VON NULL AUF MATHE

Grundlagen der Statistik: Daten, Diagramme, Kennzahlen und Zusammenhänge

Blick ins Buch

CHRISTOPH GERBER, PETER JANKOVICS

- Daten erfassen un
- Häufigkeit und Diag
- Klassierte Daten
- Summenhäufigkeiten
- Mittelwert, Modus, Median
- Standardabw. und Varianz
- Boxplots
- Bivariate Daten
- Streudiagramm
- Korrelation und Regression

compendio
Bildungsmedien

VON NULL AUF MATHE

Grundlagen der Statistik: Daten, Diagramme, Kennzahlen und Zusammenhänge

CHRISTOPH GERBER, PETER JANKOVICS

compendio
Bildungsmedien

VON NULL AUF MATHE
Grundlagen der Statistik: Daten, Diagramme, Kennzahlen und Zusammenhänge
Christoph Gerber, Peter Jankovics

Satz: Reemers Publishing Services GmbH, Krefeld
Grafisches Konzept: icona basel gmbh
Coverbild: © Stefan Schmitz, Hamburg
Druck: Edubook AG, Merenschwand
Redaktion und didaktische Bearbeitung: Peter Jankovics

Printausgabe
ISBN: 978-3-7155-0136-9
Artikelnummer: 19536
Auflage: 1. Auflage 2026
Ausgabe: 01N26
Sprache: DE
XMA 509

E-Book-Ausgabe
ISBN: 978-3-7155-0135-2
Artikelnummer: E-19537
Auflage: 1. Auflage 2026
Ausgabe: 01N26
Sprache: DE
XMA 509

Compendio Bildungsmedien AG
Neunbrunnenstrasse 50
CH-8050 Zürich
Tel. +41 44 368 21 11
info@compendio.ch
www.compendio.ch

Die Printausgabe dieses Buchs ist klimaneutral in der Schweiz gedruckt worden. Die Druckerei Edubook AG hat sich einer Klimaprüfung unterzogen, die primär die Vermeidung und Reduzierung des CO_2-Ausstosses verfolgt. Verbleibende Emissionen kompensiert das Unternehmen durch den Erwerb von CO_2-Zertifikaten eines Schweizer Klimaschutzprojekts. Mehr zum Umweltbekenntnis von Compendio Bildungsmedien finden Sie unter: www.compendio.ch/Umwelt

Inhaltsverzeichnis

Vorwissen und Lernziele

Vorwissen

Erforderliches Vorwissen für **«Grundlagen der Statistik: Daten, Diagramme, Kennzahlen und Zusammenhänge»:**

- Grundrechenarten
- Koordinatensystem

Lernziele

Nach der Bearbeitung dieses Lehrmittels …

- kennen Sie die Grundbegriffe einer statistischen Erhebung.
- können Sie zwischen quantitativen und qualitativen Merkmalen unterscheiden.
- können Sie den Rang, die absolute und die relative Häufigkeit eines Stichprobenwerts bestimmen.
- können Sie die Häufigkeitsverteilung einer Stichprobe in einer Häufigkeitstabelle darstellen.
- kennen Sie verschiedene Diagramme, um absolute und relative Häufigkeitsverteilungen grafisch darzustellen.
- erkennen Sie Eigenschaften von Häufigkeitsverteilungen anhand ihrer Diagramme.
- können Sie Stichprobenwerte in Klassen einteilen.
- können Sie für klassierte Daten ein Histogramm erstellen.
- wissen Sie, wie man die Summenhäufigkeiten ermittelt, und können diese grafisch darstellen und interpretieren.
- können Sie die Lagemasse Mittelwert, Modus und Median einer Stichprobe ermitteln.
- wissen Sie, was ein Ausreisser ist, und können einschätzen, wie dieser die Lagemasse beeinflusst.
- können Sie die Streumasse Spannweite, mittlere absolute Abweichung, Varianz und Standardabweichung einer Stichprobe ermitteln.
- kennen Sie eine praktische Faustregel für die Interpretation der Standardabweichung.
- können Sie den Boxplot einer Stichprobe zeichnen.
- können Sie die Punktwolke bivarianter Daten erstellen.
- erkennen Sie lineare Zusammenhänge bivarianter Daten.
- können Sie anhand einer Trendlinie Prognosen abgeben.

1.2 Sammeln von Daten

Bei der statistischen Erhebung in der Einstiegsaufgabe wurden Daten zu einer bestimmten Eigenschaft, nämlich dem Alter, gesammelt. In der Statistik nennt man diese zu untersuchende Eigenschaft **Merkmal.** Ein Objekt oder eine Person, die diese Eigenschaft besitzt, nennt man **Merkmalsträger.** In unserem Fall ist das ein Einwohner des Dorfs. Alle Merkmalsträger zusammen sind die **Grundgesamtheit.** Das sind alle 1 500 Einwohnern des Dorfs. Bei einer statistischen Erhebung kann es vorkommen, dass der Umfang der Grundgesamtheit zu gross ist, um die Daten von allen Merkmalsträgern zu bekommen. Deshalb wählt man eine Teilmenge der Grundgesamtheit, eine sog. **Stichprobe,** aus. In der Einstiegsaufgabe waren das 50 Einwohner. Die Stichprobe hatte also einen **Stichprobenumfang** von 50. Die Daten, die man erhält, hier das jeweilige Alter, nennt man **Stichprobenwerte** oder **Merkmalsausprägungen.** Diese Stichprobenwerte werden in der Reihenfolge ihrer Erfassung in einer **Urliste** gesammelt.

Merken Sie sich die Begriffe:

Grundbegriffe einer statistischen Erhebung

- **Merkmal:** Das Merkmal ist die Eigenschaft, die im Rahmen einer statistischen Erhebung untersucht wird.
- **Merkmalsträger:** Die Merkmalsträger sind die Personen oder Objekte, deren Merkmal untersucht wird.
- **Grundgesamtheit:** Die Grundgesamtheit ist die Menge aller zu untersuchenden Merkmalsträger.
- **Stichprobe:** Eine Stichprobe ist eine Teilmenge der Grundgesamtheit.
- **Stichprobenumfang:** Der Stichprobenumfang gibt die Anzahl der Merkmalsträger der Stichprobe an.
- **Stichprobenwerte / Merkmalsausprägungen:** Die Stichprobenwerte oder Merkmalsausprägungen sind die Werte, die das Merkmal annimmt. Das sind die Daten, die erhoben werden.
- **Urliste:** Die Urliste ist die ursprüngliche unbearbeitete Aufzeichnung der Stichprobenwerte.

Allgemein geht man bei statistischen Erhebungen mit grossen Datenmengen so vor:

Durchführung einer statistischen Erhebung

1. Gewünschtes **Merkmal** festlegen, das bei allen Personen oder Objekten einer **Grundgesamtheit** vorhanden ist.
2. Aus der Grundgesamtheit eine möglichst repräsentative **Stichprobe** mit einem bestimmten **Stichprobenumfang** wählen.
3. **Stichprobenwerte** ermitteln und in einer **Urliste** sammeln.

Von welcher Art die erhobenen Daten (Stichprobenwerte) sind, hängt vom jeweiligen Merkmal ab, das untersucht wird. Man unterscheidet zwischen Merkmalen wie dem Alter, bei denen nur Zahlenwerte erfasst werden, und solchen wie der Augenfarbe, bei denen keine Zahlenwerte erfasst werden.

