

Statistik

Beschreibende Statistik



Skript 2024/2025
Mathematik - GYM 3

Erstellt von
Krisanth Vyithiyalingam
basierend auf Skripten von P.Marti (MAP)
www.vyk-mip.ch

gym|THUN

Eine Institution des Kantons Bern

1 Daten erheben und darstellen

1.1 Einführung

In der Statistik werden Massenerscheinungen in Gesellschaft, Naturwissenschaften, Technik und Medizin mit Mitteln der Wahrscheinlichkeitsrechnung untersucht. Bevor man allerdings die Realität durch Wahrscheinlichkeitsmodelle beschreibt, muss man sie durch Messen, Zählen und Visualisieren «erfassen». Dies geschieht in der beschreibenden Statistik.

Die wichtigsten Fragen und Probleme werden im folgenden Beispiel erläutert:

Beispiel 1:

Ein Bauer will die Wirkung eines neuen biologischen Düngers auf das Wachstum der Erbsenpflanze untersuchen. Er erhofft sich eine Steigerung der Ernte durch den Dünger. Er unterteilt sein Versuchsfeld in zwei gleiche Stücke und pflanzt gleich viele Erbsen in beiden Teilen. Beide Teile werden gleich behandelt, mit Ausnahme des Düngers.

Eine zufällige Stichprobe von 150 Hülsenfrüchten wird eingesammelt von beiden Teilen, zur selben Zeit. Anschliessend wird gezählt wie viele Erbsen in jeder Hülsenfrucht vorhanden waren. Die Resultate sind:

Ohne Dünger:

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

Mit Dünger:

6	7	7	4	9	5	5	5	8	9	8	9	7	7	5	8	7	6	6	7	9	7
7	7	8	9	3	7	4	8	5	10	8	6	7	6	7	5	6	8	7	9	4	4
9	6	8	5	8	7	7	4	7	8	10	6	10	7	7	7	9	7	7	8	6	8
6	8	7	4	8	6	8	7	3	8	7	6	9	7	6	9	7	6	8	3	9	5
7	6	8	7	9	7	8	4	8	7	7	7	6	6	8	6	3	8	5	8	7	6
7	4	9	6	6	6	8	4	7	8	9	7	7	4	7	5	7	4	7	6	4	6
7	7	6	7	8	7	6	6	7	8	6	7	10	5	13	4	11	5				

Zum darüber nachdenken:

- Können Sie das Problem genau beschreiben welches der Bauer lösen möchte?
- Mit welchen Mitteln versuchte der Bauer einen fairen Vergleich zu erreichen?
- Wie stellte der Bauer sicher, dass die gewählten Hülsenfrüchte zufällig gepflückt wurden?
- Wie könnte man die Daten am sinnvollsten auflisten?
- Wie könnte man die Daten am besten darstellen?
- Gibt es Ausreisser nach oben oder unten? Wie sollten wir solche behandeln?
- Wie wird am besten die typische Grösse einer Hülsenfrucht beschrieben?
- Wie können wir die Variationen der Hülsenfrüchte beschreiben?
- Können wir eine zufriedenstellende Aussage treffen?

1.2 Daten erheben

Befragungen von Personen oder das Zählen von Gegenständen sind Beispiele von statistischen Erhebungen, bei denen Daten gesammelt werden. Bei einer solchen Erhebung wird an Merkmalsträgern (z. B. Personen oder Autos) ein bestimmtes Merkmal (z. B. Grösse oder Geschwindigkeit) festgestellt. Alle Merkmalsträger zusammen nennt man Grundgesamtheit. Werte, die von den Merkmalen angenommen werden können, heissen Merkmalsausprägungen. Die folgende Tabelle zeigt einige Beispiele.

Grundgesamtheit	Merkmalsträger	Merkmal	Merkmalsausprägungen
Einwohner einer Stadt	Person	Alter	..., 24, 25, 26, ...
Bäume eines Waldes	Baum	Baumart	Buche, Fichte, ...
Bäume eines Waldes	Baum	Höhe in m	..., 8, 9, 10, ...
Tiere eines Bauernhofs	Tier	Geschlecht	männlich, weiblich

Man unterscheidet zwei Arten von Merkmalen. Ein Merkmal, dessen Ausprägung nur durch Zahlen dargestellt werden kann, nennt man **quantitatives Merkmal** (z. B. Alter, Grösse), andernfalls heisst es **qualitatives Merkmal** (z. B. Baumart, Geschlecht). Wenn die Ausprägungen eines quantitativen Merkmals nur isolierte Zahlenwerte sind (z. B. Anzahl der Schweine eines Bauernhofs, Alter in Jahren), so spricht man von einem **diskreten Merkmal** oder **diskreter Zufallsvariable**. Im Gegensatz dazu können **stetige Merkmale (stetige Zufallsvariablen)** prinzipiell alle Zahlen eines Intervalles annehmen (z. B. Höhe eines Baumes, Bremsweg).

Bei einer statistischen Erhebung kann es vorkommen, dass der Umfang der Grundgesamtheit zu gross ist (z. B. Gesamtbevölkerung) oder die Untersuchungsobjekte zerstört werden (z.B. Ausbrennen einer Glühbirne beim Testen der Lebensdauer). In diesen Fällen untersucht man nur die Merkmalsträger einer Stichprobe (Teilmenge der Grundgesamtheit). Die Anzahl n der Merkmalsträger in der Stichprobe heisst Stichprobenumfang.

Das Vorgehen einer statistischen Erhebung soll am Beispiel eines Betriebes, in dem die Altersstruktur untersucht wird, beschrieben werden. Aus einer Stichprobe von 50 Angestellten werden für das Merkmal «Alter» die Werte in der Reihenfolge ihrer Erfassung in einer Urliste (Abb. 1.2.1) gesammelt.

23	18	27	28	39	46	23	59	35	43
33	35	56	61	27	20	31	23	50	19
23	60	42	41	43	33	27	19	52	46
23	35	20	53	27	27	43	33	52	20
27	55	30	30	21	44	46	49	28	57

Abb. 1.2.1: Urliste

Um eine bessere Übersicht zu gewinnen, ordnet man die Werte der Grösse nach und zählt, z. B. mithilfe einer Strichliste (Abb. 1.2.2), wie oft jeder einzelne Wert vorkommt. Als Ergebnis erhält man die absolute Häufigkeit jeder Ausprägung. Dividiert man die absolute Häufigkeit durch den Stichprobenumfang, so erhält man die relative Häufigkeit.

Die Zuordnung von Häufigkeiten zu den verschiedenen Merkmalsausprägungen nennt man Häufigkeitsverteilung.

Alter	Strichliste	absolute Häufigkeit	relative Häufigkeit
18		1	0,02 = 2%
19		2	0,04 = 4%
20		3	0,06 = 6%
21		1	0,02 = 2%
23		5	0,10 = 10%
25		2	0,04 = 4%
27		6	0,12 = 12%
28		2	0,04 = 4%
30		2	0,04 = 4%
31		1	0,02 = 2%
33		3	0,06 = 6%
35		1	0,02 = 2%
39		1	0,02 = 2%
41		1	0,02 = 2%

Alter	Strichliste	absolute Häufigkeit	relative Häufigkeit
42		1	0,02 = 2%
43		3	0,06 = 6%
44		1	0,02 = 2%
46		3	0,06 = 6%
49		1	0,02 = 2%
50		1	0,02 = 2%
52		2	0,04 = 4%
53		1	0,02 = 2%
55		1	0,02 = 2%
56		1	0,02 = 2%
57		1	0,02 = 2%
59		1	0,02 = 2%
60		1	0,02 = 2%
61		1	0,02 = 2%

Abb. 1.2.2

Soll das Wesentliche einer Häufigkeitsverteilung deutlicher herausgearbeitet oder eine grössere Übersichtlichkeit erzielt werden, so kann man die Merkmalsausprägungen in Klassen (Intervallen, Kategorien) zusammenfassen (Abb. 1.2.3). Der Gewinn an Übersichtlichkeit durch die Klasseneinteilung bringt jedoch einen Verlust an Information.

