Quartile $x_{0.25}$, $\tilde{x}_{0.50}$, $x_{0.75}$ teilen eine Verteilung in drei Abschnitte gleicher Häufigkeit ein.			
Wenn n*p nicht ganzzahlig ist, dann sind p-Quantile x_p realisierte Merkmalswerte.			
In einem Box-Plot teilt der Median (Quartil $\tilde{x}_{0.50}$) den Raum zwischen den Werten $x_{(1)}$ und $x_{(n)}$ symmetrisch.			
Wenn der Wert n*p ganzzahlig ist, dann ist die Intervallmitte zwischen den beiden benachbarten Werten $x_{(np)}$ und $x_{(np+1)}$ der geordneten Reihe zu betrachten: $x_p = \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right)$			
Für klassierte Daten können Quantile nicht bestimmt werden.			

Multiple Choice

Übung QBX-15

Das untere Quartil der monatlichen Gehälter von 100 Befragten in Berlin ist 1550 €. Welche Aussagen sind falsch und welche richtig?

	Richtig	Falsch	Auswertung
Ein Viertel der monatlichen Gehälter der 100 Befragten in Berlin ist gleich oder höher als 1550 Euro.			
Drei Viertel der monatlichen Gehälter der 100 Befragten in Berlin sind weniger als 1550 Euro.			
Ein Viertel der monatlichen Gehälter der 100 Befragten in Berlin ist gleich oder weniger als 1550 Euro.			
Drei Viertel der monatlichen Gehälter der 100 Befragten in Berlin sind gleich oder höher als 1550 Euro.			
Drei Viertel der monatlichen Gehälter der 100 Befragten in Berlin sind höher als 1550 Euro.			

Multiple Choice

Übung QBX-16

Die Fünf-Punkte-Zusammenfassung einer Datenreihe besteht aus folgenden Komponenten: $x_{min} = 21$, $Q_1 = 45$, $\tilde{x}_{0.5} = 60$, $Q_3 = 87$, $x_{max} = 105$.

Welche Aussagen sind falsch und welche richtig?

	Richtig	Falsch	Auswertung
Die Hälfte der Beobachtungswerte ist gleich oder kleiner als 60.			
Ein Viertel der Beobachtungswerte ist größer als 45.			
Drei Viertel der Beobachtungswerte sind gleich oder kleiner als 87.			
Ein Viertel der Beobachtungswerte ist größer als 87.			
Die Hälfte der Beobachtungswerte ist größer als 60.			
Ein Viertel der Beobachtungswerte ist gleich oder kleiner als 87.			
Drei Viertel der Beobachtungswerte sind größer als 45.			
Drei Viertel der Beobachtungswerte sind größer als 87.			
Ein Viertel der Beobachtungswerte ist gleich oder kleiner als 45.			
Die Hälfte der Beobachtungswerte ist gleich oder größer als 60.			

Übungen mit der Statistiksoftware R

Die in der Lerneinheit behandelten Themen können Sie anhand der folgenden Übungsaufgaben mit der Statistiksoftware **R** bearbeiten. Die Aufgabenstellung finden Sie in der jeweiligen Übung. Um die Übungen zu bearbeiten, muss die Software „**R**“ auf Ihrem Rechner installiert sein.

Installationshinweise [Manuals | **R** Installation and Administration]

Für einige Übungen stehen auch Musterlösungen für das Programm Excel bereit.

Statistiksoftware R

Übung QBX-17a

Angeln

Unter den Freunden des Angelsports stellt sich oft die Frage nach den Besten ihrer Zunft. 12 Angler und Anglerinnen wollten es genau wissen. Sie führten über fünf Jahre hinweg ein Fangbuch, in dem sie das Gewicht der von ihnen gefangenen Fische (sie einigten sich auf eine Fischart, den Dorsch) in Kilogramm verzeichneten. In der Datei **`angeln.txt`** sind je Person (A01-A012) die 20 schwersten gefangenen Dorsche zu finden.

`angeln.txt` (2 KB)

Alle sind der Meinung, dass nicht der größte Fisch den besten Angler bzw. die beste Anglerin ausmacht, sondern das regelmäßige Fangen schwerer Fische.

Aufgabe

1. Veranschaulichen Sie sich die Fangergebnisse mit Hilfe des Boxplot!
2. Wer ist ihrer Meinung nach der beste Angler bzw. die beste Anglerin? Warum?
3. Erstellen Sie einen neuen Datensatz, der als Beobachtungen den Median der schwersten gefangenen Fische pro Person enthält! Stellen Sie ihn im Boxplot dar! Was sagt dieser Boxplot aus?

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Statistiksoftware R

Übung QBX-17b

Unfallklinik

Im folgenden Datensatz ist ein Auszug einer Krankenhausstatistik zu sehen. Erfasst wurde von den Patienten, die wegen eines Verkehrsunfalls eingeliefert wurden, das Alter, das Geschlecht, der Grad der Verletzung (1=leichte, 2=mittlere, 3=schwere), die Aufenthaltsdauer in Tagen und die Art des Fahrzeuges (1=Fußgänger, 2=Fahrradfahrer, 3=Motorradfahrer und 4=Autofahrer).

i	Alter	Geschlecht	Verletzungsgrad	Verweildauer	Verkehrsmittel
1	17	1	1	0	1
2	18	1	3	22	4
3	22	2	3	12	3
4	22	2	3	8	3
5	27	2	3	26	3
6	28	2	1	1	2
7	29	2	3	4	4
8	33	1	2	5	3
9	34	1	3	20	2
10	35	1	2	8	2
11	38	2	1	1	4
12	41	2	3	12	4
13	54	2	2	7	1

Aufgabe

1. Stellen Sie jeweils die Häufigkeitsverteilungen der Verkehrsmittel, des Geschlechts und des Grades der Verletzungen im Säulendiagramm dar!
2. Bestimmen Sie Minimum, Maximum, Quantile und Median beim Grad der Verletzung.
3. Lassen Sie jeweils einen Boxplot erstellen zu den Variablen Dauer und Alter! Welche Maßzahlen müsste man berechnen, wollte man diesen Boxplot von Hand zeichnen? Wie lauten diese Maßzahlen konkret?

Lösung mit R und mit Excel (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Statistiksoftware R

Übung QBX-17c

Wasserstände

In folgendem aufgeführt finden Sie ein paar Wasserstandsmesswerte vom (Hoch-) Wasserstand des Mains bei Schweinfurt. Sie sind in Metern angegeben.

2,08	2,05	2,1	2,08	2,07	2,08	2,09	2,1	2,08	2,11	2,06	2,12
2,11	2,08	2,17	2,12	2,06	2,17	2,12	2,05	2,21	2,09	2,02	2,24
2,06	2,03	2,26	2,04	2,03	0,28	2,04	2,08	2,28	2,1	2,25	

Aufgaben

Analysieren Sie die Daten nach folgenden Fragestellungen:

1. Ordnen Sie den Datensatz und erstellen Sie eine Häufigkeitstabelle. Wo würden Sie den Median der Daten anhand der Häufigkeitstabelle sehen?
2. Berechnen sie die Quantile X_p (p = 0,1 , 0,25 , 0,5 , 0,75 , 0,9)!

   Was sagen sie jeweils aus?
3. Lassen Sie zur graphischen Übersicht einen Boxplot über die Daten zeichnen! Welche der Quantile finden Sie im Boxplot wieder?

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Statistiksoftware R

Übung QBX-17d

Prosecco

Die Inhaberin des Weinfachgeschäftes Maestro hat eine neue Sorte Prosecco in ihr Sorti- ment aufgenommen und interessiert sich für die Anzahl der pro Tag verkauften Flaschen dieser Sorte. Vier Wochen lang hat sie täglich die Anzahl der verkauften Flaschen notiert und in der folgenden Tabelle zusammengefasst.

Anzahl Flaschen	Tage
0	3
2	5
3	7
5	4
7	3
8	2

Aufgaben

1. Geben Sie die Quartile der Verteilung an und zeichnen Sie ein Boxplot.
2. Interpretieren Sie kurz das Ergebnis.

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Berechnen

Übung QBX-17e

Störfälle

In einem Zulieferbetrieb wurden in den Monaten Januar 2021 bis Dezember 2022 folgende Anzahl von Störfällen registriert.

Monat	Jahr 2021	Jahr 2022
Jan	0	7
Feb	4	1
Mae	1	6
Apr	4	1
Mai	2	9
Jun	8	7
Jul	2	8
Aug	4	8
Sep	3	4
Okt	2	3
Nov	1	7
Dez	6	0

Aufgabe:

1. Bestimmen Sie die Quartile der einzelnen Jahre und des gesamten Erhebungszeitraums.
2. Zeichnen Sie dazugehörigen die Boxplots.

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Berechnen

Übung QBX-17f

Grundstücke

Die Inhaberin des Weinfachgeschäftes Maestro hat 10 neue Angebote von Grundstücken in Veneto bekommen. Die Grundstücke haben folgende Preise umgerechnet in Tausend Euro:

10,5	5	4,3	12	14	10,5	6,2	5	7,3	10,5

Aufgabe:

1. Geben Sie die empirische Verteilungsfunktion graphisch an.
2. Bestimmen Sie grafisch den Wert der Verteilungsfunktion an der Stelle 13 und interpretieren Sie diesen.
3. Bestimmen Sie graphisch den Median der Preise von Angeboten.
4. Bestimmen Sie graphisch das 20 %-Quantil und interpretieren Sie das Ergebnis.

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Zusätzliche Übungsaufgaben

Zum Vertiefen der vorgestellten Inhalte finden Sie hier zusätzliche Übungsaufgaben zur freiwilligen Bearbeitung.