Arten von Merkmalen

Ein **Merkmal** heisst …

- **quantitatives Merkmal,** wenn die erhobenen Daten nur aus Zahlenwerten bestehen.
- **qualitatives Merkmal,** wenn die erhobenen Daten nicht aus Zahlenwerten bestehen.

Ein **quantitatives Merkmal** heisst zudem …

- **diskret,** wenn die erhobenen Daten nur aus isolierten Zahlenwerten bestehen.
- **stetig,** wenn die erhobenen Daten prinzipiell alle Zahlenwerte eines Intervalls annehmen können.

Beispiele:

Grundgesamtheit	Merkmal	Daten / Stichprobenwerte	Art des Merkmals
Einwohner einer Stadt	Alter (in Jahren)	…; 31; 32; 33; …	quantitatives, diskretes Merkmal
Schüler einer Klasse	Lieblingsfarbe	grün, blau, rot, …	qualitatives Merkmal
Schüler einer Klasse	Körpergrösse (in m)	1.7; 1.78; 1.789; …	quantitatives, stetiges Merkmal
Städte eines Lands	Anzahl der Einwohner	11 251; 213 624; …	quantitatives, diskretes Merkmal
Tiere eines Zoos	Tierart	Zebra, Löwe, Giraffe …	qualitatives Merkmal

Aufgabe 2

Kreuzen Sie alle statistischen Erhebungen an, bei denen das Merkmal quantitativ ist.

☐	In einem Wald werden die Höhen der Bäume gemessen.
☐	In einer Schule wird die Anzahl der Tische in den Klassenräumen gezählt.
☐	Auf einem Bauernhof wird das Geschlecht der Tiere ermittelt.
☐	Die Schüler einer Klasse werden nach ihrem Lieblingseis gefragt.
☐	Die Schüler einer Klasse geben an, wie viel Taschengeld sie bekommen.
☐	In einem Wald werden die Baumarten ermittelt.
☐	Ein Umfrageinstitut befragt die Wähler nach ihrem möglichen Wahlverhalten.
☐	Ein Würfel wird mehrfach geworfen. Die gewürfelten Augenzahlen werden erfasst.

2.2 Strichliste und Häufigkeiten

Die geordnete Liste von Alice ist passend, um den Rang eines Stichprobenwerts zu ermitteln. Will man aber zählen, wie oft ein Wert in der Stichprobe vorkommt, so ist die **Strichliste** von Bob geeigneter. Die Anzahl der Striche bei einem Stichprobenwert ist dann jeweils dessen **absolute Häufigkeit.** Dividiert man die absolute Häufigkeit durch den Stichprobenumfang, so erhält man die **relative Häufigkeit:**

$$\text{relative Häufigkeit} = \frac{\text{absolute Häufigkeit}}{\text{Stichprobenumfang}}$$

Die beiden Soziologen müssen also die absolute Häufigkeiten jeweils durch 50 teilen, um die zugehörigen relativen Häufigkeiten zu erhalten. Diese werden auch häufig in Prozent angegeben:

Alter	Striche	Abs. H.	Rel. H.
1	|	1	0.02 = 2%
4	|||	3	0.06 = 6%
8	||	2	0.04 = 4%
11	|	1	0.02 = 2%
13	||	2	0.04 = 4%
14	|	1	0.02 = 2%
17	||||||	6	0.12 = 12%
22	|	1	0.02 = 2%
26	|	1	0.02 = 2%
27	|	1	0.02 = 2%
31	|||||	5	0.1 = 10%
32	|	1	0.02 = 2%
35	||	2	0.04 = 4%
36	|	1	0.02 = 2%
41	|	1	0.02 = 2%

Alter	Striche	Abs. H.	Rel. H.
44	||	2	0.04 = 4%
46	|	1	0.02 = 2%
49	|	1	0.02 = 2%
50	|	1	0.02 = 2%
54	||	2	0.04 = 4%
55	|	1	0.02 = 2%
56	|	1	0.02 = 2%
59	||||	4	0.08 = 8%
63	|	1	0.02 = 2%
64	||	2	0.04 = 4%
73	|	1	0.02 = 2%
79	|	1	0.02 = 2%
83	|	1	0.02 = 2%
84	|	1	0.02 = 2%
96	|	1	0.02 = 2%

In dieser **Häufigkeitstabelle** gibt es 30 verschiedene Stichprobenwerte (Alter): $x_1 = 1$, $x_2 = 4$, $x_3 = 8$, …, $x_{29} = 84$, $x_{30} = 96$. Und für **jeden** Stichprobenwert x_i können wir seine absolute Häufigkeit ablesen bzw. zuordnen. Diese Zuordnung nennt man **absolute Häufigkeitsverteilung** H. Es gilt z. B. $H(4) = 3$. Das bedeutet, dass der Stichprobenwert 4 genau dreimal vorkommt. In einer Häufigkeitsverteilung ergibt die Summe aller absoluten Häufigkeiten immer den Stichprobenumfang n, in unserem Fall $n = 50$.

Genauso können wir für **jeden** Stichprobenwert x_i seine relative Häufigkeit berechnen. Diese Zuordnung nennt man **relative Häufigkeitsverteilung** h. Es gilt z. B. $h(17) = 0.12$. Das heisst, 12% der Stichproben haben den Wert 17. In einer Häufigkeitsverteilung ergibt die Summe aller relativen Häufigkeiten immer 1 bzw. 100%.

Wir fassen die elementaren Begriffe zusammen.

Diagramme für absolute und relative Häufigkeitsverteilungen

Die **absolute** und die **relative Häufigkeitsverteilung** einer Stichprobe können durch folgende **Diagramme** grafisch dargestellt werden:

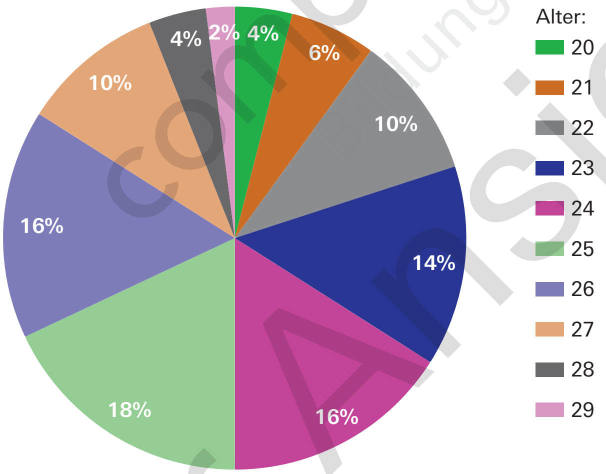
- Stabdiagramm
- Säulendiagramm
- Liniendiagramm
- Balkendiagramm

Hierfür zeichnet man zwei senkrecht aufeinanderstehende Achsen. Beim Stab-, Säulen- und Liniendiagramm wird die horizontale Achse mit den Stichprobenwerten beschriftet und die vertikale Achse mit den absoluten bzw. relativen Häufigkeiten. Beim Balkendiagramm ist es umgekehrt.