Alter X	abs. Häufigkeit	rel. Häufigkeit
$10 \leq X < 20$	3	6%
$20 \leq X < 30$	19	38%
$30 \leq X < 40$	8	16%
$40 \leq X < 50$	10	20%
$50 \leq X < 60$	8	16%
$60 \leq X < 70$	2	4%

Abb. 1.2.3

1.3 Grafische Darstellungen von Daten

Bei einer Häufigkeitsverteilung in Tabellenform fällt es oft schwer, die Verteilung von Ausprägungen eines Merkmals schnell zu erfassen. Deshalb werden Häufigkeitsverteilungen auch grafisch dargestellt.

Will man die tabellarische Häufigkeitsverteilung der einzelnen Ausprägungen (Abb. 1.2.2) grafisch darstellen, kann man ein Säulendiagramm oder Stabdiagramm (Abb. 1.3.1) erstellen. Dabei wird auf der waagerechten Achse die Ausprägung (Alter) des Merkmals markiert. Über den einzelnen Ausprägungen werden dann Säulen oder Stäbe mit der Länge der entsprechenden relativen bzw. absoluten Häufigkeiten gezeichnet.

Die durch die Klasseneinteilung entstandene Häufigkeitsverteilung (Abb. 1.2.3) ist durch ein Kreisdiagramm (Abb. 1.3.2) oder durch ein Histogramm (Abb. 1.3.3) darstellbar.

Beim Kreisdiagramm sind die Flächen der Kreissektoren proportional zu den Häufigkeiten. Beim Histogramm wird über jedem Teilintervall (Klasse) ein Rechteck errichtet, dessen Fläche proportional zur jeweiligen Klassenhäufigkeit ist. Bei gleich grossen Klassenbreiten ergibt sich die Rechteckhöhe direkt aus den relativen bzw. absoluten Häufigkeiten.

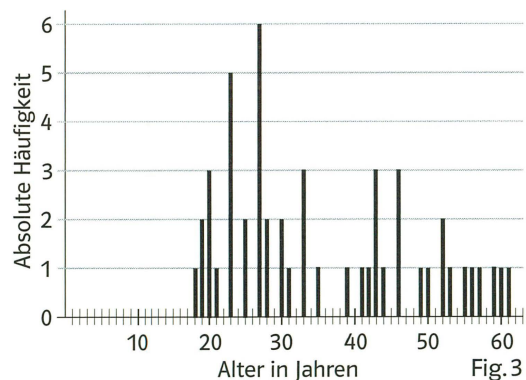


Abb. 1.3.1

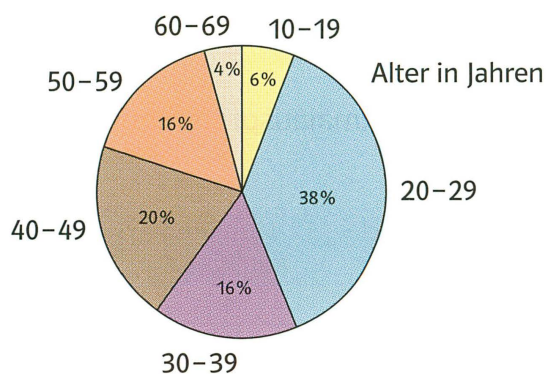


Abb. 1.3.2

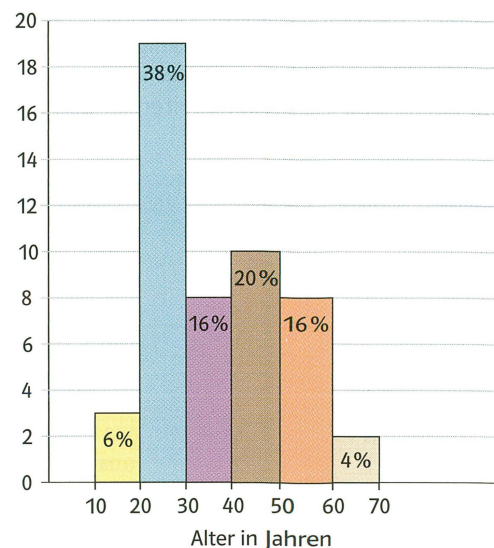


Abb. 1.3.3

Vorgehen bei der Durchführung einer statistischen Erhebung

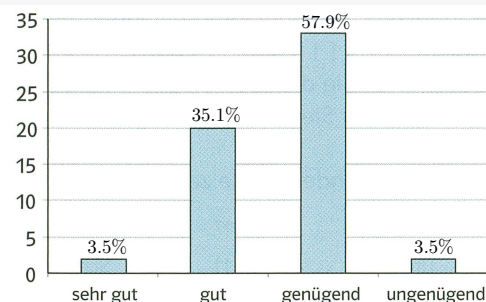
1. Merkmal festlegen.
2. Urliste einer Stichprobe erfassen.
3. Eventuell Klassen festlegen.
4. Häufigkeiten der Merkmalsausprägungen bzw. der Klassen bestimmen.
5. Diagramm zeichnen.

Beispiel 2:

In einer Stichprobe aus 57 Klausuren eines Jahrganges wurde das Merkmal «Punkte» erhoben. Es ergab sich nebenstehende Urliste. Ermitteln Sie die Häufigkeiten, die zu den Merkmalsklassen «sehr gut» ($13 \leq X \leq 15$), «gut» ($10 \leq X \leq 12$), «genügend» ($7 \leq X \leq 9$) und «ungenügend» ($4 \leq X \leq 6$) gehören. Stellen Sie die Verteilung in einem Säulendiagramm dar.

7	9	10	8	12
12	7	11	8	9
10	8	6	7	9
11	12	10	7	8
9	13	9	12	8
5	11	8	7	11
12	9	11	7	8
7	8	12	15	9
8	11	8	7	8
8	10	7	9	12
10	10	7	9	8
12	7			

Klasse	sehr gut	gut	gen.	ungen.
Strichliste		 	 	
abs. H.	2	20	33	2
rel. H.	3,51%	35,09%	57,89%	3,51%



1.3.1 Suggestive Darstellung von Statistiken

«Ein Bild sagt mehr als tausend Worte.» Dieses Sprichwort befolgend werden uns täglich statistische Informationen in Form von Diagrammen präsentiert, die Botschaften übermitteln möchten. Die meisten statistischen Grafiken sind einwandfrei, mitunter wird aber auch gemogelt, um beim Betrachter einen gewünschten Eindruck hervorzurufen.

1.3.2 Trick #1: Wirkung von Skalierungen

Durch die Wahl von Skalierungen, insbesondere durch das Ein- und Ausblenden des Koordinatenursprungs, kann man mit Diagrammen sehr verschiedene Eindrücke erzeugen. Während Abb. 1.3.4 einen dramatischen Mitgliederschwund in einem Verein signalisiert die Mitgliederzahlen scheinen auf ein Achtel zurückgegangen zu sein -, macht Abb. 1.3.5 deutlich, dass der Rückgang um 0,2% eigentlich unbedeutend ist.

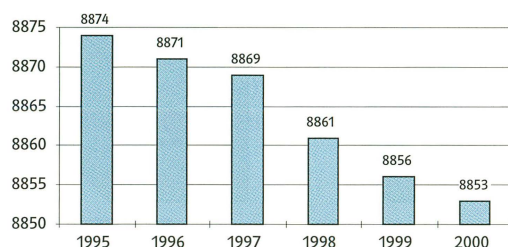


Abb. 1.3.4

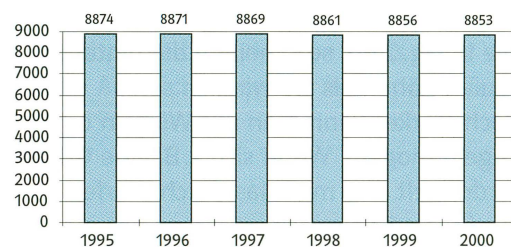


Abb. 1.3.5

Abb. 1.3.7 zeigt die Ergebnisse einer Modellrechnung zur Entwicklung der Weltbevölkerung für

ausgeht. Sie hinterlässt einen eher harmlosen Eindruck. Abb. 1.3.6 zeigt die Ergebnisse der gleichen Modellrechnung für die Jahre 0 bis 2020. Diese hinterlässt den Eindruck einer gegen Ende des betrachteten Zeitraumes dramatischen «Bevölkerungsexplosion».