Berechnen

Übung QBX-18

Körperlänge von Studierenden

Wir betrachten folgenden geordneten Datensatz des Merkmals „Körpergröße“ von 55 Studierenden der Berliner Hochschule für Technik:

158, 160, 160, 160, 160, 162, 163, 163, 163, 164, 165, 165, 165, 165, 165, 166, 166, 166, 167, 168, 168, 169, 170, 170, 171, 172, 172, 172, 172, 173, 174, 175, 175, 175, 178, 179, 180, 181, 181, 183, 183, 183, 185, 185, 186, 186, 186, 190, 190, 192, 194, 194, 198, 201, 204

(Alle Angaben in Zentimeter)

a) Bestimmen Sie die Quartile Q_1 und Q_3.

b) Bestimmen Sie den Median und die Quartilsmitte.

c) Interpretieren Sie ihre Ergebnisse aus Aufgabe a) und b).

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Berechnen

Übung QBX-19

Spiele der Europameisterschaft

Während der Fußball-Europameisterschaft 2012 in der Ukraine und Polen fanden 31 Spiele statt. Im Folgenden finden Sie eine Liste mit der Anzahl der Tore je Spiel:

2, 5, 3, 2, 1, 1, 1, 1, 5, 3, 3, 3, 2, 4, 2, 4, 2, 1, 2, 3, 2, 5, 1, 2, 1, 6, 2, 6, 6, 3, 4

(Quelle: www.sportschau.de)

a) Erstellen Sie einen geordneten Datensatz.

b) Bestimmen Sie den Median sowie das untere und das obere Quartil.

c) Interpretieren Sie ihre Ergebnisse.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-20

Alter von Studierenden der Mathematik

Aus den Studenten des ersten Semesters des Studiengangs Mathematik wurden 25 Studenten zufällig ausgewählt und nach ihrem Alter befragt.

22, 20, 21, 23, 23, 24, 25, 28, 25, 23, 22, 21, 23, 24, 25, 21, 24, 23, 26, 24, 23, 25, 23, 21, 25

(Alle Angaben in Jahren)

Berechnen Sie den Median, das untere Quartil, das obere Quartil und geben Sie den Boxplot an.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-21

Kraftstoffverbrauch

Um den Spritverbrauch zu kontrollieren, schreibt Herr Salzmann stets auf, wieviel Liter Diesel sein Auto auf 100 km verbraucht.

6,3	6,7	7,6	5,8	5,9	5,9	7,4	5,6	7,2	6,7	6,3
6,5	7,2	7,1	5,7	6,4	6,8	7,2	5,8	6,4	6,2	

(Alle Angaben in Liter)

Zeichnen Sie den Boxplot und interpretieren Sie ihre Ergebnisse.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-22

Jana radelt zur Arbeit

Jana schreibt immer auf, wie lange sie mit ihrem Fahrrad für den Weg zur Arbeit braucht. Sie erhalten folgende Werte:

12, 14, 13, 15, 19, 15, 17, 12, 20, 18, 14, 15, 23, 21, 13, 15, 17, 27, 20

(Angaben in Minuten)

Stellen Sie die Fahrzeiten in einem Boxplot dar.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-23

Der Sushi-Lieferant

Anja bestellt häufig bei Sushi-Yuko. Sie notiert sich jedes Mal die Zeit zwischen der Bestellung und Lieferung der Sushi.

20, 19, 35, 23, 26, 24, 33, 45, 39, 35, 25, 28, 23, 41, 38, 23, 43, 35, 23, 29, 32, 30, 22

(Angaben in Minuten)

Stellen Sie die Lieferzeiten in einem Boxplot dar und interpretieren Sie kurz ihre Ergebnisse.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-24

Versicherungsverträge

Für die vergangenen 20 Arbeitstage eines Versicherungsmaklers wurden hinsichtlich des
Merkmals „Anzahl der abgeschlossenen Verträge“ folgende Zahlen notiert:

12, 15, 20, 23, 30, 13, 18, 17, 19, 25, 26, 14, 26, 19, 30, 32, 26, 28, 25, 20

(Anzahl abgeschlossener Verträge)

Zeichnen Sie den Boxplot und interpretieren Sie kurz ihre Ergebnisse.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung QBX-25

Bruttoverdienst

Die folgende Häufigkeitstabelle zeigt den monatlichen Bruttoverdienst von 100 Befragten:

i	$x_{i-1} < X \leq x_i$	n_i	h_i
1	0 < X ≤ 500	10	0,1
2	500 < X ≤ 1000	18	0,18
3	1000 < X ≤ 1500	30	0,3
4	1500 < X ≤ 2000	32	0,32
5	2000 < X ≤ 3000	7	0,07
6	3000 < X ≤ 6000	3	0,03

a. Geben Sie die empirische Verteilungsfunktion $F_n(x)$ an.
b. Bestimmen Sie $F_n(1740)$ und interpretieren Sie kurz ihre Ergebnisse.
c. Bestimmen Sie welcher Verdienst von 70 Prozent der Haushalte nicht überschritten wird.
d. Interpretieren Sie kurz ihre Ergebnisse aus c).

Lösung (Siehe Anhang)

Bearbeitungszeit: 15 Minuten

Berechnen

Übung QBX-26

Lebensdauer von Sparlampen

Die Firma Licht GmbH stellt Sparlampen her. Um deren Lebensdauer besser einschätzen zu können, wird eine Stichprobe aus der täglichen Produktion gezogen und die Brenndauer dieser Glühbirnen beobachtet. Die Ergebnisse wurden in der folgenden Häufigkeitstabelle zusammengefasst (das Merkmal X gibt die Brenndauer der Glühbirnen in Stunden an):

i	$x_{i-1} < X \leq x_i$	n_i
1	0 < X ≤ 4000	7
2	4000 < X ≤ 6000	18
3	6000 < X ≤ 7000	21
4	7000 < X ≤ 9000	41
5	9000 < X ≤ 10000	10
6	10000 < X ≤ 12000	3

a) Erweitern Sie die Tabelle um die relativen Häufigkeiten und um die kumulierten relativen Häufigkeiten.

b) Berechnen Sie den Anteil der Sparlampen die mindestens 6001 Stunden gebrannt haben.

c) Berechnen Sie den Anteil der Sparlampen die höchstens 9000 Stunden gebrannt haben.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Berechnen

Übung QBX-27

Geburtensituation

In einem mittelgroßen Stadt wurden im Jahr 2021 eine altersspezifische Geburtensituation untersucht. Es liegen folgende Werte vor:

Alter	Anzahl der Geborenen
[15; 20)	70
[20; 25)	289
[25; 30)	369
[30; 35)	178
[35; 40)	83
[40; 45)	11

a) Erweitern Sie die Tabelle um die relativen Häufigkeiten und kumulierten relativen Häufigkeiten.

b) Bestimmen Sie die zugehörigen Quartile.

c) Bestimmen Sie den Interquartilsabstand.

d) Interpretieren Sie kurz ihre Ergebnisse.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Appendix

Exkurs Gaußklammer

Das Symbol $\lfloor \ldots \rfloor$ steht hier für die sogenannte Gaußklammer. Mit der Gaußklammer wird für jede reelle Zahl z die nächst kleinere ganze Zahl angegeben, d. h. es wird auf die nächste ganze Zahl abgerundet. Mathematisch ausgedrückt wird z die größte ganze Zahl, die kleiner oder gleich z ist, zugeordnet:

$\lfloor z \rfloor := max\{k \in \mathbb{Z} | k \leq z\}$

So geht's:

$\lfloor 11/3 \rfloor = \lfloor 3 + 2/3 \rfloor = 3$
$\lfloor 3{,}1415 \rfloor = 3$
$\lfloor 47{,}11 \rfloor = 47$
$\lfloor 0{,}5 * 99 \rfloor = \lfloor 49{,}5 \rfloor = 49$
$\lfloor 1{,}99999999999999999 \rfloor = 1$
$\lfloor 1 \rfloor = 1, \lfloor 2 \rfloor = 2, \lfloor 3 \rfloor = 3$, ...
$\lfloor 2{,}00000000000000001 \rfloor = 2.$

Die Funktion $y(z) = \lfloor z \rfloor$ ist eine Treppenfunktion, mit Sprungstellen an den ganzzahligen z-Werten.

Lösung für Übung QBX-01

Mietpreise und Quantilberechnungen

Unteres Quartil $x_{0,25}$ (Q_1):

Es gilt $n \cdot p = 99 \cdot 0{,}25 = 24{,}75$, also ist $n \cdot p$ nicht ganzzahlig. Laut Definition müssen wir um $\lfloor n \cdot p + 1 \rfloor$ bestimmen und erhalten mit $\lfloor 99 \cdot 0{,}25 + 1 \rfloor = \lfloor 25{,}75 \rfloor = 25$ den Wert von $x_{(25)}$ also 368 als unteres Quartil $x_{0,25}$.

Die Interpretation des unteren Quartils $x_{0,25}$: ein Viertel (25 %) der Mieten der 99 freien Einzimmerwohnungen in Berlin ist gleich oder weniger als 368 € bzw. drei Viertel (75 %) der Mieten der Zweizimmerwohnungen sind gleich oder höher als 368 €.

Oberes Quartil $x_{0,75}$ (Q_3):

Die entsprechende Betrachtung führt zu $x_{0,75} = x_{(75)} = 546.$

Die Interpretation des oberen Quartils, $x_{0,75}$: drei Viertel (75 %) der Mieten der 99 freien Zweizimmerwohnungen in Berlin sind gleich oder weniger als 546 € bzw. ein Viertel (25 %) der Mieten der Einzimmerwohnungen ist gleich oder höher als 546 €.

Die Quartilsmitte ist folglich: $\frac{1}{2}(x_{0,25} + x_{0,75}) = 457$

Lösung für Übung QBX-02

Median und Quartilsmitte

Die Quartilsmitte ist aus Aufgabe 1 bekannt.