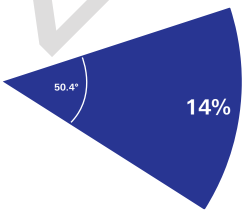
Eine andere Möglichkeit der Darstellung einer **relativen** Häufigkeitsverteilung ist das **Kreis-** oder auch **Kuchendiagramm.**

Kreisdiagramm

Im Kreisdiagramm wird jedem Stichprobenwert ein Kreissektor zugeteilt. Die Flächeninhalte der einzelnen Kreissektoren entsprechen dabei dem Anteil der zugehörigen relativen Häufigkeiten:



Man erhält den Zentriwinkel α_i des Kreissektors eines Stichprobenwerts x_i mithilfe seiner relativen Häufigkeit $h(x_i)$ durch die Formel $\alpha_i = h(x_i) \cdot 360°$. Beispielsweise ergibt sich der Zentriwinkel α des Kreissektors für das Alter 23 durch die relative Häufigkeit $h(23) = 0.14$ mit $\alpha = 0.14 \cdot 360° = 50.4°$:



Dabei wird also der 360°-Vollwinkel des Kreises anteilmässig auf die Stichprobenwerte verteilt: 14% von 360° sind 50.4°.

Aufgabe 20

Bei der Wahl für die Studentenvertretung bekam Christoph 37 Stimmen, Sonja 112 Stimmen, Remy 219 Stimmen, Selina 47 Stimmen und Maurice 9 Stimmen. 43 Studenten enthielten sich.

a) Legen Sie eine Häufigkeitsverteilung der erhaltenen Stimmen in Tabellenform an. Benutzen Sie dabei für die relativen Häufigkeiten Prozentangaben und runden Sie diese auf ganze Zahlen.

b) Stellen Sie die relative Häufigkeitsverteilung der Stichprobe mit einem Säulendiagramm dar.

c) Stellen Sie die relative Häufigkeitsverteilung der Stichprobe mit einem Kreisdiagramm dar.

d) Was ist der Zentriwinkel bei den 10% von Selina?

3.2 Eigenschaften von Häufigkeitsverteilungen aus Diagrammen lesen

Einige Eigenschaften von Häufigkeitsverteilungen lassen sich einfach an ihren Diagrammen erkennen. So spricht man von einer **symmetrischen** oder **schiefen** Häufigkeitsverteilung, je nachdem, ob das zugehörige Diagramm eine Symmetrieachse hat oder nicht:

symmetrisch — linksschief — rechtsschief

Das Liniendiagramm einer symmetrischen Verteilung besitzt eine Symmetrieachse und es ist dadurch gekennzeichnet, dass die Stichprobenwerte, die gleich weit vom zentralen maximalen Häufigkeitswert entfernt sind, dieselbe Häufigkeit besitzen. Bei einer **linksschiefen** Verteilung ist die Kurve links vom maximalen Häufigkeitswert länger als rechts davon. Bei **rechtsschiefen** Verteilungen ist es umgekehrt.

In einem Stab-, Säulen- bzw. Liniendiagramm gibt es oft einen längsten Stab, eine längste Säule bzw. einen Hochpunkt. Dieser maximale Häufigkeitswert liegt naturgemäss bei dem Stichprobenwert, der am häufigsten vorkommt. Diesen Stichprobenwert nennt man dann **Modus.** Die Anzahl der Modi entscheidet, ob eine Häufigkeitsverteilung als **unimodal, bimodal** oder **multimodal** bezeichnet wird. An den Diagrammen können wir diese Unterscheidung schnell erkennen:

unimodal bimodal multimodal

Das Liniendiagramm der unimodalen Verteilung hat einen Hochpunkt, das der bimodalen hat genau zwei Hochpunkte und das der multimodalen Verteilung hat mehrere Hochpunkte.

Aufgabe 21 Wahr oder falsch? Begründen Sie Ihre Entscheidung.

a) Bei einem achsensymmetrischen Liniendiagramm kommen alle Häufigkeitswerte paarweise vor.

b) Für die Diagramme einer Häufigkeitsverteilung gilt: Die Säulen im Säulendiagramm sind bei gleicher Skalierung genauso lang wie die Stäbe im Stabdiagramm.

c) Beim 100-maligen Werfen einer Münze (Merkmale K[opf] oder Z[ahl]) ist das Liniendiagramm eine Gerade.

d) Beim 100-maligen Werfen eines Würfels (Merkmale: Augenzahlen 1, 2, 3, 4, 5, 6) ist die Häufigkeitsverteilung unimodal.

e) Jede nichtsymmetrische Häufigkeitsverteilung ist schief.

f) Eine Häufigkeitsverteilung, bei der alle Stichprobenwerte dieselbe absolute Häufigkeit besitzen, hat keinen Modus.

g) Jede symmetrische Häufigkeitsverteilung ist unimodal.

h) Die Häufigkeitsverteilung der Stichprobe 3, 2, 1, 2, 3, 2, 1 ist symmetrisch.

4.3 Klassenintervalle mit verschiedenen Breiten

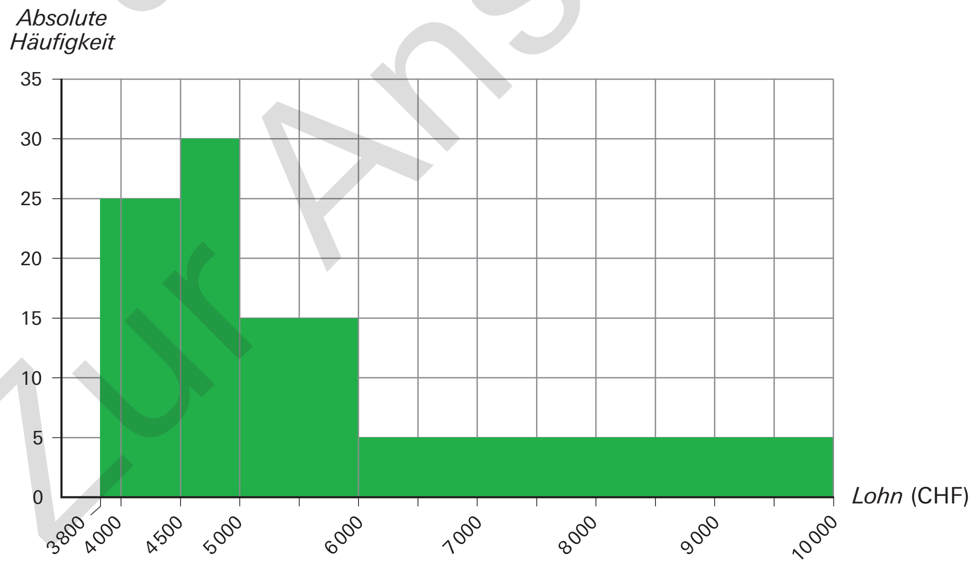
Die Gruppierung von Daten in Klassen kann verallgemeinert werden, indem man bei gewissen Datenstrukturen die Klassenbreiten unterschiedlich gross macht. Dies ist insbesondere dann angezeigt, wenn die Daten nicht gleichmässig über alle Merkmalsausprägungen verteilt sind. Wir betrachten dazu das folgende Beispiel.

In einem Betrieb mit 100 Angestellten werden monatliche Löhne von 3800 bis 10000 Franken ausbezahlt. Die Löhne verteilen sich wie folgt:

Lohn (CHF)	Klassenintervalle	Klassenbreiten	Abs. H.
von 3800 bis unter 4000	$I_1 = [3800; 4000[$	$b_1 = 200$	$H(I_1) = 25$
von 4000 bis unter 4500	$I_2 = [4000; 4500[$	$b_2 = 500$	$H(I_2) = 25$
von 4500 bis unter 5000	$I_3 = [4500; 5000[$	$b_3 = 500$	$H(I_3) = 30$
von 5000 bis unter 6000	$I_4 = [5000; 6000[$	$b_4 = 1000$	$H(I_4) = 15$
von 6000 bis 10000	$I_5 = [6000; 10000]$	$b_5 = 5000$	$H(I_5) = 5$

Da die Anzahl der höheren Löhne erwartungsgemäss kleiner ist als diejenige der niedrigeren, ist es sinnvoll, bei der Klassenbildung die höheren Lohnklassen breiter zu machen.