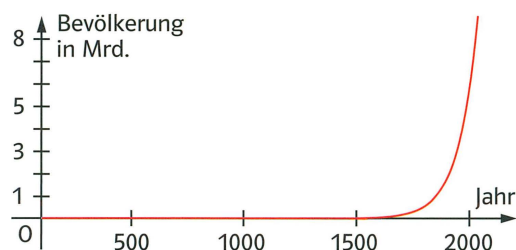


Abb. 1.3.6

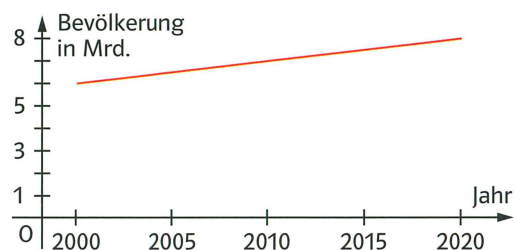


Abb. 1.3.7

Bei jedem Diagramm sollte deshalb die Skalierung der Achsen kontrolliert werden.

1.3.3 Trick #2: Weglassen «unbequemer» Daten

Den Leser kann man auch dadurch täuschen, dass man einfach nicht alle Daten darstellt.

Um mögliche Investoren anzulocken, zeigt man nur die rechte Grafik, in welcher der Aktienverlauf stoppt, sobald der Wert der Aktie zu fallen beginnt.

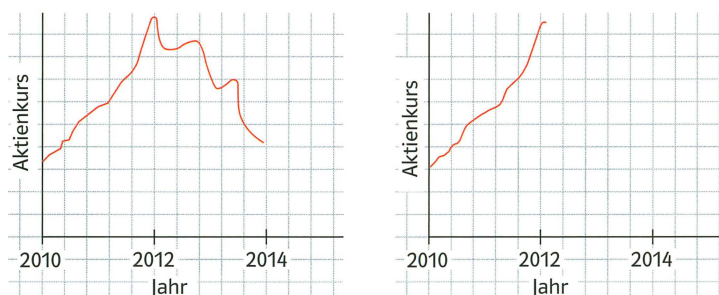


Abb. 1.3.8

1.3.4 Trick #3: Falsche zwei- und dreidimensionale Darstellung

In Abb. 1.3.9 sank die Anzahl neuzugelassener Autos vom September 2008 zum Oktober 2008 um die Hälfte. Werden Länge und Höhe des grossen Autos halbiert, resultiert das kleine Auto, dessen Fläche nur ein Viertel der Fläche des grossen Autos ist. Dadurch erscheint die Abnahme drastischer, als sie in Wirklichkeit ist.

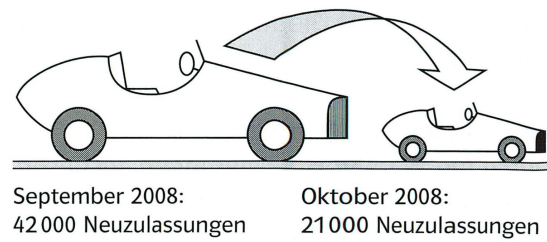


Abb. 1.3.9

Noch stärker kann man mit Würfeln oder Kugeln mogeln. Wenn sich z. B. eine Grösse verdoppelt und man das durch Würfel wie in Abb. 1.3.10 a.) veranschaulicht, so hat man sogar einen 8-mal so grossen Würfel dargestellt. Der zweite Würfel dürfte aber nur eine $\sqrt[3]{2}$ -mal so lange Kante haben (Abb. 1.3.10 b.)).

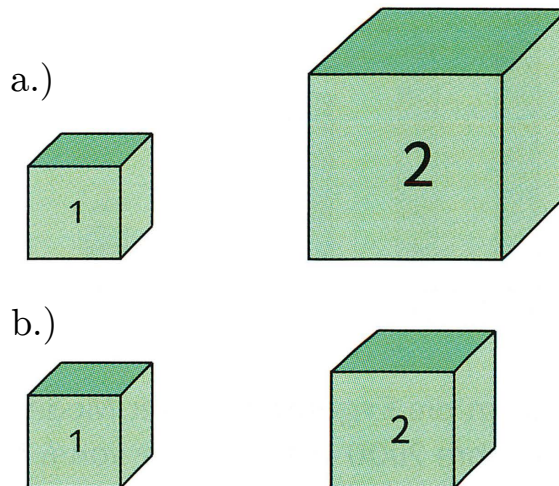
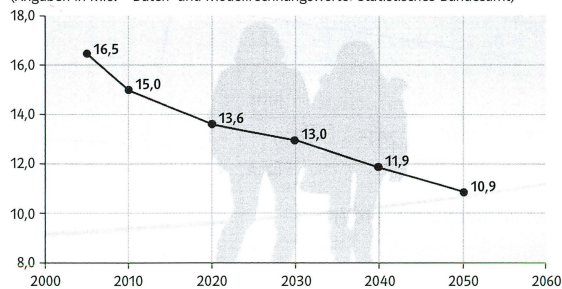


Abb. 1.3.10

Untersuchen Sie die folgenden Grafiken im Hinblick auf Mogeleyen:

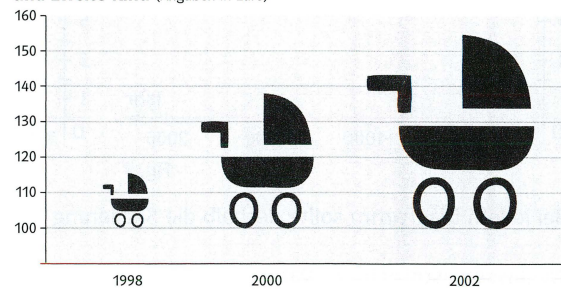
Anzahl der Jugendlichen unter 20 Jahren

(Angaben in Mio. – Daten- und Modellrechnungswerte: Statistisches Bundesamt)



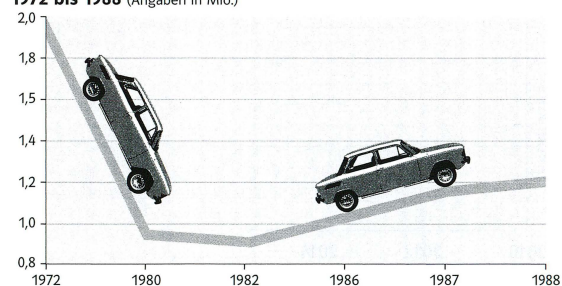
Entwicklung des Kindergeldes für das erste und zweite Kind

(Angaben in Euro)

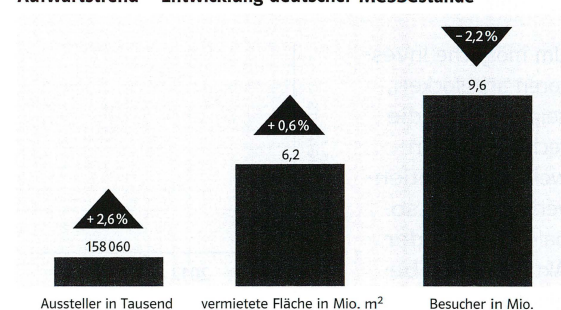


Produktion von Personenwagen in Grossbritannien

1972 bis 1988 (Angaben in Mio.)