Da n = 99 ungerade ist erhalten wir mit $\frac{n+1}{2} = 50$ den Median als $x_{(50)}$. Wir finden die Zahl 423 als 50. Wert in der geordneten Datenreihe.
Wir erhalten also für unser Beispiel unterschiedliche Werte für die Quartilsmitte 457 und Median 423.

Lösung für Übung QBX-03

Mietpreise und Quantilberechnungen (noch einmal)

Unteres Quartil $x_{0,25}$:
In diesem Fall ist $n \cdot p = 50 \cdot 0{,}25 = 12{,}5$ also keine ganze Zahl.
Nun ist $n \cdot p + 1 = 13{,}5$, also die Zahl 13.

Folglich ist x_p = $x_{0,25}$= $x_{(13)} = 322.$
Oberes Quartil $x_{0,75}$: $n \cdot p = 50 \cdot 0{,}75 = 37{,}5$. Also: $x_{0,25}$ = $x_{(38)} = 388.$

Quartilsmitte:

Quartilsmitte:	$\frac{1}{2}(322 + 388) = 355$
Median:	In diesem Fall ist $n \cdot p = 50 \cdot 0{,}50 = 25$ also ganzzahlig.

Median:

Der Median ist somit

$$\frac{1}{2}[x_{(np)} + x_{(np+1)}] = \frac{1}{2}[x_{(25)} + x_{(26)}] = \frac{1}{2}[(368 + 369) = 368{,}5$$

Er ist also nicht identisch mit der Quartilsmitte.

Lösung für Übung QBX-04

Grundstückspreise: Verteilung

Zur Berechnung der Verteilungsfunktion $F_n(x)$ werden die N=10 Angebote nach dem Merkmal X = „Preis" aufsteigend geordnet und die absoluten und relativen Häufigkeiten (n_i und h_i) der realisierten Merkmalsausprägungen x_i ermittelt. Verteilungsfunktion $F_n(x)$ ergibt sich durch Kumulation der relativen Häufigkeiten h_i.

Häufigkeitstabelle mit der empirischen Verteilungsfunktion: Preise der Angebote.

Index i	Merkmalsausprägung x_i	absolute Häufigkeit n_i	relative Häufigkeit h_i	Verteilungsfunktion $F_n(x)$
1	4,3	1	0,1	0,1
2	5,0	2	0,2	0,3
3	6,2	1	0,1	0,4
4	7,3	1	0,1	0,5
5	10,5	3	0,3	0,8
6	12,0	1	0,1	0,9
7	14,0	1	0,1	1

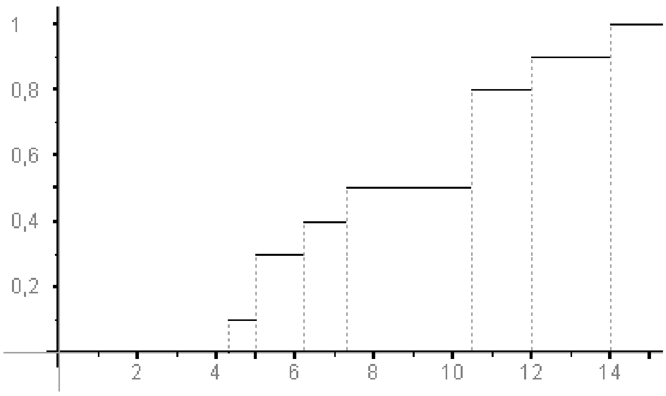


Abb.: Grafische Darstellung der empirischen Verteilungsfunktion

Lösung für Übung QBX-05

Grundstückspreise: Verteilung 2

$F_n(13) = 0{,}90$, d. h. mindestens 90 % der Angebotspreise sind gleich oder niedriger als 13.000 € bzw. 10 % der Preise sind gleich oder höher als 13.000 €.

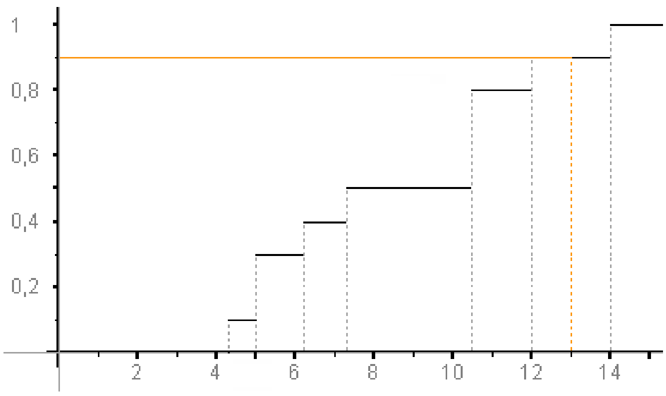


Abb.: Grafische Bestimmung von $F_n(13)$.

Lösung für Übung QBX-06

Grundstückspreise: Median

Aus der Graphik ergibt sich $\tilde{x}_{0{,}50} = 8{,}9$, d. h. mindestens die Hälfte der Preise sind gleich oder niedriger als 8.900 € bzw. mindestens die Hälfte der Preise ist gleich oder höher als 8.900 €.

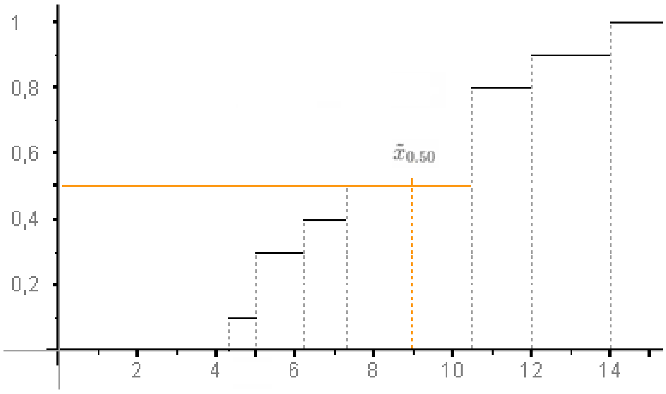


Abb.: Grafische Bestimmung von $\tilde{x}_{0{,}50}$

Lösung für Übung QBX-07

Grundstückspreise: Quantil

Aus der Graphik erhält man $x_{0{,}2} = 5$
d. h. mindestens 20 % der Angebotspreise sind gleich oder niedriger als 5.000 € bzw. maximal 80 % sind gleich oder höher als 5.000 €.

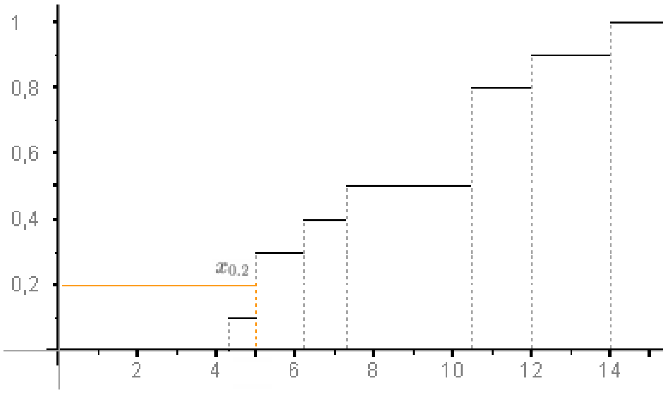


Abb.: Grafische Bestimmung von $x_{0{,}2}$

Lösung für Übung QBX-08

Verkaufszahlen: Quartile

Da es sich bei den Verkaufszahlen um ein diskretes absolutskaliertes Merkmal handelt, ist es günstig zunächst die kumulierten Häufigkeiten zu bilden. (Es kommen viele Werte mehrfach vor.)

Flasche	Tage
0	3
2	5
3	7
5	4
7	3
8	2

Stichprobenumfang $n = 24$. Damit folgt sofort wegen $n \cdot 0{,}25 = 6$

$$Q_1 = \frac{\left(x_{(6)} + x_{(7)}\right)}{2} = 2$$

Median = $\frac{\left(x_{(12)} + x_{(13)}\right)}{2} = 3$

$$Q_3 = \frac{\left(x_{(18)} + x_{(19)}\right)}{2} = 5$$

Interpretation:
Die mittleren 50 % der Verkäufe liegen zwischen zwei und fünf Flaschen Prosecco pro Tag. Es gibt Tage an denen keine Flasche verkauft wird. Das Maximum liegt bei acht verkauften Flaschen pro Tag. Der Median und damit ein mittlerer Wert sind drei verkaufte Flaschen täglich.

Lösung für Übung QBX-09

Mietpreise: ..... (5 Punkte)

Als Quartile erhält man folgende Werte:

$x_{0,25}$ = 368,
$\tilde{x}_{0,50}$ = 423,
$x_{0,75}$ = 546.

Der zugehörige Boxplot sieht folgendermaßen aus:

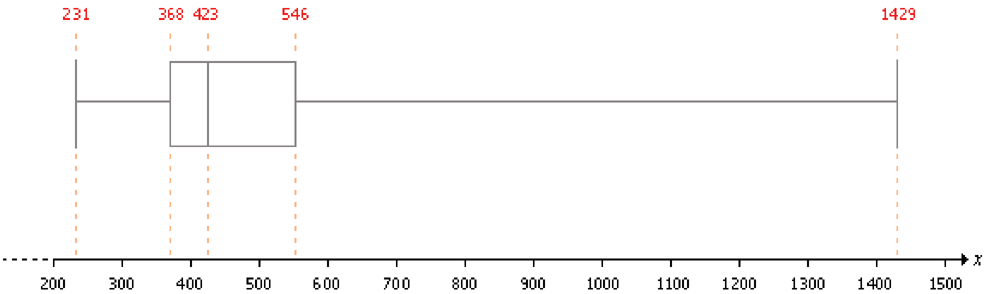
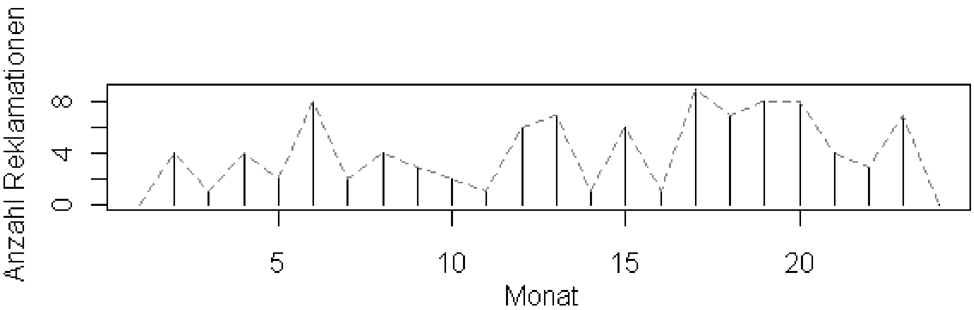


Abb.: Boxplot für die Mieten der Einzimmerwohnungen

Lösung für Übung QBX-10

Reklamationen: Daten

Genau genommen haben wir es mit einer einzigen Realisierung einer Zeitreihe, den Reklamationen in den Jahren 2020 und 2021 zu tun. Merkmalsträger ist die Getränkehandlung.