Wenn wir nun, wie bei konstanten Klassenbreiten, einfach über jedes Intervall ein Rechteck mit der Höhe der jeweiligen absoluten Klassenhäufigkeit zeichnen, entsteht folgendes Diagramm:



Dieses Diagramm ist irreführend. Es erweckt den Eindruck, dass die Häufigkeit der Klasse $I_1 = [3800; 4000[$ kleiner wäre als die Häufigkeit der Klasse $I_2 = [4000; 4500[$. Und auch das Rechteck der Klasse $I_5 = [6000; 10000]$ wirkt überproportional gross, man bekommt den Eindruck, es gäbe sehr viele Angestellten in der höchsten Lohnklasse. Das liegt daran, dass sich das menschliche Auge an der Grösse der **Flächen** der Rechtecke orientiert und nicht an deren Höhe.

Um diesem Umstand Rechnung zu tragen, müssen die Höhen der Rechtecke im Histogramm so gewählt werden, dass die **Flächeninhalte** der Rechtecke den Klassenhäufigkeiten entsprechen. Dies erreicht man z. B. für die Klasse $I_1 = [3800;\, 4000[$, indem man die absolute Klassenhäufigkeit $H(I_1) = 25$ durch die Klassenbreite $b_1 = 200$ dividiert (Normierung der Häufigkeiten):

$$f_1 = \frac{H(I_1)}{b_1} = \frac{25}{200} = 0.125$$

Diese Grösse f_1 entspricht der Rechteckshöhe über dem Intervall I_1 im Histogramm. Allgemein bekommt man die Klassenhöhen f_i für das Intervall I_i durch:

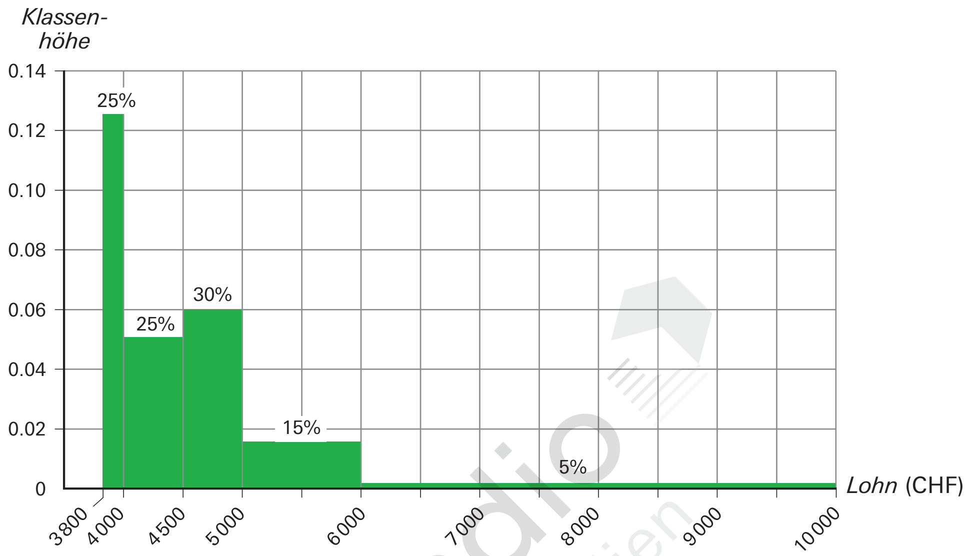
$$f_i = \frac{H(I_i)}{b_i}$$

Die Klassenhöhe f_i nennt man auch **Häufigkeitsdichte.**

Wir erweitern nun die obige Tabelle mit der Spalte «Klassenhöhen»:

Klassenintervalle	Klassenbreiten	Abs. H.	Klassenhöhen
$I_1 = [3800;\, 4000[$	$b_1 = 200$	$H(I_1) = 25$	$f_1 = \frac{25}{200} = 0.125$
$I_2 = [4000;\, 4500[$	$b_2 = 500$	$H(I_2) = 25$	$f_2 = \frac{25}{500} = 0.05$
$I_3 = [4500;\, 5000[$	$b_3 = 500$	$H(I_3) = 30$	$f_3 = \frac{30}{500} = 0.06$
$I_4 = [5000;\, 6000[$	$b_4 = 1000$	$H(I_4) = 15$	$f_4 = \frac{15}{1000} = 0.015$
$I_5 = [6000;\, 10000]$	$b_5 = 5000$	$H(I_5) = 5$	$f_5 = \frac{5}{4000} = 0.00125$

Das zugehörige Histogramm sieht jetzt so aus:



Wir haben noch die relativen Klassenhäufigkeiten in Prozent an die Rechtecke geschrieben. Diese entsprechen dann auch dem jeweiligen Anteil eines Rechtecks an der Gesamtfläche aller Rechtecke.

Histogramm bei verschiedenen Klassenbreiten

Das Anlegen eines **Histogramms** bei verschiedenen Klassenbreiten erfolgt in folgenden Schritten:

1. Klassen festlegen: Unterteilung der Stichprobenwerte in aneinandergrenzenden Intervallen $I_1, I_2, I_3, \ldots, I_m$ mit den Breiten $b_1, b_2, b_3, \ldots, b_m$.
2. Klassenhäufigkeiten bestimmen: Die absolute Häufigkeit $H(I_i)$ jeder Klasse bestimmen.
3. Klassenhöhen bestimmen: Für jede Klasse I_i die Klassenhöhe (Häufigkeitsdichte) f_i bestimmen, indem man die absolute Klassenhäufigkeit $H(I_i)$ durch die Klassenbreite b_i dividiert: $f_i = \dfrac{H(I_i)}{b_i}$.
4. Histogramm zeichnen: Über jedes Intervall I_i ein Rechteck mit der Breite b_i und der Höhe f_i zeichnen.

Aufgabe 35

Benutzen Sie in der vorherigen Tabelle die relativen Summenhäufigkeiten und beantworten Sie die folgenden Fragen. Wie viel Prozent der Befragten sind …

a) höchstens 35 Jahre alt?

b) mindestens 36 Jahre alt?

c) mindestens 35 Jahre alt?