Aufwärtstrend – Entwicklung deutscher Messestände



Übungen: Daten erheben und darstellen

- 1.1: Die Häufigkeitstabelle für die Grössen der Spieler in einem Basketballteam ist gegeben als:

Grösse (cm)	Häufigkeit
170 - <175	1
175 - <180	8
180 - <185	9
185 - <190	11
190 - <195	9
195 - <200	3
200 - <205	3

- a) Erklären Sie, weshalb die Grösse eine stetige Variable ist.
- b) Zeichnen Sie ein Histogramm für die Daten. Die Achsen sollten beschriftet sein und ein Titel gesetzt.
- c) Was ist die am häufigsten vorkommende Klasse?
- 1.2: Eine Schule wollte wissen welche Zeit die Schüler für den Schulweg aufwenden. Dazu wurden 60 Individuen befragt. Die Daten geben die Dauer gerundet auf Minuten an:

12 15 16 8 10 17 25 34
42 18 24 18 45 33 38 45
40 3 20 12 10 10 27 16
37 45 15 16 26 32 35 8
14 18 15 27 19 32 6 12
14 20 10 16 14 28 31 21
25 8 32 46 14 15 20 18
8 10 25 22

- a) Ist die Dauer auf dem Schulweg eine diskrete oder stetige Variable?
- b) Zeichnen Sie ein Histogramm
- c) Was ist die häufigste Zeit?
- 1.3: Entscheiden Sie für die folgenden Datensätze ob ein Histogramm oder ein Säulendiagramm besser geeignet ist und zeichnen Sie die Diagramme.
- a) Die Anzahl an Streichhölzern in 30 Schachteln:

Anzahl Hölzer pro Schachtel	Häufigkeit
47	1
49	1
50	9
51	12
52	4
53	2
55	1

b) Die Grösse von 25 Hockeyspielern (gerundet auf cm):

Grösse (cm)	Häufigkeit
120 - 129	1
130 - 139	2
140 - 149	7
150 - 159	14
160 - 169	1

1.4: Ein Pflanzeninspektor wählt zufällig sechs Monate alte Setzlinge und misst die Höhe auf einen mm genau. Die Resultate sind angegeben:

Höhe (mm)	Häufigkeit
300 - 324	12
325 - 349	18
350 - 374	42
375 - 399	28
400 - 424	14
425 - 449	6

a) Zeichnen Sie die Daten in ein Histogramm.

b) Wie viele Setzlinge sind grösser als 400 mm?

c) Wie viel Prozent der Setzlinge sind zwischen 349 und 400 mm?

d) Insgesamt befinden sich 1462 Setzlinge in der Pflanzenzucht. Schätzen Sie die Anzahl an Setzlingen welche weniger als 400 mm höhe haben, bzw. welche eine Höhe zwischen 375 und 425 mm haben.

2 Statistische Kennzahlen

«Die Mathematik- und die Französisch-Prüfungen sind gleich ausgefallen, da in beiden Prüfungen der Klassendurchschnitt eine 4 ist», meint Lejla. «Für mich gibt es aber schon Unterschiede zwischen den beiden Prüfungen», entgegnet Max.

Note	1	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	Schnitt
Mathematik	0	1	0	2	2	3	4	4	2	2	1	4
Französisch	0	0	0	0	3	5	6	5	3	0	0	4

Will man zwei Stichproben vergleichen, so kann man entsprechende Diagramme erstellen und daraus Schlussfolgerungen ziehen. Häufig sind solche ausführlichen Vorgehensweisen nicht notwendig und es reicht, wenn man Zahlen gegenüberstellt, die möglichst viele Informationen über die jeweilige Stichprobe enthalten. Solche Zahlen heißen Kennzahlen einer Stichprobe.

Besonders aussagekräftig sind die Kennzahlen, die über die Lage des Zentrums einer Verteilung (**Lagemasse**), sowie Kennzahlen, die über die Abweichungen der Werte vom Zentrum Auskunft geben (**Streuungsmaße**).

2.1 Das Zentrum der Daten bestimmen

Betrachtet man die Messreihe einer Stichprobe, kann man sich fragen: «Wo liegen die meisten Werte?», «Wo liegt die Mitte der Verteilung?» oder «Was ist ein typischer Wert?». Solche Fragen beantworten Lagemasse.

2.1.1 Mittelwert

Das in der Praxis am häufigsten verwendete Lagemass ist das **arithmetische Mittel**, auch **Mittelwert** oder **Durchschnitt** genannt.

Arithmetisches Mittel einer Stichprobe

Es seien $x_1; x_2; \dots; x_n$ Werte einer Stichprobe mit Stichprobenumfang n . Dann ist das arithmetische Mittel $\bar{x}$ gegeben durch:

$$\bar{x} = \frac{(x_1 + x_2 + \dots + x_n)}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Treten in der Stichprobe $x_1; x_2; \dots; x_n$ gewisse Werte x_i mehrmals auf (n_i) oder ist die relative Häufigkeit h_i der Werte x_i bekannt, muss nicht jedes x_i einzeln aufsummiert werden.

Arithmetisches Mittel aus einer Häufigkeitsverteilung

Nimmt eine Stichprobe die k verschiedenen Werte $x_1; x_2; \dots; x_k$ mit den absoluten Häufigkeiten $n_1; n_2; \dots; n_k$ bzw. mit den relativen Häufigkeiten $h_1; h_2; \dots; h_k$ an, so wird das arithmetische Mittel $\bar{x}$ bestimmt durch:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n (n_i x_i) = \sum_{i=1}^k (h_i x_i)$$

Bemerkung: Wenn die Messwerte zahlreicher sind bzw. bereits in Klassen eingeteilt wurden, verwendet man die Klassenmitteln (Mitte der Klassenintervalle) als Messwerte x_i und die zugehörigen Klassenhäufigkeiten h_i in den Formeln.

2.1.2 Median

Oft liegen die Stichprobenwerte in der Nähe ihres arithmetischen Mittels. Der Mittelwert repräsentiert dann den «typischen» Wert eines Merkmals gut. Das ist allerdings nicht mehr der Fall,

wenn sich unter den Werten einzelne Ausreisser befinden, die sehr stark von den anderen Werten abweichen. Die Verteilung ist dann nicht mehr symmetrisch, man spricht von einer (links- oder rechts-)schiefen Verteilung.

Wohnen beispielsweise in einem kleinen Dorf neun Bauern, die monatlich jeweils ungefähr 4000 Fr. verdienen, und ein Reicher, der ein Einkommen von 300 000 Fr. hat, so ist der durchschnittliche Verdienst etwas über 30'000 Fr. Verlässt man sich nur auf diese Kennzahl, so erscheint das Dorf wohlhabend. Das arithmetische Mittel repräsentiert hier nicht die tatsächliche Situation.

Bei Messreihen mit Ausreissern, die den Mittelwert stark beeinflussen, zieht man den **Median (Zentralwert)** dem arithmetischen Mittel vor. Zur Bestimmung des Medians betrachtet man eine **geordnete Stichprobe**, bei der die Werte der Grösse nach geordnet sind.

Der Median einer geordneten Stichprobe ist

- bei ungeradem Stichprobenumfang der in der Mitte stehende Wert
- bei geradem Stichprobenumfang das arithmetische Mittel der beiden in der Mitte stehenden Werte.

Die so erhaltene Zahl hat die Eigenschaft, dass die Hälfte der Werte darunter, die Hälfte darüber liegt.

Beispiel 3:

Der Median im obigen Beispiel mit dem kleinen Dorf lässt sich also wie folgt bestimmen: Wir ordnen die Liste der Verdienste der Grösse nach

3'700, 3'800, 3'900, 4'000, 4'100, 4'200, 4'300, 4'400, 4'500, 300'000

und nehmen den Wert in der Mitte. Da es sich um einen geraden Stichprobenumfang handelt (10 Werte) bilden wir das arithmetische Mittel der beiden in der Mitte stehenden Werte:

$$\frac{4'100 + 4'200}{2} = 4150$$

Der Median ist also 4150 Fr., was die Situation viel besser repräsentiert.