Vergisst man diesen Bezug, dann handelt es sich um 24 Beobachtungen und Merkmalsträger ist dann jeder Monat im betrachteten Zeitraum.
n = 24

Merkmalsträger:	Monat; 24 Monate;
statistische Gesamtheit:	Monat (sachlich),
Identifikationsmerkmale:	2020 - 2021 (zeitlich);
Erhebungsmerkmal:	Anzahl der Reklamation im Monat;
Skala:	kardinal bzw. absolut skaliert;
Urliste:	alle n=24 erfassten Anzahlen

Lösung für Übung QBX-11

Reklamationen: Skalierung

Das Erhebungsmerkmal ist absolut skaliert und diskret.

Lösung für Übung QBX-12

Reklamationen: Quartile

Man ordnet die 24 Werte zunächst der Größe nach an:

0 0 1 1 1 1 2 2 2 3 3 4 4 4 4 6 6 7 7 7 8 8 8 9.

Unteres Quartil $x_{0,25}$:

$n \cdot p = 24 \cdot 0{,}25 = 6$ (ganzzahlig) und $x_{0,25}$ ist somit also

$$x_{0,25} = \frac{1}{2}(x_{(np)} + x_{(np+1)}) = \frac{1}{2}(x_{(6)} + x_{(7)}) = \frac{1}{2}(1+2) = 1{,}5$$

Interpretation:

$x_{0,25} = 1{,}5$, d. h. in mindestens einem Viertel (25 %) des Untersuchungszeitraums wurden 1,5 oder weniger Reklamationen pro Monat beobachtet bzw. in mindestens drei Viertel (75 %) des Untersuchungszeitraums wurden mehr als 1,5 Störfälle pro Monat beobachtet.

Median $\tilde{x}_{0,50}$:

$n \cdot p = 24 \cdot 0{,}50 = 12$ also ganzzahlig.

Der Median ist somit:

$$\tilde{x}_{0,50} = \frac{1}{2}(x_{(np)} + x_{(np+1)}) = \frac{1}{2}(x_{(12)} + x_{(13)}) = \frac{1}{2}(4+4) = 4$$

Interpretation:

Median $X_{0,50}$ = 4, d. h. in mindestens 50 % des Untersuchungszeitraums wurden 4 oder weniger Reklamationen pro Monat beobachtet bzw. in mindestens 50 % des Untersuchungszeitraums wurden mehr als 4 Reklamationen pro Monat beobachtet.

Oberes Quartil $x_{0,75}$:

$n \cdot p = 24 \cdot 0{,}75 = 18$ (ganzzahlig) und $x_{0,75}$ ist somit also

$$x_{0,75} = \frac{1}{2}(x_{(np)} + x_{(np+1)}) = \frac{1}{2}(x_{(18)} + x_{(19)}) = \frac{1}{2}(7+7) = 7$$

Interpretation:

$x_{0,75} = 7$, d. h. in mindestens drei Viertel (75 %) des Untersuchungszeitraums wurden 7 oder weniger Reklamationen pro Monat beobachtet bzw. in mindestens einem Viertel (25 %) des Untersuchungszeitraums wurden mehr als 7 Reklamationen pro Monat beobachtet.

Lösung für Übung QBX-13

Reklamationen: Boxplot

Boxplot

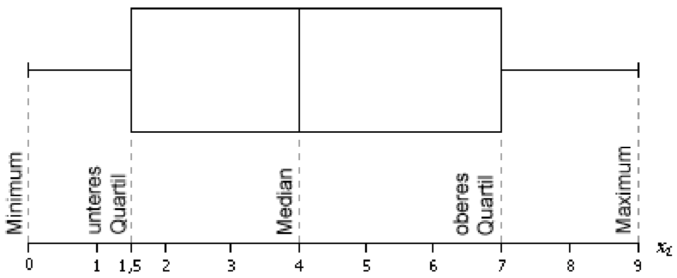
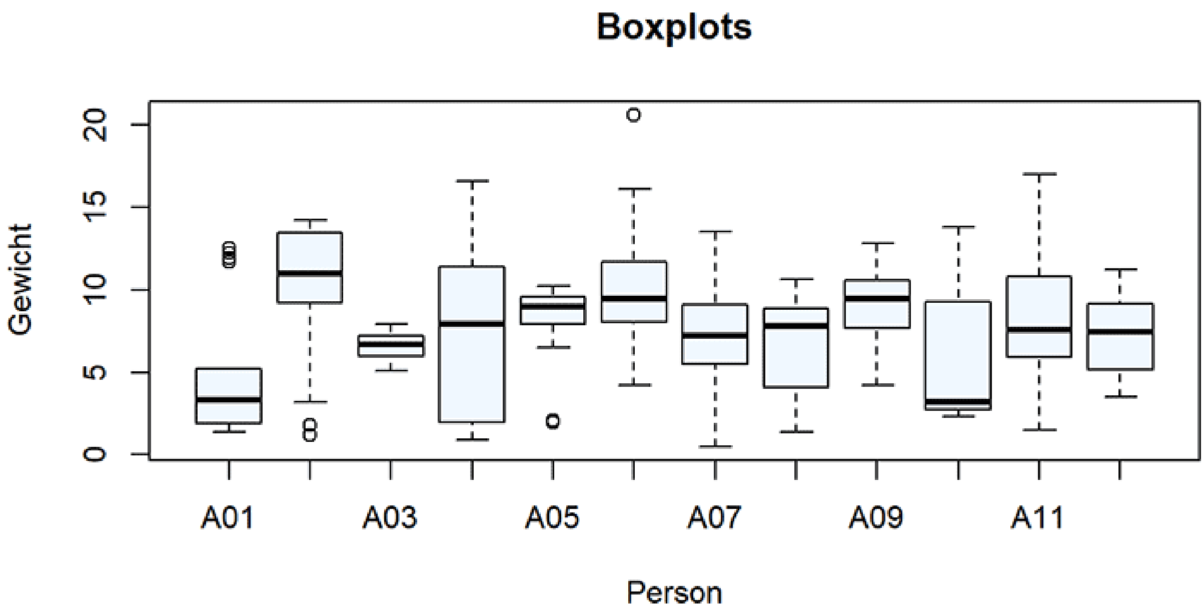


Abb.: Boxplot für die Anzahl der Reklamationen im Monat

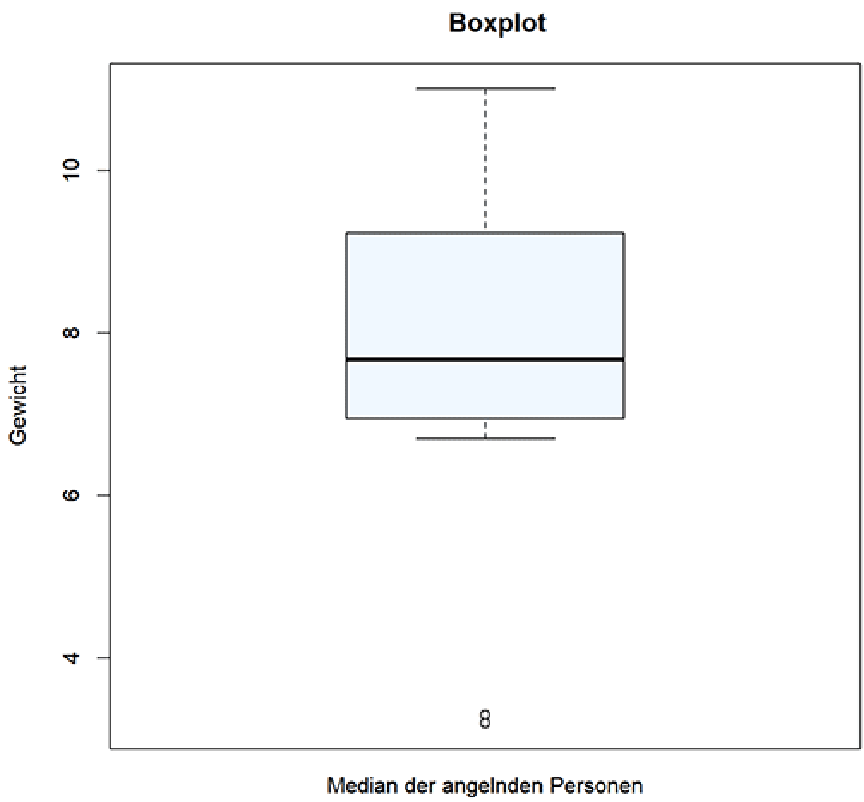
Lösung Übung: QBX-17a

Angeln

1. Boxplot der Daten **`angeln.txt`**



2. An den Boxplots oben ist zu erkennen, dass die Person mit der Nummer A02 am besten abschneidet. Die zentralen 50 % der Fänge sowie der Median liegen im Vergleich zu den anderen Anglern am höchsten. Zwar haben einige vereinzelt größere Fänge gemacht, doch da es bei dem Vergleich um die Regelmäßigkeit schwerer Fänge geht, liegt der Angler oder die Anglerin A02 vorne.
3. In dem Boxplot unten sind die Quantile der einzelnen Mediane der Angelnden zu erkennen. Die zentralen 50 % liegen recht dicht beieinander. Außerdem gibt es nur zwei „Ausreißer“. Der Median liegt bei 7,675 kg.