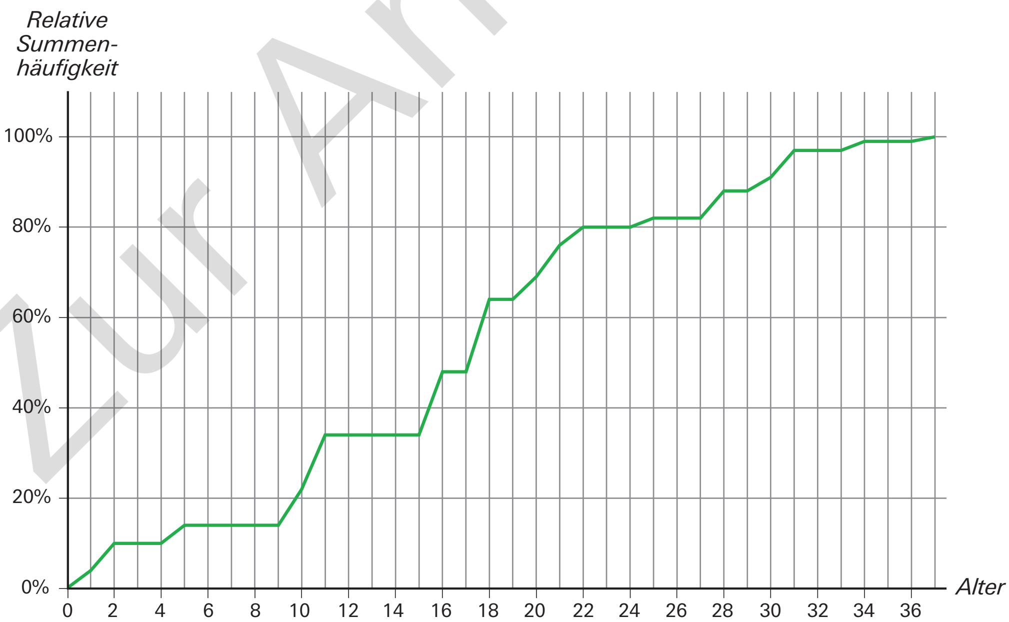
d) mindestens 22 Jahre alt?

e) mindestens 22, aber höchstens 30 Jahre alt?

f) mindestens 6, aber höchstens 34 Jahre alt?

5.2 Grafische Darstellung von Summenhäufigkeiten

Genauso wie die einzelnen relativen Häufigkeiten können wir auch die relativen Summenhäufigkeiten für die gesamte Stichprobe grafisch darstellen. Für die Stichprobe aus der Einstiegsaufgabe ergibt sich z. B. folgendes Liniendiagramm:



Die Kurve beginnt bei null und steigt bei jedem beobachteten Wert sprunghaft um die jeweilige relative Häufigkeit an. Zwischen den einzelnen Werten bleibt sie konstant, bis sie beim grössten Wert die Summenhäufigkeit 100% erreicht. So lässt sich direkt ablesen, welcher Anteil der Daten kleiner oder gleich einem bestimmten Wert ist.

Mittelwert einer Stichprobe

Gegeben sei eine Stichprobe mit den Stichprobenwerten $x_1, x_2, \ldots, x_n$ und dem Stichprobenumfang n.

Der **Mittelwert** $\bar{x}$ (auch **arithmetisches Mittel** oder **Durchschnitt**) der Stichprobe berechnet sich aus der Summe aller Stichprobenwerte geteilt durch den Stichprobenumfang:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Weniger Aufwand zur Berechnung des Mittelwerts hat man, wenn die Stichprobe als absolute oder relative Häufigkeitsverteilung vorliegt. Für die Befragung aus der Einstiegsaufgabe wäre das:

x_i	20	23	24	26	29	30	32	35	36	40
$H(x_i)$	2	3	5	7	8	9	8	5	2	1
$h(x_i)$	0.04	0.06	0.1	0.14	0.16	0.18	0.16	0.1	0.04	0.02

Dann braucht man nämlich nicht jeden Wert einzeln zu addieren, sondern kann ihn mit seiner absoluten Häufigkeit multiplizieren:

$$\bar{x} = \frac{2 \cdot 20 + 3 \cdot 23 + 5 \cdot 24 + 7 \cdot 26 + 8 \cdot 29 + 9 \cdot 30 + 8 \cdot 32 + 5 \cdot 35 + 2 \cdot 36 + 1 \cdot 40}{50} = 29.12$$

Der Mittelwert ist auch direkt mit den relativen Häufigkeiten berechenbar:

$$\begin{aligned}\bar{x} &= \frac{2}{50} \cdot 20 + \frac{3}{50} \cdot 23 + \frac{5}{50} \cdot 24 + \ldots + \frac{5}{50} \cdot 35 + \frac{2}{50} \cdot 36 + \frac{1}{50} \cdot 40 \\ &= 0.04 \cdot 20 + 0.06 \cdot 23 + 0.1 \cdot 24 + \ldots + 0.1 \cdot 35 + 0.04 \cdot 36 + 0.02 \cdot 40 = 29.12\end{aligned}$$

Mittelwert aus einer Häufigkeitsverteilung

Gegeben sei eine Stichprobe mit dem Stichprobenumfang n, mit m verschiedenen Stichprobenwerten $x_1, x_2, \ldots, x_m$ und den zugehörigen absoluten Häufigkeiten $H(x_1), H(x_2), \ldots, H(x_m)$ bzw. relativen Häufigkeiten $h(x_1), h(x_2), \ldots, h(x_m)$.

Der **Mittelwert** $\bar{x}$ der Stichprobe berechnet sich durch:

$$\bar{x} = \frac{H(x_1) \cdot x_1 + H(x_2) \cdot x_2 + \ldots + H(x_m) x_m}{n}$$

bzw.

$$\bar{x} = h(x_1) \cdot x_1 + h(x_2) \cdot x_2 + \ldots + h(x_m) \cdot x_m$$

Und wie wird der **Mittelwert klassierter Daten** berechnet? Man verwendet dann in den Formeln die **Klassenmitten** als Stichprobenwerte x_i mit den zugehörigen Klassenhäufigkeiten $H(x_i)$ bzw. $h(x_i)$.

Wir halten allgemein fest:

Mittlere absolute Abweichung

Gegeben sei eine Stichprobe mit den Stichprobenwerten $x_1, x_2, \ldots, x_n$, dem Stichprobenumfang n und dem Mittelwert $\bar{x}$.

Die **mittlere absolute Abweichung** d ist die Summe aller absoluten Abweichungen zum Mittelwert geteilt durch den Stichprobenumfang:

$$d = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}$$

Wir erkennen, dass die mittlere absolute Abweichung d ein Mass für die Streuung einer Stichprobe ist. Gilt $d = 0$, so sind alle Stichprobenwerte gleich und die Stichprobe streut überhaupt nicht. Je mehr bzw. grössere oder kleinere Ausreisser es in der Stichprobe gibt, umso grösser ist d.

7.3 Die Standardabweichung und die Varianz

Das wichtigste Streuungsmass in der Statistik ist die **Standardabweichung** s. Zur Berechnung der Standardabweichung verwendet man nicht die absoluten Abweichungen zum Mittelwert, sondern die **Quadrate** der absoluten Abweichungen.

Für die Stichprobe 17, 45, 20, 23, 24, 23, 24, 32, 30, 2, 24, 29, 54 mit dem Mittelwert $\bar{x} = 26.7$ aus der Einstiegsaufgabe ergibt sich dann:

$$\begin{aligned}
&(17-26.7)^2+(45-26.7)^2+(20-26.7)^2+(23-26.7)^2+(24-26.7)^2\\
&+(23-26.7)^2+(24-26.7)^2+(32-26.7)^2+(30-26.7)^2\\
&+(2-26.7)^2+(24-26.7)^2+(29-26.7)^2+(54-26.7)^2\\
&=9.7^2+18.3^2+6.7^2+3.7^2+2.7^2+3.7^2+2.7^2+5.3^2\\
&+3.3^2+24.7^2+2.7^2+2.3^2+27.3^2\\
&=1922.8
\end{aligned}$$

Das ist die Summe der Quadrate der absoluten Abweichungen. Diese Summe teilen wir nun durch den um 1 verminderten Stichprobenumfang[1], also durch $13 - 1 = 12$, und erhalten die (empirische) **Varianz** s^2 der Stichprobe: $s^2 = \frac{1922.8}{12} = 160.2$.