2.1.3 Modus

Das einfachste und seltenste Lagemass schliesslich ist der sogenannte Modus (Modalwert). Für stetige Daten können wir keinen Modus finden, denn es gibt keine zwei *exakt* identische Werte. Daher sprechen wir bei stetigen Daten über die **Modalklasse**, das ist die Gruppe welche am häufigsten auftritt.

Der Modus einer Stichprobe ist der Wert, der am häufigsten auftritt.

Wenn die Daten zwei Werte haben, die am häufigsten vorkommen, sprechen wir von einer **bimodalen** Verteilung. Wenn eine Verteilung mehr als 2 Modi hat, verwenden wir den Modus nicht mehr als Mass für die Mitte.

Mittelwert, Median und Modus in verschiedenen Verteilungen

Bei symmetrischen Verteilungen ist der Mittelwert und der Median in etwa identisch.

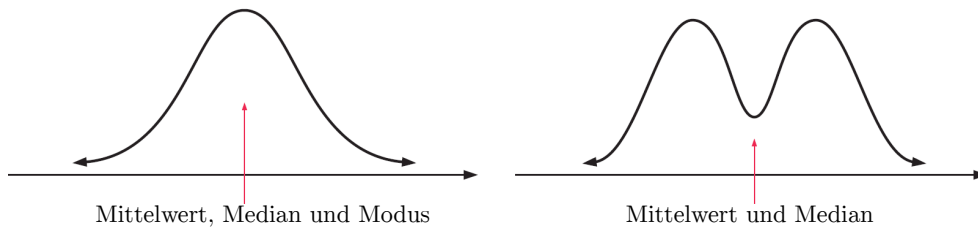


Abb. 2.1.1: Symmetrische Verteilungen

Wenn die Verteilung nicht symmetrisch ist, könnte diese Links- oder Rechtsschief sein:

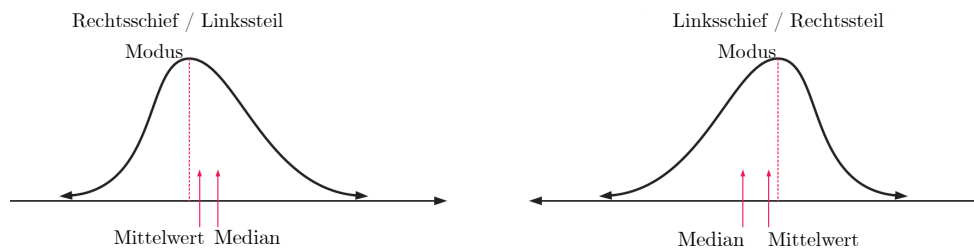


Abb. 2.1.2: Schiefe Verteilungen

Beachten Sie: Median und Mittelwert sind deutlich unterschiedlich.

Einsatz des Taschenrechners

Bei grossen Datensätzen macht es Sinn, den TR als Hilfsmittel zu verwenden. Am besten wird die Vorgehensweise anhand von Beispielen klar.

Daten in einer Häufigkeitstabelle

Beispiel 4:

Die Messung der Grösse (in cm) von 103 Schülern wurde in einer Strichliste notiert:

161	166		171		176		181		186	
162	167		172		177		182		187	
163	168		173		178		183		188	
164	169		174		179		184		189	
165	170		175		180		185		190	

Bestimmen Sie den Modalwert und berechnen Sie Mittelwert und Median.

Die Daten werden unter Lists & Spreadsheet in eine Tabelle eingegeben. Nun macht es Sinn, eine weitere Spalte hinzuzufügen, in der das Produkt $f \cdot x$ steht.

Werte x	Häufigkeit h	Produkt $h \cdot x$
163	1	$1 \cdot 163 = 163$
164	1	$1 \cdot 164 = 164$
165	1	$1 \cdot 165 = 165$
166	3	$3 \cdot 166 = 498$
167	3	$3 \cdot 167 = 501$
...	...	...
Total	$\sum h = 103$	$\sum(h \cdot x) = 18111$

Modus: Aus der Strichliste ist direkt ersichtlich, dass 175, 176 und 179 mit je der Häufigkeit 7 auftreten. Da dies mehr als zwei Werte sind, haben wir keinen Modalwert in dieser Verteilung.

Mittelwert: $\bar{x} = \frac{\sum(h \cdot x)}{\sum h} = \frac{18111}{103} = 175.84$

Median: Es gibt insgesamt 103 Werte. Der Median ist der mittlere Wert, also der 52. Wert, also 176.

Daten in Klassen

Falls Informationen schon in Klassen eingeteilt gesammelt wurden, dann verwenden wir den **Mittelpunkt** der Klasse um alle Werte in diesem Intervall zu repräsentieren.

Wir gehen davon aus, dass alle Werte gleichmässig auf das Intervall verteilt sind. Daher werden die Resultate auch nur angenähert sein.

Beispiel 5:

Berechnen Sie den Mittelwert des Alters der folgenden Busfahrer und die Modalklasse:

Alter	21-25	26-30	31-35	36-40	41-45	46-50	51-55
Häufigkeit	11	14	32	27	29	17	7

Alter	Häufigkeit (h)	Mittelpunkt (x)	hx
21-25	11	23	253
26-30	14	28	392
31-35	32	33	1056
36-40	27	38	1026
41-45	29	43	1247
46-50	17	48	816
51-55	7	53	371
Total	$\sum h = 137$		$\sum hx = 5161$

Modalklasse: Sie kann einfach in der Tabelle abgelesen werden: 31-35.

Mittelwert:

$$\bar{x} = \frac{\sum(h \cdot x)}{\sum h} = \frac{5161}{137} = 37.7$$

2.1.4 Kumulierte Häufigkeit

Oftmals ist es nützlich wenn wir die Anzahl Daten kennen welche über, bzw. unter einem Wert liegen. In solchen Situationen können wir eine **kumulierte Häufigkeitstabelle** erstellen und einen Graphen, die **Häufigkeitsverteilungskurve**.

Beispiel 6:

Die Daten geben das Gewicht von 120 Fussballspielern an.

Gewicht (kg)	Häufigkeit
55 - <60	2
60 - <65	3
65 - <70	12
70 - <75	14
75 - <80	19
80 - <85	37
85 - <90	22
90 - <95	8
95 - <100	2
100 - <105	1

a) Erstellen Sie eine Tabelle mit kumulierter Häufigkeit.

b) Stellen Sie die Daten in einer Häufigkeitsverteilungskurve dar.

c) Benutzen Sie den Graphen um folgendes abzuschätzen:

i) Durchschnittliches Gewicht

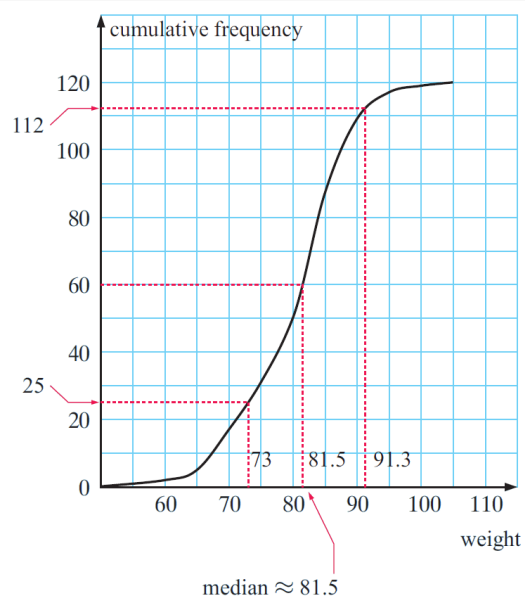
ii) Anzahl Männer mit Gewicht unter 73 kg

iii) Anzahl Männer mit Gewicht grösser als 92 kg.

a)

Gewicht (kg)	Häufigkeit	kumulierte Häufigkeit
55 - < 60	2	2
60 - < 65	3	5
65 - < 70	12	17
70 - < 75	14	31
75 - < 80	19	50
80 - < 85	37	87
85 - < 90	22	109
90 - < 95	8	117
95 - < 100	2	119
100 - < 105	1	120

b)



c)

i) Der Median wird abgeschätzt durch das 50ste Perzentil.