Lösung mit R

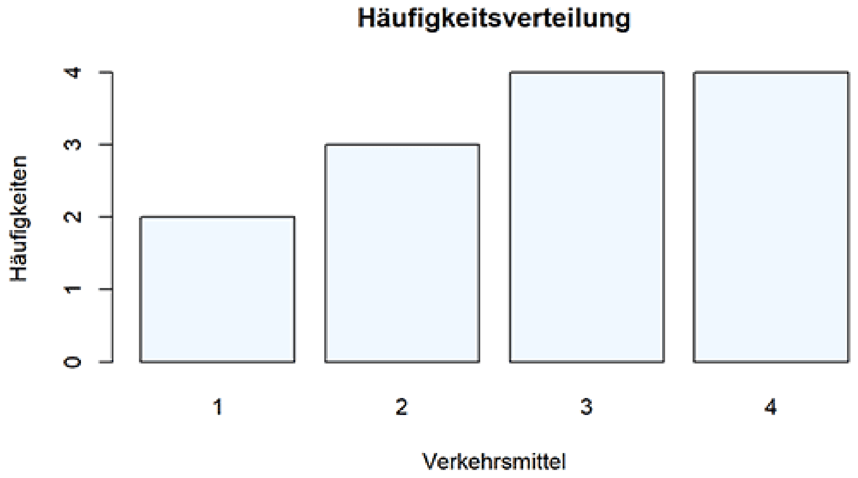
angeln_loesung.R

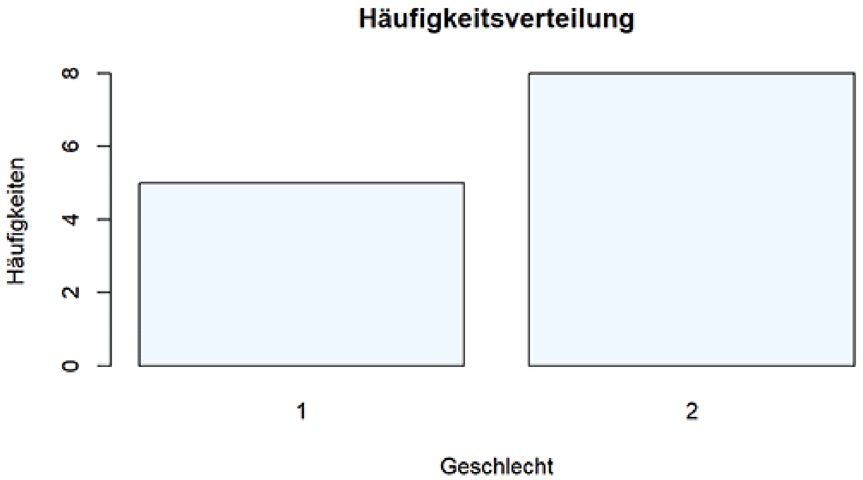
```
# Einlesen der Daten aus der Datei angeln.txt
angeln<-read.table("angeln.txt",sep="\t",header=TRUE)

attach(angeln)

#  1)
boxplot(
  angeln,
  main = "Boxplots",
  ylab = "Gewicht",
  xlab = "Person",
  col = "aliceblue"
)


#  3) Werte der einzelnen Mediane einlesen und als Boxplot darstellen
angeln.median <-
  c(
    median(A01),
    median(A02),
    median(A03),
    median(A04),
    median(A05),
    median(A06),
    median(A07),
    median(A08),
    median(A09),
    median(A10),
    median(A11),
    median(A12)
  )
boxplot(
  angeln.median,
  xlab = "Median der angelnden Personen",
  ylab = "Gewicht",
  main = "Boxplot",
  col = "aliceblue"
)
median(angeln.median)
```

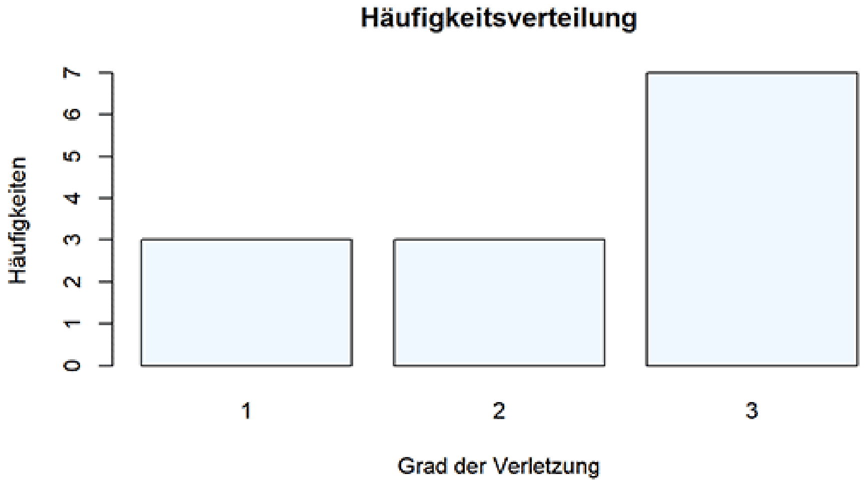
Lösung für Übung QBX-17b

Unfallklinik

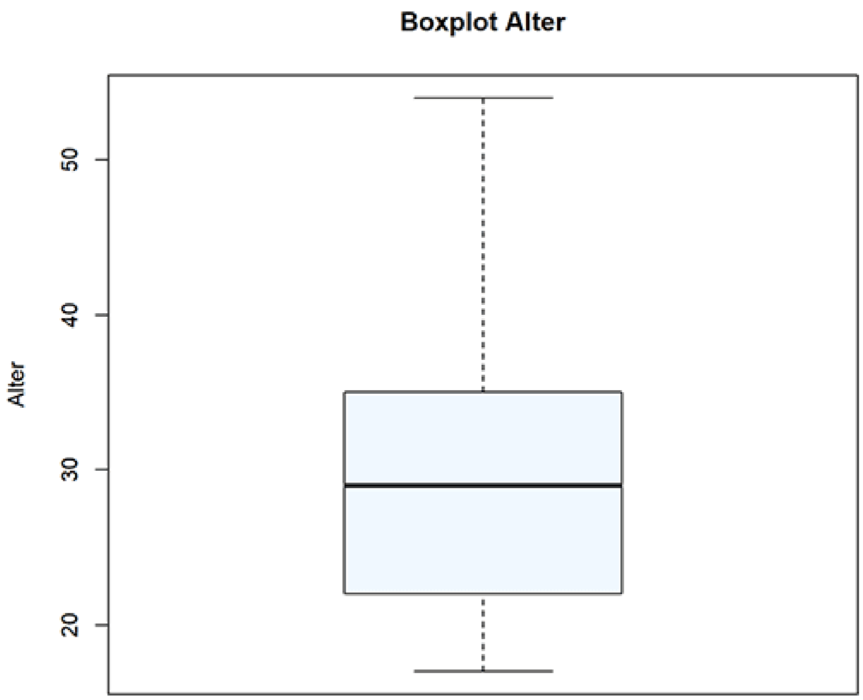
1. Säulendiagramme der Häufigkeitsverteilungen der Verkehrsmittel, des Geschlechts und des Grades der Verletzungen.

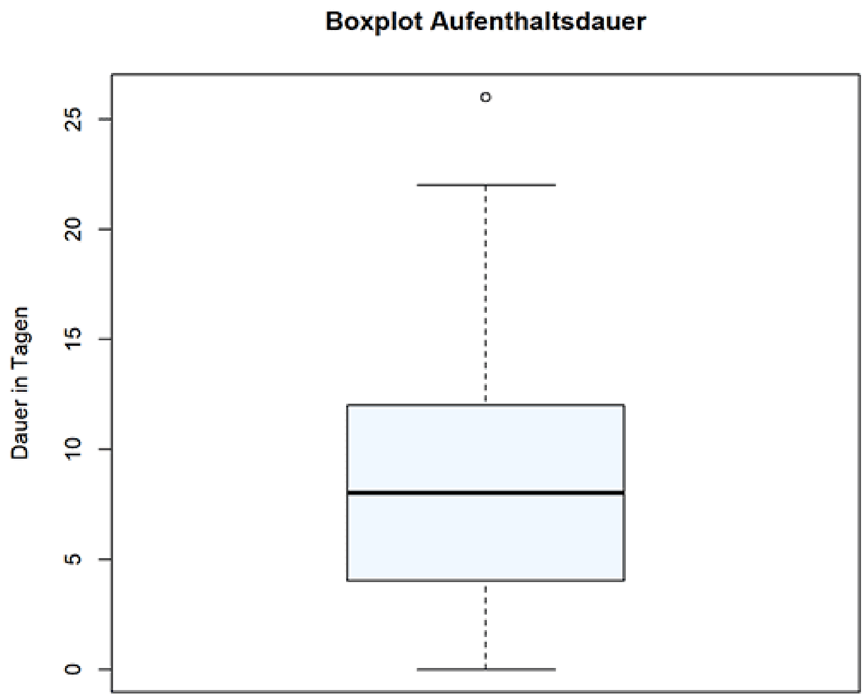






2. Beim Grad der Verletzung beträgt das Minimum 1, das Maximum beträgt 3, $x_{0,25} = 2$, Median $\tilde{x}_{0,50} = 3$ und $x_{0,75} = 3$
3. Je ein Boxplot für Aufenthaltsdauer und für das Alter. Als Maßzahlen werden das 25 %, 50 % und das 75 % Quantil benötigt.