Nehmen wir daraus noch die Wurzel, erhalten wir das wichtigste Streuungsmass, die (empirische) **Standardabweichung:** $s = \sqrt{160.2} = 12.7$.

Wir halten allgemein fest:

[1] Auf die Gründe, warum hier durch $n-1$ und nicht durch den Stichprobenumfang n geteilt wird, kann hier nicht eingegangen werden.

Varianz und Standardabweichung

Gegeben sei eine Stichprobe mit den Stichprobenwerten $x_1, x_2, \ldots, x_n$, dem Stichprobenumfang n und dem Mittelwert $\bar{x}$.

Die (empirische) **Varianz** s^2 der Stichprobe ist die Summe aller Quadrate der absoluten Abweichungen geteilt durch den um 1 verminderten Stichprobenumfang:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

Die (empirische) **Standardabweichung** s der Stichprobe ist die Quadratwurzel aus der Varianz:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

Wegen der Quadrate der Abweichungen gewichtet die Standardabweichung Ausreisser stärker als die mittlere absolute Abweichung.

Beispiel:

In einem Grossbetrieb wurde ein fakultativ zu absolvierender Intelligenztest unter 11 CEOs des Betriebs durchgeführt. Es ergaben sich die folgenden IQ-Punkte: 117, 108, 110, 108, 81, 120, 114, 122, 89, 104, 93. Wir berechnen die Standardabweichung.

Zuerst den Mittelwert bestimmen:

$$\bar{x} = \frac{117 + 108 + 110 + 108 + 81 + 120 + 114 + 122 + 89 + 104 + 93}{11} = \frac{1166}{11} = 106$$

Dann die Summe der quadratischen Abweichungen:

$$\begin{aligned}&(117-106)^2 + (108-106)^2 + (110-106)^2 + (108-106)^2 + (81-106)^2 + (120-106)^2\\&+ (114-106)^2 + (122-106)^2 + (89-106)^2 + (104-106)^2 + (93-106)^2\\&= 11^2 + 2^2 + 4^2 + 2^2 + 25^2 + 14^2 + 8^2 + 16^2 + 17^2 + 2^2 + 13^2\\&= 1748\end{aligned}$$

Geteilt durch $11 - 1 = 10$ ergibt (als Zwischenergebnis) die Varianz s^2:

$$s^2 = \frac{1748}{10} = 174.8$$

Und daraus die Wurzel ergibt die gesuchte Standardabweichung s:

$$s = \sqrt{174.8} = 13.2$$

8.1 Median und Quartile

Der Median $\tilde{x}$ einer Stichprobe hat die Eigenschaft, dass 50% der Stichprobenwerte kleiner und 50% der Stichprobenwerte grösser als $\tilde{x}$ sind.

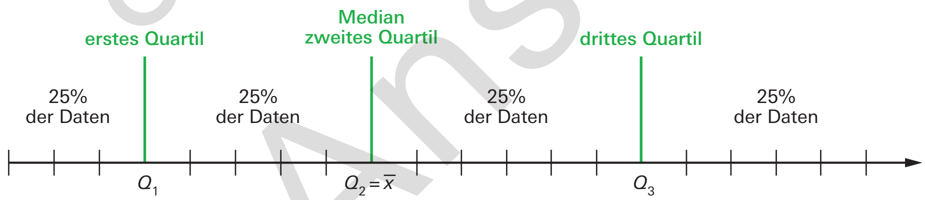
Beispiele:

Die Stichprobe aus der Einstiegsaufgabe hat den ungeraden Stichprobenumfang 13, daher ist der Median der mittlere Wert der geordneten Liste: 2, 17, 20, 23, 23, 24, **25,** 26, 29, 30, 32, 45, 54. Also $\tilde{x} = 25$.

Kommt noch der Wert 56 hinzu, so hat die Stichprobe aus der Einstiegsaufgabe den geraden Stichprobenumfang 14, daher ist der Median der Mittelwert der beiden mittleren Werte der geordneten Liste: 2, 17, 20, 23, 23, 24, **25,** **26,** 29, 30, 32, 45, 54, 56. Also $\tilde{x} = \frac{25+26}{2} = 25.5$.

Statt in zwei gleiche Teile können wir die Stichprobe auch in vier gleiche Teile teilen. Dazu brauchen wir zwei weitere Lagemasse. Erstens einen Wert Q_1, sodass 25% der Stichprobenwerte kleiner und 75% grösser als Q_1 sind. Und zum anderen einen Wert Q_3, sodass 75% der Stichprobenwerte kleiner und 25% grösser als Q_3 sind.

Die Werte Q_1 und Q_3 nennt man **erstes** und **drittes Quartil**[1] der Stichprobe. Zusammen mit dem Median, der das **zweite Quartil** Q_2 bildet, teilen sie die Stichprobe in vier gleiche Teile, in denen je 25% der Stichprobenwerte liegen:



Zwischen dem ersten und dem dritten Quartil liegen 50% aller Stichprobenwerte. Die Differenz aus drittem und erstem Quartil nennt man **Quartilsdifferenz** IQR (engl. «interquartile range»). Die Quartilsdifferenz ist ein Streuungsmass, das die Spannweite der mittleren Hälfte aller Stichprobenwerte angibt.

[1] Quartil (lat.): «Viertelwert».

Aufgabe 63 Gegeben sei die Stichprobe mit den Stichprobenwerten: 118, 123, 188, 176, 136, 150, 137, 177, 174, 165, 171, 172.

a) Bestimmen Sie den Median, das erste und das dritte Quartil und die Quartilsdifferenz.

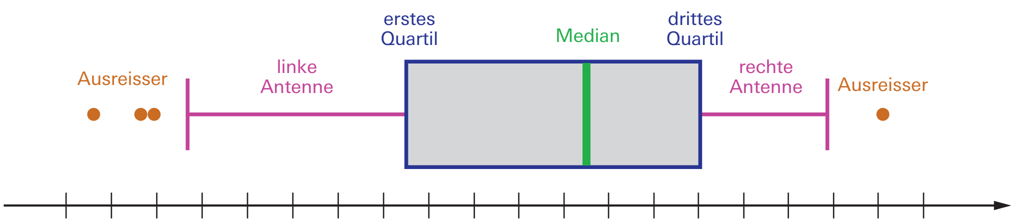
b) Zwischen welchen Werten liegen 75% aller gesammelten Daten?