Also 50% von 120 = 60, Wir starten bei der kumulierten Häufigkeit bei 60 und finden das entsprechende Gewicht.

Also ist der Median ≈ 81.5 .

ii) Es gibt 25 Männer mit weniger als 73 kg Gewicht.

iii) Es gibt $120 - 112 = 8$ Männer mit Gewicht mehr als 91.3 kg.

Übungen: Das Zentrum der Daten bestimmen

2.1: Berechnen Sie den Mittelwert, den Median und den Modus für die folgenden Datensammlungen:

a) 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9

b) 10, 12, 12, 15, 15, 16, 16, 17, 18, 18, 18, 18, 19, 20, 21

c) 22.4, 24.6, 21.8, 26.4, 24.9, 25.0, 23.5, 26.1, 25.3, 29.5, 23.5

2.2: Betrachten Sie die folgenden beiden Sammlungen von Daten:

Sammlung A: 3, 4, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 10

Sammlung B: 3, 4, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 15

a) Berechnen Sie den Mittelwert für beiden Sammlungen A und B.

b) Berechnen Sie den Median für beide.

c) Weshalb ist der Mittelwert von A kleiner als jener von B?

d) Weshalb ist der Median für beide identisch?

2.3: Der Jährliche Lohn von zehn Büroangestellten ist:

CHF 23000, CHF 46000, CHF 23000, CHF 38000, CHF 24000,

CHF 23000, CHF 23000, CHF 38000, CHF 23000, CHF 32000

a) Berechnen Sie den Median, den Modus und den Mittelwert für diese Gruppe.

b) Weshalb ist der Modus nicht geeignet als Mass der Mitte?

c) Ist der Median ein gutes Mass für die Mitte?

2.4: Die folgenden Messwerte geben den täglichen Niederschlag (gemessen in mm) für den Monat Juli 2007 in der Wüste:

3, 1, 0, 0, 0, 0, 0, 2, 0, 0, 3, 0, 0, 0, 7, 1, 1, 0, 3, 8, 0, 0, 0, 42, 21, 3, 0, 3, 1, 0, 0

a) Berechnen Sie den Mittelwert, den Median und den Modus für diese Daten.

b) Weshalb ist der Median nicht der beste Wert um das Zentrum zu beschreiben?

c) Weshalb ist der Modus nicht der beste Wert um das Zentrum zu beschreiben?

d) Gibt es Ausreisser in diesen Daten?

- e) Je nach Situation werden Ausreisser aus den Daten entfernt, da diese offensichtlich falsch sind aufgrund Messfehler oder falscher Berechnungen. Wenn Ausreisser jedoch korrekt gemessen wurden sollten diese aus den Daten entfernt werden bevor man irgend ein Mass der Mitte berechnet?

2.5: Ein Basketballteam erspielte 43, 55, 41 und 37 Punkte in den ersten vier Spielen.

- a) Was ist der Mittelwert der Punkte in den ersten vier Spielen?
- b) Welchen Punktwert braucht das Team im nächsten Spiel damit Sie denselben Mittelwert halten können?
- c) Das Team erspielt nur 25 Punkte im fünften Spiel. Was ist der Mittelwert der Punkte in den ersten fünf Spielen?
- d) Das Team erspielt 41 Punkte im sechsten und dem finalen Spiel. Erhöht oder verkleinert das den Mittelwert für die sechs Spiele? Was ist der Mittelwert aller sechs Spiele?

2.6: Der Mittelwert von 10 Spielen ist 11.6. Was ist die Summe aller Punkte?

2.7: Während einer Safari fährt Bill durchschnittlich 262 km am Tag während 12 Tagen. Wie weit ist Bill während der Safari gefahren?

2.8: Der Mittelwert der Einnahmen eines Kleidergeschäfts für einen Monat beträgt CHF 15467. Berechnen Sie den Gesamtgewinn für ein Jahr.

2.9: Berechnen Sie x wenn 5, 9, 11, 12, 13, 14, 17 und x einen Mittelwert von 12 haben.

2.10: Berechnen Sie a wenn 3, 0, a , a , 4, a , 6, a und 3 einen Mittelwert von 4 haben.

2.11: Während einem Semester erreichte Aruna einen Durchschnitt von 35 aus 40 Punkten. Als Sie die einzelnen Tests suchte fand Sie nur noch 7 der 8 Tests. In diesen erreichte Sie 29, 36, 32, 38, 35, 34 und 39 Punkte. Berechnen Sie die Punktzahl im achten Test.

2.12: Eine Stichprobe von 10 Messwerten hat ein Mittelwert von 15.7 und eine Stichprobe von 20 Messwerten hat einen Mittelwert von 14.3. Was ist der Mittelwert aller 30 Messwerte gemeinsam?

2.13: Jana hatte sieben Wörtchenproben, jede mit zwölf Worten. Sie findet nur noch die Resultate von fünf dieser Proben: 9, 5, 7, 9 und 10. Sie fragte den Lehrer für die anderen zwei Resultate und der Lehrer antwortete mit: Der Modus der Punkte sei 9 und der Mittelwert ist 8. Was sind die beiden fehlenden Werte wenn Jana weiss, dass das schlechteste Resultat 5 war?

2.14: Die Tabelle enthält die Resultate wenn drei Münzen gleichzeitig geworfen werden, 30 mal.

Anzahl Kopf	Häufigkeit
0	4
1	12
2	11
3	3
Total	30

Berechnen Sie:

a) Den Modus

b) Den Mittelwert

c) Den Median.

2.15: Die folgende Häufigkeitstabelle zeigt die Anzahl Telefonanrufen während einem Tag, gemacht von 50 Fünfzehnjährigen.

Anzahl Anrufe	Häufigkeit
0	5
1	8
2	13
3	8
4	6
5	3
6	3
7	2
8	1
11	1

a) Berechnen Sie den Median, den Modus und den Mittelwert.

b) Zeichnen Sie ein Säulendiagramm und kennzeichnen Sie die drei Werte aus der vorherigen Aufgabe auf der x -Achse.

c) beschreiben Sie die Verteilung.

d) Weshalb ist der Mittelwert grösser als der Median für diese Daten?

e) Welches Mass der Mitte ist in diesem Fall das beste?

2.16: Eine Firma behauptet, dass ihre Zündholzschächtelchen, im Mittel, 50 Zündhölzer beinhalten. In einer Untersuchung wurden folgende Resultate gefunden:

Anzahl Zündhölzer	Häufigkeit
47	5
48	4
49	11
50	6
51	3
52	1
Total	30

a) Berechnen Sie den Mittelwert, den Median und den Modus.

b) Unterstützt die Untersuchung die Firma?

2.17: In einer Untersuchung wurden die Kinder in den Familien einer Schule gezählt.

Anzahl Kinder	Häufigkeit
1	5
2	28
3	15
4	8
5	2
6	1
Total	59

a) Berechnen Sie den Mittelwert, den Median und den Modus.

b) Im Mittel hat eine Familie 2.2 Kinder. Wie ist der Vergleich dieser Familien zum nationalen Durchschnitt?

c) Die Daten sind Schief; Links- oder Rechtsschief?

d) Wie hat diese Schiefeit die Masse der Mitte beeinflusst?

2.18: Die Immobilienpreise der letzten 10 Gebäude in einem Quartier waren:
CHF 146400, CHF 127600, CHF 211000, CHF 192500, CHF 256400,
CHF 132400, CHF 148000, CHF 129500, CHF 131400, CHF 162500

a) Berechnen Sie den Mittelwert und den Median.

b) Welchen Wert würden Sie verwenden, wenn Sie ihr Haus verkaufen möchten; Welchen wenn Sie ein Haus dort kaufen möchten?

2.19: In der Tabelle aufgelistet finden wir das Gewicht von Meerschweinchen (A-H) nach der Geburt und im Alter von 2 Wochen.