Boxplot Alter
Alter
20
30
40
50


Boxplot Aufenthaltsdauer
Dauer in Tagen
0
5
10
15
20
25

Lösung mit R

unfallklinik_loesung.R

```
# Einlesen der Daten
alter<-c(17,18,22,22,27,28,29,33,34,35,38,41,54)
geschlecht<-c(1,1,2,2,2,2,2,1,1,1,2,2,2)
grad<-c(1,3,3,3,3,1,3,2,3,2,1,3,2)
aufenthalt<-c(0,22,12,8,26,1,4,5,20,8,1,12,7)
verkehrsmittel<-c(1,4,3,3,3,2,4,3,2,2,4,4,1)

# Zusammenführen der Daten in einen Data Frame
khstatistik <-
  data.frame(
    Alter = alter,
    Geschlecht = geschlecht,
    Grad = grad,
    Aufenthalt = aufenthalt,
    Verkehrsmittel = verkehrsmittel
  )


#  1) Säulendiagramme der Häufigkeitsverteilung der Merkmale Verkehrsmittel,
#     Geschlecht und Grad der Verletzung.
barplot(
  table(verkehrsmittel),
  main = "Häufigkeitsverteilung",
  xlab = "Verkehrsmittel",
  ylab = "Häufigkeiten",
  col = "aliceblue"
)
barplot(table(geschlecht),
        main = "Häufigkeitsverteilung",
        xlab = "Geschlecht",
        ylab = "Häufigkeiten",
        col = "aliceblue"
)
barplot(table(grad),
        main = "Häufigkeitsverteilung",
        xlab = "Grad der Verletzung",
        ylab = "Häufigkeiten",
        col = "aliceblue"
)

#  2)
min(grad)
max(grad)
quantile(grad,c(0.25,0.5,0.75))

#  3)
#  Boxplot für Aufenthaltsdauer
#  Benötigte Maßzahlen: 25%-, 50%- und 75%-Quantil
quantile(aufenthalt,c(0.25,0.5,0.75))
boxplot(aufenthalt,
        main = "Boxplot Aufenthaltsdauer",
        ylab = "Dauer in Tagen",
        col = "aliceblue")

#  Boxplot für Alter
#  Benötigte Maßzahlen: 25%-, 50%- und 75%-Quantil
quantile(alter,c(0.25,0.5,0.75))
boxplot(alter,
        main = "Boxplot Alter",
        ylab = "Alter",
        col = "aliceblue")
```

Lösung mit Excel

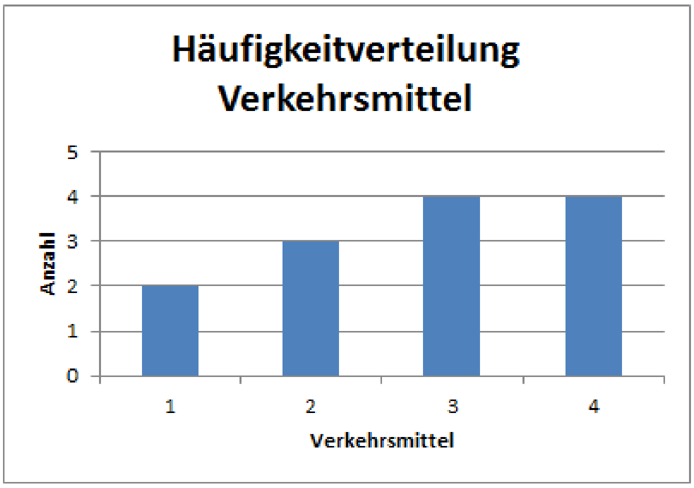
`WWMS_QBX_17_Krankenhaus.xlsx` (14 KB)

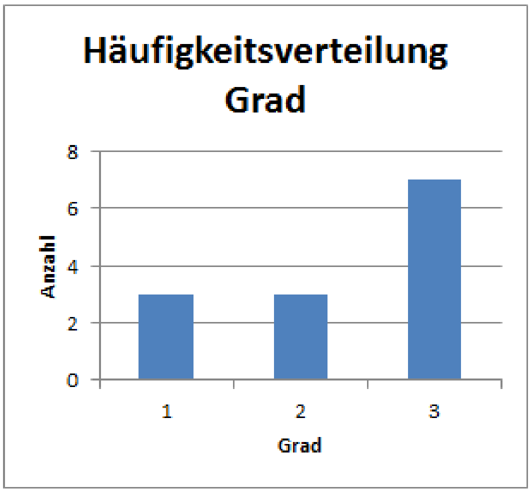
Aufgabe 1*: Stellen Sie jeweils die Häufigkeitsverteilungen der Verkehrsmittel und des Grades der Verletzungen im Säulendiagramm dar!*

Um die Häufigkeitsverteilungen des Verkehrsmittel und des Grades der Verletzungen im Säulendiagramm darzustellen, gehen wir vor, wie wir es in LE03 gelernt haben und verwenden die Funktion `HÄUFIGKEIT`. So erhalten wir folgende Häufigkeitsverteilungen:

Diagramm 2

	A	B	C
19			
20	Häufigkeitsverteilung Verkehrsmittel:		
21	Verkehrsmit	H	
22	1	2	
23	2	3	
24	3	4	
25	4	4	
26			
27	Häufigkeitsverteilung Grad:		
28	Grad	H	
29	1	3	
30	2	3	
31	3	7	





Aufgabe 2*: Bestimmen Sie Minimum, Maximum, Quartile und Median beim Grad der Verletzung.*

Zur Bestimmung von Minimum, Maximum, den Quartilen und dem Median werden die Funktionen `MIN`, `MAX`, `QUANTILE.INKL` und `MEDIAN` verwendet. Die Aufrufe sehen so aus:
•Minimum: `=MIN(D2:D14)`
•Maximum: `=MAX(D2:D14)`
•1. Quartil: `=QUANTIL.INKL(D2:D14;0.25)`
•3. Quartil: `=QUANTIL.INKL(D2:D14;0.75)`
•Median: `=MEDIAN(D2:D14)`

Genauer zu sehen ist das in folgenden Abbildungen:

B38 | =MIN(D2:D14)

	A	B	C	D	E
38	Minimum:	1			
39	1. Quartil:	2			
40	Median:	3			
41	3. Quartil:	3			
42	Maximum	3			
43					

B39 | =QUANTIL.INKL(D2:D14;0,25)

	A	B	C	D	E	F
38	Minimum:	1				
39	1. Quartil:	2				
40	Median:	3				
41	3. Quartil:	3				
42	Maximum	3				
43						

Dabei stehen in den Zellen D2:D14 die beobachteten Grade der Verletzungen. Für die Funktion QUANTIL muss noch angegeben werden, welches Quantil berechnet werden soll. Hier also 0,25 und 0,75. Alternativ können Quantile mit der Funktion `QUARTILE.INKL` berechnet werden. Hier würde dann nicht 0,25 und 0,75 stehen, sondern 1 und 3 (für das 1. und 3. Quartil). Entsprechend kann der Median mit `QUANTIL.INKL(D2:D14;0.5)` oder mit `QUARTILE.INKL(D2:D14;2)` berechnet werden. Die Ergebnisse in Excel sehen dann so aus:

Minimum:	1
1. Quartil:	2
Median:	3
3. Quartil:	3
Maximum	3

***Aufgabe 3**: Lassen Sie jeweils einen Boxplot erstellen zu den Variablen Dauer und Alter! Welche Maßzahlen müsste man berechnen, wollte man diesen Boxplot von Hand zeichnen? Wie lauten diese Maßzahlen konkret?*

Excel biete keine Möglichkeit, Boxplots zu erstellen. Sollen die Boxplots von Hand gezeichnet werden, so werden die Maßzahlen Minimum, 1. Quartil, Median, 3. Quartil und Maximum benötigt.
Für die Variablen Dauer und Alter lauten diese Maßzahlen konkret:

	Dauer	Alter
Minimum:	0	17
1. Quartil:	4	22
Median:	8	29
3. Quartil:	12	35
Maximum	26	54

Lösung für Übung QBX-17c

Wasserstände

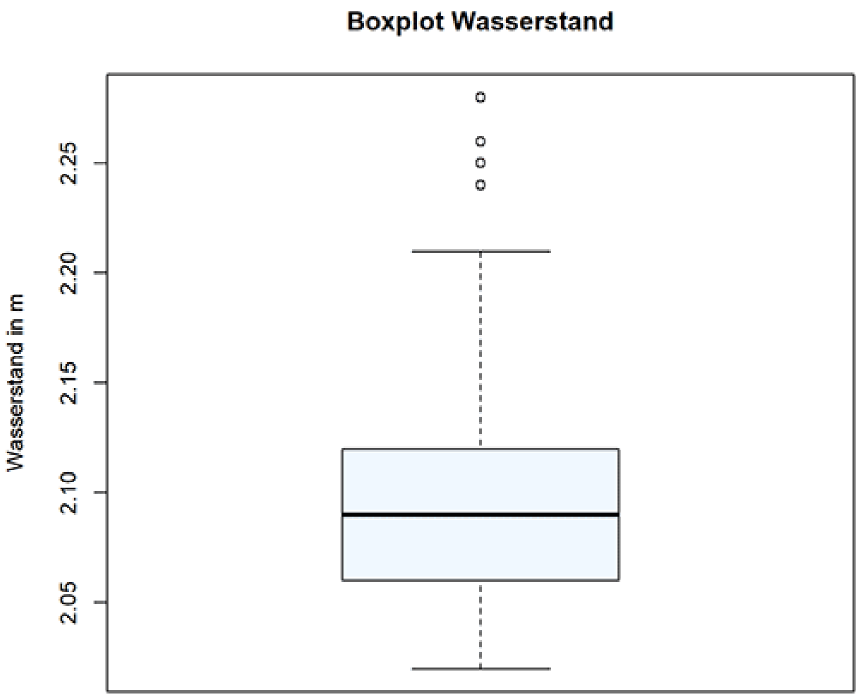
1. Die Häufigkeitstabelle lässt vermuten dass der Median bei etwa 2,08 Metern liegt.