8.2 Erstellen eines Boxplots

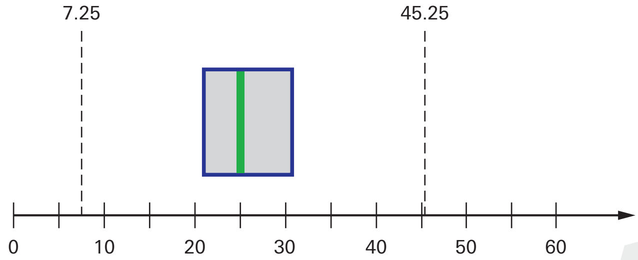
Um einen **Boxplot** für eine Stichprobe zu zeichnen, braucht man folgende Masszahlen der Stichprobe:

- den **Median,**
- **erstes** und **drittes Quartil** und
- die **Quartilsdifferenz.**

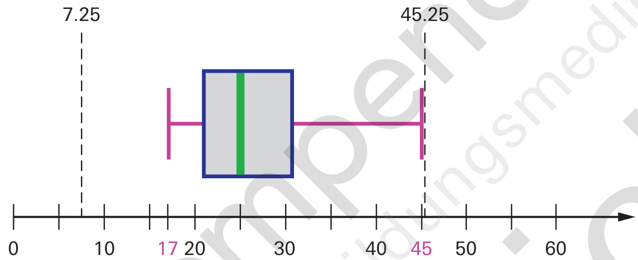
Der Boxplot sieht grundsätzlich so aus:



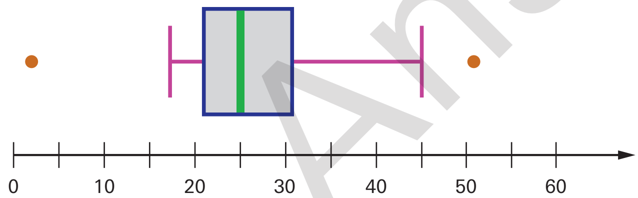
Innerhalb dieses Intervalls $[7.25; 45.25]$ müssen also die Antennen enden. Wir zeichnen diese gedachten Schranken als gestrichelte Linien:



Alle Stichprobenwerte ausserhalb der Schranken bezeichnen wir als Ausreisser. Diese werden für die Antennen nicht berücksichtigt. Wir stellen nun fest, dass der kleinste «normale» Stichprobenwert, der innerhalb der Schranken ist, der Wert 17 ist. Bis dahin geht die linke Schranke. Und der grösste Wert innerhalb der Schranken ist 45, bis dahin geht die rechte Antenne:

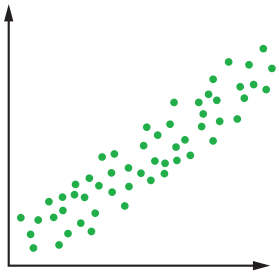


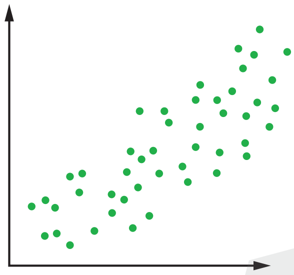
Das ist der Boxplot. Es ist auch noch üblich, die Ausreisser 2 und 52 als einzelne Punkte einzuzeichnen:

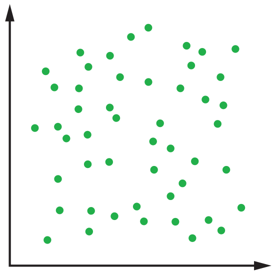


Der Boxplot hilft dabei, sehr schnell einen Überblick über die Daten zu erhalten. So erkennt man direkt, dass der Median bei 25 liegt und dass je 25% der Daten unter 21.5 und über 32 liegen, denn dies sind genau die Abmessungen der Box, in der 50% der Stichprobenwerte enthalten sind. Folglich ist auch die Quartilsdifferenz, die der Länge der Box entspricht, genau 9.5. Auch sind die Ausreisser sofort zu identifizieren.

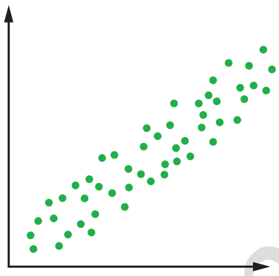
Je näher sich die Punkte an der gedachten Geraden befinden, umso grösser ist die Korrelation:

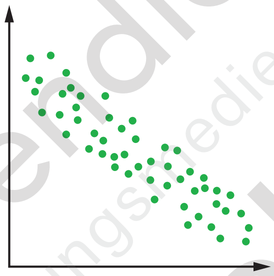

starke Korrelation


schwache Korrelation


keine Korrelation

Je nachdem, ob die gedachte Gerade von links nach rechts steigt oder fällt, ist die Korrelation positiv oder negativ:


positive Korrelation


negative Korrelation

Man spricht von einer **positiven Korrelation,** wenn gilt: Wird das Merkmal X grösser, so wird auch das Merkmal Y tendenziell grösser. Die Punkte im Streudiagramm liegen dann um eine steigende Gerade. Man spricht von einer **negativen Korrelation,** wenn gilt: Wird das Merkmal X grösser, so wird das Merkmal Y tendenziell kleiner. Die Punkte liegen dann um eine fallende Gerade.

Die Korrelation, also die Stärke des Zusammenhangs zweier Merkmale, kann auch durch eine Masszahl, den sog. **Korrelationskoeffizienten** quantitativ beschrieben werden. Wir beschränken uns hier auf eine grafische Beschreibung.

Korrelation

Die **Korrelation** beschreibt die Stärke des Zusammenhangs bivariater Daten.

Werden die bivariaten Daten im Streudiagramm dargestellt, so besteht eine **(lineare) Korrelation** zwischen den Merkmalen, wenn sich die Punkte annähernd auf einer Geraden befinden. Je näher die Punkte sich an der gedachten Geraden befinden, umso stärker ist die Korrelation.

Steigt diese Gerade von links nach rechts, so spricht man von **positiver Korrelation,** wenn sie fällt, von **negativer Korrelation.**

c) Zweites Quartil, Median: $Q_2 = \tilde{x} = \frac{8+9}{2} = 8.5$

Erstes Quartil: $Q_1 = \frac{4+5}{2} = 4.5$

Drittes Quartil: $Q_3 = \frac{12+13}{2} = 12.5$

Quartilsdifferenz: $IQR = 12.5 - 4.5 = 8$

d) Zweites Quartil, Median: $Q_2 = \tilde{x} = \frac{50+51}{2} = 50.5$

Erstes Quartil: $Q_1 = \frac{25+26}{2} = 25.5$

Drittes Quartil: $Q_3 = \frac{75+76}{2} = 75.5$

Quartilsdifferenz: $IQR = 75.5 - 25.5 = 50$

e) Zweites Quartil, Median: $Q_2 = \tilde{x} = 500$

Erstes Quartil: $Q_1 = 250$

Drittes Quartil: $Q_3 = 750$

Quartilsdifferenz: $IQR = 750 - 250 = 500$

Aufgabe 63

a) Geordnete Stichprobe: 118, 123, 136, 137, 150, 165, 171, 172, 174, 176, 177, 188.

Zweites Quartil, Median: $Q_2 = \tilde{x} = \frac{165+171}{2} = 168$

Erstes Quartil: $Q_1 = \frac{136+137}{2} = 136.5$

Drittes Quartil: $Q_3 = \frac{174+176}{2} = 175$

Quartilsdifferenz: $IQR = 175 - 136.5 = 38.5$

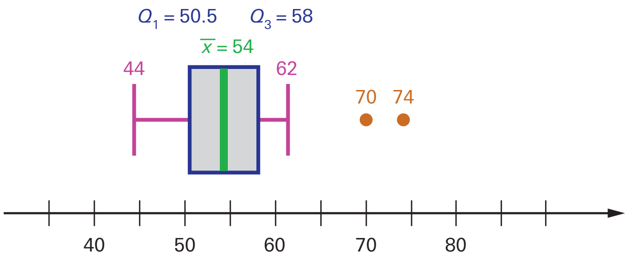
b) 75% aller Daten liegen sowohl zwischen dem kleinsten Stichprobenwert $x_1 = 118$ und dem dritten Quartil $Q_3 = 175$ als auch zwischen dem ersten Quartil $Q_1 = 136.5$ und dem grössten Stichprobenwert $x_{12} = 188$, also je in den beiden Intervallen [118; 175] und [136.5; 188].