	Geburtsgewicht (g)	Gewicht nach 2 Wochen (g)
A	75	210
B	70	200
C	80	200
D	70	220
E	74	215
F	60	200
G	55	206
H	83	230

a) Was war der Durchschnitt bei der Geburt?

b) Was war der Durchschnitt nach 2 Wochen?

c) Was war der mittlere Zuwachs während den 2 Wochen?

2.20: 15 von 31 Messwerten sind unter 10 cm und 12 Messwerte sind über 11cm. Berechnen Sie den Median wenn die anderen 4 Messwerte 10.1 cm, 10.4 cm, 10.7 cm und 10.9 cm sind.

2.21: Zwei Zahnstocherproduzenten behaupten in den Boxen seien im Mittel 50 Zahnstocher. Eine Untersuchung lieferte folgende Daten:

Marke A

Anzahl in der Box	Häufigkeit
46	1
47	1
48	2
49	7
50	10
51	20
52	15
53	3
55	1

Marke B

Anzahl in der Box	Häufigkeit
48	3
49	17
50	30
51	7
52	2
53	1
54	1

a) Berechnen Sie den Mittelwert für Marke A und B.

b) Wäre es Fair aufgrund dieser Untersuchung einen der Produzenten zu bestrafen?

2.22: Gegen Ende der Saison erreicht ein Spieler in 14 Spielen einen Mittelwert von 16.5 Punkten. In den letzten beiden Spielen erspielte er noch 21 und 24 Punkte. Was ist der neue Mittelwert der Punkte für diesen Spieler?

2.23: Der Mittelwert und der Median einer Stichprobe von 9 Messwerten beträgt 12. Wenn 7 der Messwerte bekannt sind:
7, 9, 11, 13, 14, 17 und 19, berechnen Sie die anderen beiden Messwerte.

2.24: In einem Büro von 20 Personen gibt es 4 Lohnniveaus:

CHF 50000 (1 Person), CHF 42000 (3), CHF 35000 (6), CHF 28000 (10).

- a) Berechnen Sie: Den Median, den Modalwert und den Mittelwert.
- b) Welchen Wert verwendet der Chef um eine Lohnerhöhung abzuwenden?

2.25: 50 Studenten haben in einem Mathematiktest folgende Resultate erhalten:

Punkte	0-9	10-19	20-29	30-39	40-49
Häufigkeit	2	5	7	27	9

Berechnen Sie den Mittelwert.

2.26: Die Tabelle stellt die verkaufte Benzinmenge in einem Tag der Tankstellen in einer Stadt.

Liter (L)	Häufigkeit
2000 to 2999	4
3000 bis 3999	4
4000 bis 4999	9
5000 bis 5999	14
6000 bis 6999	23
7000 bis 7999	16

- a) Wie viele Tankstellen nehmen an der Umfrage teil?
- b) Wie viel Benzin wurde verkauft?
- c) Berechnen Sie den Mittelwert an verkauftem Benzin an einem Tag.

2.2 Streuungsmasse

Der Mittelwert reicht im Allgemeinen nicht aus, um die Verteilung von Messwerten ausreichend zu beschreiben. Wichtig ist auch, wie stark die Werte vom Mittelwert abweichen, wie sehr sie «streu»en. Liegen die Werte nahe beim Mittelwert, streuen sie schwach. Weichen die Werte deutlich vom Mittelwert ab, streuen sie stark.

2.2.1 Spannweite

Es gibt verschiedene Arten, die Streuung zu messen. Die naheliegendste ist die Differenz zwischen dem grössten und dem kleinsten Messwert, die sogenannte **Spannweite** oder **Variationsbreite**. Da nur zwei Messwerte herangezogen werden, ist die Spannweite sehr grob und wird deshalb selten verwendet.

2.2.2 Mittlere absolute Abweichung

Eine nächste Möglichkeit zur Messung der Streuung ist die durchschnittliche absolute Abweichung vom Mittelwert. Dazu werden sämtliche Abweichungen der Messwerte x_i von $\bar{x}$ im Betrag addiert und durch die Anzahl der Messwerte dividiert.

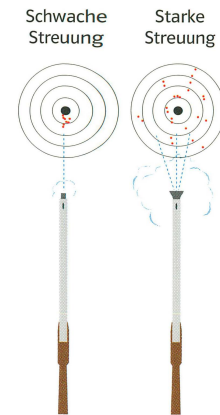


Abb. 2.2.1

Die mittlere absolute Abweichung s_a einer Stichprobe $x_1; x_2; \dots; x_n$ wird berechnet durch:

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

Beispiel 7:

Die Tabelle zeigt den Zeitaufwand in Minuten von Erwerbstätigen für den Arbeitsweg.

28	73	56	44	23	23	41	51	7	12
21	0	47	63	21	71	30	11	18	23
96	8	31	27	86	45	66	19	55	57

Bestimmen Sie den Mittelwert und die mittlere absolute Abweichung.

$$\bar{x} = \frac{28 + 73 + 56 + \dots + 57}{30} = 38.4\bar{3} \text{ min}$$

$$s_a = \frac{|28 - 38.4\bar{3}| + |73 - 38.4\bar{3}| + |56 - 38.4\bar{3}| + \dots + |57 - 38.4\bar{3}|}{30} = 20.86 \text{ min}$$

2.2.3 Die Quartile und der Interquartilsabstand

Der Median unterteilt die geordneten Daten in zwei gleich grosse Hälften. Diese Hälften sind erneut halbiert durch die **Quartile**.

Der mittlere Wert in der unteren Hälfte wird als **unteres Quartil** oder **25 Perzentil** bezeichnet. Ein Viertel, oder 25% der Daten sind kleiner oder gleich dem unteren Quartil. 75% der Daten sind grösser oder gleich dem unteren Quartil.

Der mittlere Wert in der oberen Hälfte wird als **oberes Quartil** oder 75 Perzentil bezeichnet. Ein Viertel oder 25% der Daten sind grösser oder gleich dem oberen Quartil. 75% der Daten sind kleiner oder gleich dem oberen Quartil.

Der **Interquartilsabstand (IQA)** ist die Spannweite der mittleren Hälfte (50%) der Daten.

$$\text{Interquartilsabstand} = \text{oberes Quartil} - \text{unteres Quartil}$$

Die Daten sind daher in viertel unterteilt durch das untere Quartil (Q_1), den Median (Q_2), und das obere Quartil (Q_3).

Daher ist der Interquartilsabstand $IQA = Q_3 - Q_1$.

Beispiel 8:

Daten: 6, 4, 7, 5, 3, 4, 2, 6, 5, 7, 5, 3, 8, 9, 3, 6, 5

Berechnen Sie: Den Median, das untere Quartil, das obere Quartil, und den IQA

Die geordneten Daten: 2 3 3 3 4 4 5 5 5 5 6 6 6 7 7 8 9 (17 Werte)

Der Median ist der 9te Wert, also Median = 5.

Da der Median ein Datenwert ist ignorieren wir diesen nun und halbieren die verbleibenden Werte in zwei Teile:

unterer Teil: 2 3 3 3 4 4 5 5 $Q_1 = \text{Median der unteren Hälfte} = \frac{3+4}{2} = 3.5$
oberer Teil: 5 6 6 6 7 7 8 9 $Q_3 = \text{Median der oberen Hälfte} = \frac{6+7}{2} = 6.5$

$$IQA = Q_3 - Q_1 = 6.5 - 3.5 = 3$$

Beispiel 9:

Daten: 11, 6, 7, 8, 13, 10, 8, 7, 5, 2, 9, 4, 4, 5, 8, 2, 3, 6

berechnen Sie: Den Median, unteres Quartil, oberes Quartil und den IQA

Geordnete Daten: 2 2 3 4 4 5 5 6 6 7 7 8 8 8 9 10 11 13 (18 Werte)

Der Median ist $\frac{6+7}{2} = 6.5$

Der Median ist kein Datenwert. Wir halbieren die Werte in zwei Teile:

unterer Teil: 2 2 3 4 4 5 5 6 6 $Q_1 = 4$
oberer Teil: 7 7 8 8 8 9 10 11 13 $Q_3 = 8$

$$IQA = Q_3 - Q_1 = 8 - 4 = 4$$

Box-and-whisker plots

Ein **box-and-whisker plot** (oder einfacher ein **Boxplot**) ist eine Visualisierung einiger statistischer Werte. Es werden eingezeichnet:

- Der kleinste Wert
- Das untere Quartil Q_1
- Der Median Q_2
- Das obere Quartil Q_3
- Der maximale Wert

Diese fünf Zahlen bezeichnet man auch als **five-number summary** einer Datensammlung.