i	Wasserstände	Häufigkeit
1	2,02	1
2	2,03	2
3	2,04	2
4	2,05	2
5	2,06	3
6	2,07	1
7	2,08	6
8	2,09	2
9	2,1	3
10	2,11	2
11	2,12	3
12	2,17	2
13	2,21	1
14	2,24	1
15	2,25	1
16	2,26	1
17	2,28	2

2. Die Quantile sagen aus, dass p Prozent der Werte kleiner oder gleich dem Wert X_p sind.

$x_{0,1}$	$x_{0,25}$	$\tilde{x}_{0,5}$	$x_{0,75}$	$x_{0,9}$
2,040	2,060	2,090	2,120	2,246

3. Im Boxplot werden das 25 %-, das 50 %- und das 75 %-Quantil dargestellt (50 %-Quantil = Median).



Lösung mit R

wasserstaende_loesung.R

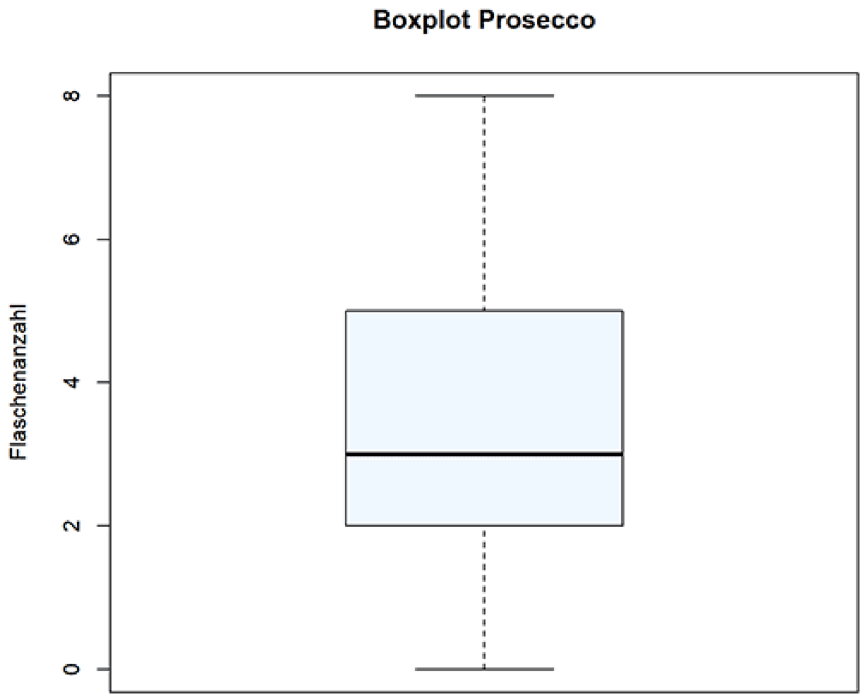
```
#  Einlesen der Wasserstände
stand<-c(2.08,2.05,2.1,2.08,2.07,2.08,2.09,2.1,2.08,2.11,2.06,2.12,
         2.11,2.08,2.17,2.12,2.06,2.17,2.12,2.05,2.21,2.09,2.02,
         2.24,2.06,2.03,2.26,2.04,2.03,2.28,2.04,2.08,2.28,2.1,2.25)

#  1)
table(stand)

#  2)
quantile(stand,c(0.1,0.25,0.5,0.75,0.9))

#  3)
boxplot(stand,
        main = "Boxplot Wasserstand",
        ylab = "Wasserstand in m",
        col = "aliceblue")
```

Lösung für Übung QBX-17d

Prosecco

1.

$x_{0,25}$	$\tilde{x}_{0,5}$	$x_{0,75}$
2	3	5



2. Wichtig sind die zentralen 50 Prozent der Daten – diese liegen zwischen 2 und 5 verkauften Flaschen sowie der Median der bei 3 Flaschen liegt. An diesen Werten lässt sich die Verkaufsstruktur des Betriebs gut ablesen.

Lösung mit R

prosecco_loesung.R

```
#  Einlesen der Daten (verkaufte Flaschen)
prosecco <- c(5,3,0,2,0,0,2,3,2,5,7,8,5,2,3,3,3,2,8,3,5,7,6,1)

#  Quartile bestimmen und Boxplot erzeugen
quantile(prosecco,c(0.25,0.5,0.75))
boxplot(prosecco,
        main = "Boxplot Prosecco",
        ylab = "Flaschenanzahl",
        col = "aliceblue")
```

Lösung für Übung QBX-17e

Störfälle

1. Störfalle 2021

$x_{0,25}$	$\tilde{x}_{0,5}$	$x_{0,75}$
1,75	2,5	4

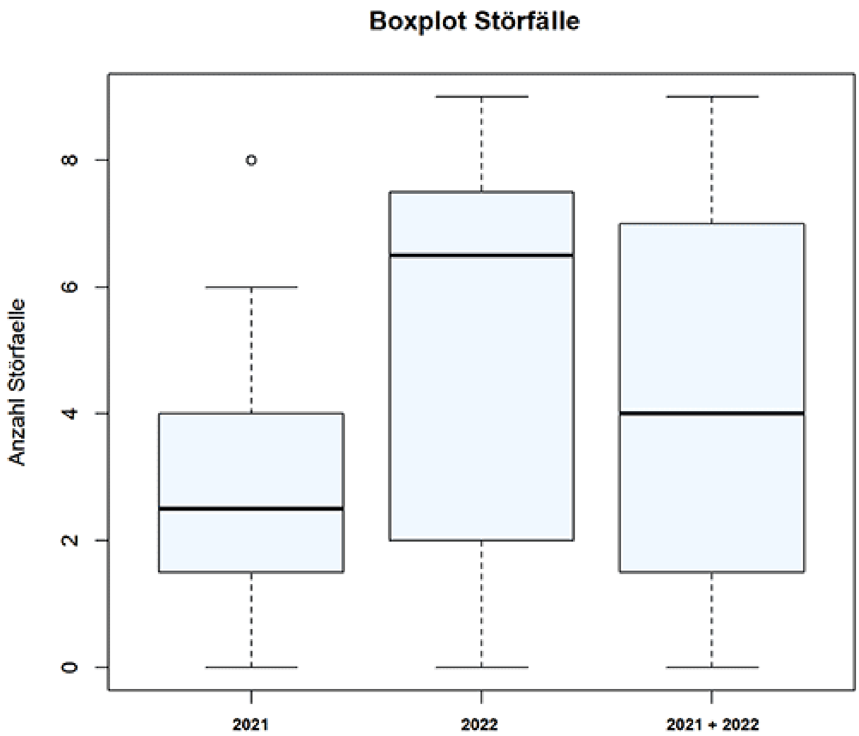
Störfälle 2022

$x_{0,25}$	$\tilde{x}_{0,5}$	$x_{0,75}$
2,5	6,5	7,25

Störfälle 2021 und 2022

$x_{0,25}$	$\tilde{x}_{0,5}$	$x_{0,75}$
1,75	4	7

2. Boxplots der Jahre 2021, 2022 und zusammen.



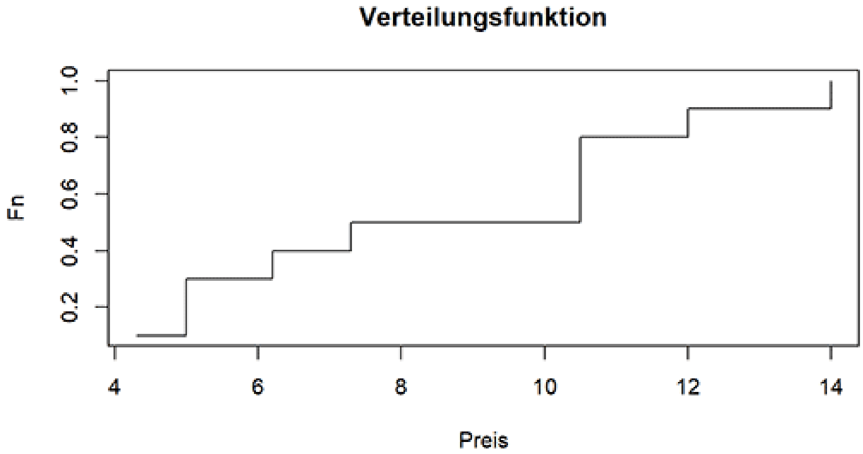
Lösung mit R

stoerfaelle_loesung.R

```
# Einlesen der Daten
stoerfaelle <- data.frame(Monat=c("Jan","Feb","Mae","Apr","Mai","Jun","Jul",
                                  "Aug","Sep","Okt","Nov","Dez"),
                          Jahr_2021=c(0,4,1,4,2,8,2,4,3,2,1,6),
                          Jahr_2022=c(7,1,6,1,9,7,8,8,4,3,7,0)
                          )

#  1) Quantile für 2021 und 2022 bestimmen
quantile(stoerfaelle$Jahr_2021,c(0.25,0.5,0.75))
quantile(stoerfaelle$Jahr_2022,c(0.25,0.5,0.75))

#  Gesamt (2021 und 2022 zusammen)
quantile(c(stoerfaelle$Jahr_2021, stoerfaelle$Jahr_2022),
         c(0.25, 0.5, 0.75))

#  2)
boxplot(
  stoerfaelle$Jahr_2021,
  stoerfaelle$Jahr_2022,
  c(stoerfaelle$Jahr_2021, stoerfaelle$Jahr_2022),
  main = "Boxplot Störfälle",
  ylab = "Anzahl Störfälle",
  xlab = "2021, 2022 und zusammen",
  col = "aliceblue"
)
```

Lösung für Übung QBX-17f

Angebote für Grundstücke

1. 