Aufgabe 64

1. Median und Quartile: $\tilde{x} = 54$; $Q_1 = 50.5$; $Q_3 = 58$.
2. Quartilsdifferenz: $IQR = Q_3 - Q_1 = 58 - 50.5 = 7.5$.
3. Die maximale Länge der Antennen: $1.5 \cdot IQR = 1.5 \cdot 7.5 = 11.25$.
4. Die Ausreisser befinden ausserhalb des Intervalls $[50.5 - 11.25;\ 58 + 11.25] = [39.25;\ 69.25]$.
5. Der kleinste und der grösste «normale» Wert innerhalb dieses Intervalls: 44 und 62. Bis dahin gehen die Antennen.

Lösungen

6. Boxplot mit den Ausreissern 70 und 74:



Aufgabe 65

a) Durch Ablesen ergeben sich die drei Mediane $\tilde{x}_{96} \approx 75$, $\tilde{x}_{97} \approx 72$, $\tilde{x}_{98} \approx 70$ und die drei Quartilsdifferenzen $IQR_{96} \approx 82-54=28$, $IQR_{97} \approx 84-51=33$, $IQR_{98} \approx 86-52=36$.

b) Die Quartilsdifferenz stellt die Streuung der Hälfte aller Stichprobenwerte dar. Aus diesem Grund ist im Jahr 1998 die Streuung der Prüfungsresultate am grössten.

c) Es ist weitgehend von der Art der Stichprobe abhängig, wie man einen Boxplot interpretiert. In der Pädagogik gibt es ganz unterschiedliche Ansichten darüber, was eine «schlechte» Prüfung ist. Wir sind der Meinung, dass bei der Prüfungsauswertung bei ungefähr gleich grossen Medianen (was hier gegeben ist) eine gute Prüfung möglichst weit im Punktebereich von 0 bis 100 streuen sollte, und dies wiederum würde darauf hindeuten, dass die Prüfung des Jahrs 1998 in diesem Sinn sogar die beste aller drei Prüfungen ist.

Aufgabe 66

Die Quartile teilen eine Stichprobe in vier gleiche Teile. Das zweite Quartil heisst auch Median . Zwischen dem ersten und dem dritten Quartil befinden sich 50% aller Stichprobenwerte. Die Hälfte aller Stichprobenwerte befindet sich auch zwischen dem kleinsten Stichprobenwert und dem Median . Um die Quartilsdifferenz auszurechnen, subtrahieren wir das erste vom dritten Quartil. Diese Differenz wird mit IQR abgekürzt. Bei dem Boxplot ist die Länge der Box gleich der Quartilsdifferenz . Die maximale Länge der Antennen ist das 1.5-Fache der Quartilsdifferenz . Ausserhalb der Antennen befinden sich die Ausreisser .

Aufgabe 67

☒	Bei Stichproben mit lauter gleichen Stichprobenwerten ist die Quartilsdifferenz 0.
☒	Addiert man zu allen Stichprobenwerten der Stichprobe –1, –1.5, –2, –2.5 –3, –3.5, –4 die Zahl 2, dann verändern sich die Quartile.
☒	Addiert man zu allen Stichprobenwerten der Stichprobe –1, –1.5, –2, –2.5 –3, –3.5, –4 die Zahl 2, so bleiben die Spannweite und die Quartilsdifferenz gleich.
☒	Multipliziert man alle Stichprobenwerte der Stichprobe –1, –1.5, –2, –2.5, –3, –3.5, –4 mit der Zahl 10, dann muss man auch die Spannweite und die Quartilsdifferenz mit 10 multiplizieren, um die neue Spannweite und die neue Quartilsdifferenz zu berechnen.
☐	Der Abstand der Antennenenden eines Boxplots gibt immer die Spannweite der Stichprobe an.
☐	Wenn zwei Boxplots eine gleich lange Box haben, dann haben sie den gleichen Median.

Grundlagen der Statistik: Daten, Diagramme, Kennzahlen und Zusammenhänge

Die Lehrmittelreihe VON NULL AUF MATHE vermittelt alle grundlegenden Kompetenzen in Algebra, Analysis, Statistik, Wahrscheinlichkeitsrechnung und Wirtschaftsmathematik. Sie führt Schritt für Schritt durch die Inhalte und steht Ihnen dabei als Lerncoach zur Seite.

Die Erklärungen beginnen bewusst bei den Grundlagen, um auch Lernenden ohne Vorkenntnisse den Einstieg zu erleichtern. Komplexe Themen werden in kleine, verständliche Einheiten gegliedert, die logisch aufeinander aufbauen. Das Konzept lädt zum aktiven Lernen ein. Sie werden motiviert, Aufgaben zu lösen und Ihren Lernerfolg durch Selbstreflexion zu überprüfen. Diese Arbeitsmethodik, unterstützt durch grafische Elemente, zieht sich wie ein roter Faden durch das gesamte Werk. In kleinen Schritten bauen Sie Ihr Wissen nachhaltig auf und schaffen so ein stabiles Fundament für Ihren Erfolg in Schule, Studium, Beruf und Alltag.

VON NULL AUF MATHE eignet sich ideal zum Selbstlernen, kann aber auch im Unterricht und in der Nachhilfe eingesetzt werden. Jeder Band ist in sich abgeschlossen und ermöglicht das gezielte Wiederholen oder Vertiefen einzelner Themen.

Die Reihe VON NULL AUF MATHE umfasst die folgenden elf Bände:

- Die Grundlagen des Rechnens: Zahlen, Brüche, Prozente und Potenzen
- Rechnen mit Buchstaben: Terme und Bruchterme
- Auf geradem Weg: lineare Gleichungen und Funktionen
- Im Bogen zum Ziel: quadratische Gleichungen und Funktionen
- Gemeinsame Lösungen: lineare Gleichungssysteme
- Das Prinzip der Zuordnung: Einführung in die Funktionen
- Funktionen mit Tiefgang: Potenzen, Wurzeln und Polynome
- Wachstum im Blick: Exponential- und Logarithmusfunktionen
- **Grundlagen der Statistik: Daten, Diagramme, Kennzahlen und Zusammenhänge**
- Dem Zufall auf der Spur: Wahrscheinlichkeitsrechnung und Kombinatorik
- Mathematik für Wirtschaft und Finanzen: Zinsen, Preise und Optimierung

Compendio Bildungsmedien
www.compendio.ch

ISBN 978-3-7155-0136-9

CHRISTOPH GERBER Dr. phil. nat., Lehrdiplom für Maturitätsschulen
Studium der Mathematik, Astronomie und Philosophie an der Universität Bern. Unterrichtete mehrere Jahre Mathematik am Gymnasium. Langjähriger Dozent für die Fachausbildung Mathematik am Sekundarlehramt der Universität und der Pädagogischen Hochschule Bern.

PETER JANKOVICS Dipl.-Math.
Studium der Mathematik und Physik an der TU Berlin. Danach Lehrbeauftragter an der Hamburger Fern-Hochschule. Als Autor, Lektor und Redaktor für zahlreiche Verlage im Bildungssektor tätig, seit 2018 für Compendio Bildungsmedien.