2. Der Wert der Verteilungsfunktion an der Stelle 13 ist 0,9. Das bedeutet: mindestens 90 % der Angebote sind niedriger oder gleich 13.000 Euro.
3. Mit Hilfe der Werte der Verteilungsfunktion lassen sich folgende Einschränkungen für den Median treffen: 7.300 $\leq$ Median $<$ 10.500
4. Aus der Grafik lässt sich ablesen: Das 20 %-Quantil liegt bei 5.000 Euro. Das bedeutet: mindestens 20 % der Angebote sind niedriger oder gleich 5.000 Euro.

Lösung mit R

grundstuecke_loesung.R

```
#  Einlesen der Angebotspreise in Tausend Euro
angebote<-c(10.5,5,4.3,12,14,10.5,6.2,5,7.3,10.5)

#  1)
abs.hkeit<-table(angebote)
rel.hkeit<-abs.hkeit/length(angebote)
fn<-cumsum(rel.hkeit)
plot(
  sort(unique(angebote)),
  fn,
  type = "n",
  xlab = "Preis",
  ylab = "Fn",
  main = "Verteilungsfunktion"
)
lines(sort(unique(angebote)),fn,type="s")
```

Lösung für Übung QBX-18

1. Wir bestimmen zuerst Q_1 (unteres Quartil): Es gilt $n \cdot p = 55 \cdot 0{,}25 = 13{,}75$ also ist das Ergebnis nicht ganzzahlig. Laut der Definition
   $Q_1 = x_{([n \cdot p + 1])} = x_{([55 \cdot 0{,}25 + 1])} = x_{([14{,}75])} = x_{(14)}$, also ist der Wert von
   $x_{(14)} = 165 = Q_1$
   Nach dem gleichen Prinzip berechnen wir das obere Quartil, Q_3 und erhalten als Ergebnis
   $x_{(42)} = 183 = Q_3$.
2. Die Quartilsmitte ist $\frac{1}{2}(Q_1 + Q_3) = 174$

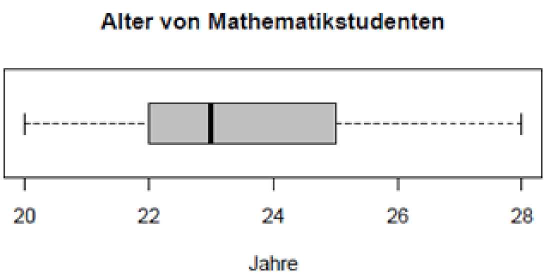
   und der Median beträgt $\frac{(n+1)}{2} = \frac{(55+1)}{2} = 172$
3. 50 Prozent der Studierenden sind kleiner als 172 cm, die restlichen 50 Prozent der Studierenden sind größer als 172 cm - diese Aussage können wir mithilfe des Medians treffen. Das untere Quartil entspricht 165 cm. Damit sind 25 Prozent der Studierenden kleiner als 165 cm. Das obere Quartil mit dem Wert 183 cm kann folglich dahingehend interpretiert werden, dass 75 Prozent der Studierenden kleiner sind als 185 cm.

Lösung für Übung QBX-19

1.
   1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6
2. Der Median ist gleich 2, Q_1 = 2 und Q_3 = 4.
3. In 50 Prozent der Spiele gab es höchstens 2 Tore pro Spiel und in den restlichen 50 Prozent gab es mindestens 3 Tore pro Spiel. In 75 Prozent der Spiele wurden weniger als 4 Tore geschossen und in 25 Prozent der Spiele weniger als 2.

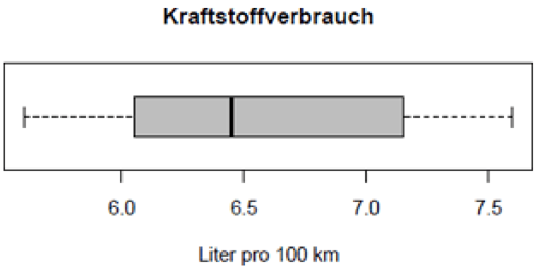
Lösung für Übung QBX-20

Der Median ist gleich 23, Q_1 = 22, Q_3 = 25, Minimum ist gleich 20 und Maximum ist gleich 28.



Lösung für Übung QBX-21

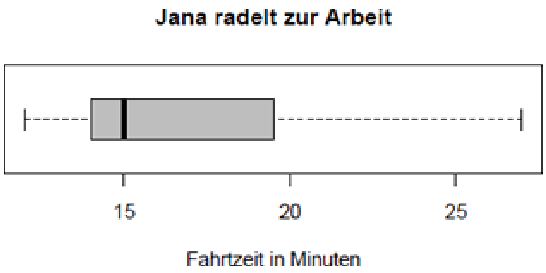
Der Median ist gleich 6,45, Q_1 = 6,125, Q_3 = 7,125, Minimum ist gleich 5,6 und Maximum ist gleich 7,6.



Das Ergebnis wurde mit Hilfe der R-Funktion **`quantile`** bestimmt. Kleine Abweichungen von der Skriptformel sind durch eine leicht unterschiedliche Interpolationsmethode erklärbar. Das stört uns nicht wirklich :-)

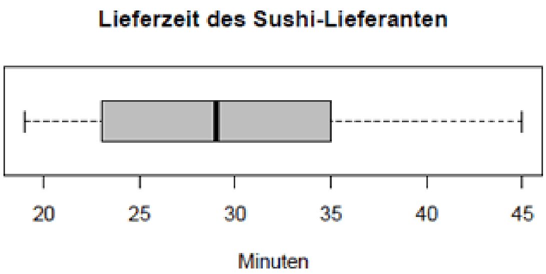
Lösung für Übung QBX-22

Der Median ist gleich 15, Q_1 = 14, Q_3 = 19,5, Minimum ist gleich 12 und Maximum ist gleich 27.



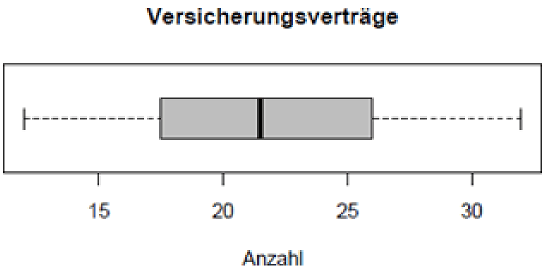
Lösung für Übung QBX-23

Der Median ist gleich 29, Q_1 = 23, Q_3 = 35, Minimum ist gleich 19 und Maximum ist gleich 45.



Lösung für Übung QBX-24

Der Median ist gleich 21,5, Q_1 = 17,75, Q_3 = 26, Minimum ist gleich 12 und Maximum ist gleich 32.



Das Ergebnis wurde mit Hilfe der R-Funktion `quantile` bestimmt. Kleine Abweichungen von der Skriptformel sind durch eine leicht unterschiedliche Interpolationsmethode erklärbar. Das stört uns nicht wirklich :-)

Lösung für Übung QBX-25

a)

i	$x_{i-1} < X \leq x_i$	n_i	h_i	$F_n(x_i)$
1	0 < X ≤ 500	10	0,1	0,1
2	500 < X ≤ 1000	18	0,18	0,28
3	1000 < X ≤ 1500	30	0,3	0,58
4	1500 < X ≤ 2000	32	0,32	0,9
5	2000 < X ≤ 3000	7	0,07	0,97
6	3000 < X ≤ 6000	3	0,03	1

b) $F(1740) = 0{,}58 + \frac{1740-1500}{500} \cdot 0{,}32 = 0{,}58 + 0{,}15 = 0{,}73$, d. h. mindestens 73 % der Befragten bekommen weniger, oder 27 % der Befragten bekommen mehr als 1.740,- Euro brutto.

c) Als erstes suchen wir die Klasse in welche dieses 0.7-Quantil fällt, in unserem Fall ist es die vierte Klasse. Nun haben wir alle notwendigen Angaben und können die entsprechenden Werte in die folgende Formel einsetzen:

$$x_p = x_{m-1} + \frac{(p - F(x_m - 1))\Delta m}{h_m}$$

$$\Rightarrow x_{0.7} = 1500 + \frac{(0{,}7 - 0{,}58)500}{0{,}32} = 1500 + 187{,}5 = 1687{,}5$$

d) Das heißt, 70 % der befragten Haushalte haben einen monatlichen Bruttoverdienst von nicht mehr als 1.687,50 €.

Lösung für Übung QBX-26

a)

i	$x_{i-1} < X \leq x_i$	n_i	h_i	$F_n(x_i)$
1	0 < X ≤ 4000	7	0,07	0,07
2	4000 < X ≤ 6000	18	0,18	0,25
3	6000 < X ≤ 7000	21	0,21	0,46
4	7000 < X ≤ 9000	41	0,41	0,87
5	9000 < X ≤ 10000	10	0,1	0,97
6	10000 < X ≤ 12000	3	0,03	1

b) 75 %

c) 87 %

Lösung für Übung QBX-27

a)

Alter	***Anzahl der Geborenen***	h_i	$F_n(x_i)$
[15; 20)	70	0,07	0,07
[20; 25)	289	0,289	0,359
[25; 30)	369	0,369	0,728
[30; 35)	178	0,178	0,906
[35; 40)	83	0,083	0,989
[40; 45)	11	0,011	1

b) Q_1 = 23,11 und Q_3 = 30,62

c) IQR = Q_3 - Q_1 = 30,62 − 23,11 = 7,51

d) 50 % der Mütter sind im Alter von 23 bis 31 Jahre.