Die Daten 11, 6, 7, 8, 13, 10, 8, 7, 5, 2, 9, 4, 4, 5, 8, 2, 3, 6 ergeben folgendes five-number summary und Boxplot:

$Minimum = 2$, $Q_1 = 4$, $Median = 6.5$, $Q_3 = 8$, $Maximum = 13$.

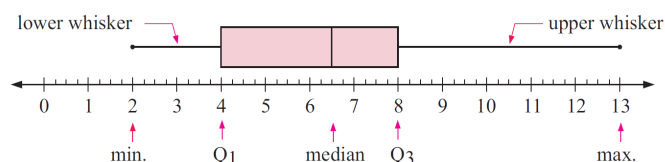


Abb. 2.2.2: Boxplot

Die rechteckige Box repräsentiert die mittlere Hälfte der Daten.

Der untere Strich (whisker) repräsentiert das 25. Perzentil mit dem kleinsten Wert.

Der obere Strich repräsentiert das 75. Perzentil mit dem Maximum.

Perzentile

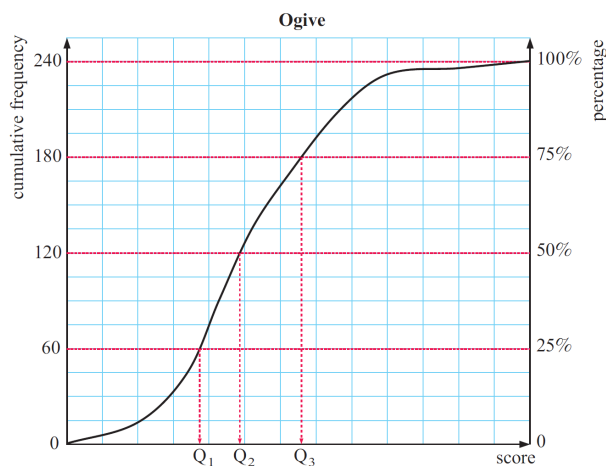
Ein Perzentil ist die Anzahl Werte, welche kleiner als der gegebene Prozentwert sind.

Beispiel:

- Das 85. Perzentil ist die Anzahl an Werten, welche kleiner oder gleich dem Wert bei 85% sind.
- Wenn Sie in einem Test einen 95. Perzentil erreichen, bedeutet das, dass 95 Prozent ihrer Klassenkameraden weniger Punkte als Sie gemacht haben.

Beachten Sie:

- Das untere Quartil (Q_1) ist das 25. Perzentil
- Der Median (Q_2) ist das 50. Perzentil
- Das obere Quartil (Q_3) ist das 75. Perzentil.



Eine Möglichkeit diese Perzentile zu zeigen ist, dass man eine separate y -Achsenbezeichnung zu einer Häufigkeitsverteilungskurve hinzufügt.

Wobei das Maximum mit 100% bezeichnet wird. Mit dieser Perzentil-Skala können wir die Quartile finden.

Übungen: Boxplot, Perzentile

2.27: Für die folgenden Stichproben berechnen Sie den Median, das obere und untere Quartil, die Spannweite und den IQA.

a) 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9

b) 10, 12, 15, 12, 24, 18, 19, 18, 18, 15, 16, 20, 21, 17, 18, 16, 22, 14

c) 21.8, 22.4, 23.5, 23.5, 24.6, 24.9, 25, 25.3, 26.1, 26.4, 29.5

2.28: Die Zeit welche 20 Menschen jeweils in der Schlange eine Bank verbrachten:

3.4 2.1 3.8 2.2 4.5 1.4 0 0 1.6 4.8 1.5 1.9 0 3.6 5.2 2.7 3.0 0.8 3.8 5.2

a) Berechnen Sie den Median für die Wartezeit und das oberen und untere Quartil.

b) Berechnen Sie die Spannweite und den IQA für die Wartezeit.

c) Vervollständigen Sie folgende Aussagen:

50% der Wartezeiten waren grösser als..... Minuten.

75% der Wartezeiten waren kleiner als Minuten.

Die minimale Wartezeit betrug Minuten und die Maximale Wartezeit war Minuten. Die Wartezeiten sind verteilt über Minuten.

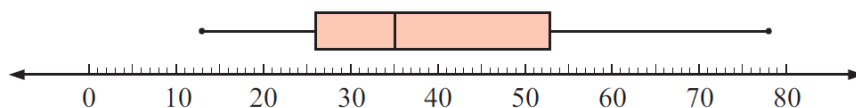
2.29: Betrachten Sie erneut das Erbsenproblem auf Seite 1.

a) für die Daten ohne Dünger berechnen Sie die Spannweite, den Median, das untere und obere Quartil und den IQA.

b) Berechnen Sie dasselbe für die Daten mit Dünger

c) Betrachten Sie die Fragen auf Seite 2 erneut und ergänzen Sie wo nötig ihre damaligen Resultate.

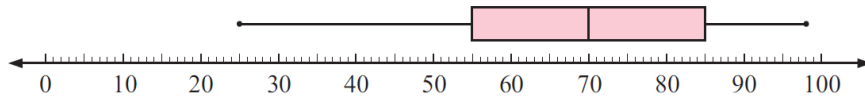
2.30: Der Boxplot gibt eine Auflistung der Punkte welche von einem Basketballteam erspielt wurden.



a) Finden Sie den Median, das Maximum, das Minimum, das obere und untere Quartil

b) Berechnen Sie den IQA

2.31: Der Boxplot zeigt die Resultate einer Klasse in einem Test mit möglichen 100 Punkten. Ergänzen Sie folgende Aussagen:

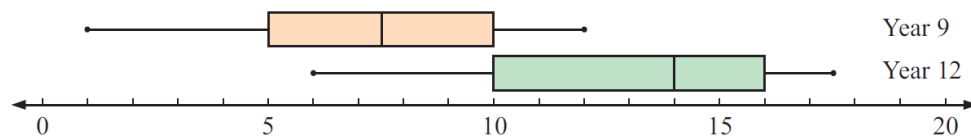


- Der höchste Testwert war ,und der tiefste
- Die Hälfte der Klasse erreichte eine Punktzahl von grösser als
- Die besten 25% der Klasse erreichten mindestens Punkte in diesem Test.
- Die mittlere Hälfte der Klasse erreichte Punkte zwischen und in diesem Test.
- Berechnen Sie die Spannweite der Daten.
- Berechnen Sie den IQA.
- Schätzen Sie die durchschnittliche Punktzahl für diesen Test.

2.32: Für die folgenden Daten zeichnen Sie einen Boxplot, finden Sie die 5-number summary und berechnen Sie den IQA

- 3, 5, 5, 7, 10, 9, 4, 7, 8, 6, 6, 5, 8, 6
- 3, 7, 0, 1, 4, 6, 8, 8, 8, 9, 7, 5, 6, 8, 7, 8, 8, 2, 9

2.33: Die folgenden Boxplots vergleichen die Zeit welche Studenten in der Quarta und Prima für Hausaufgaben verwendeten.



a) Ergänzen Sie:

Statistik	Quarta	Prima
Minimum		
Q_1		
Median		
Q_3		
Maximum		

- Berechnen Sie die Spannweite und den IQA.

- c) Wahr oder falsch:
Im Durchschnitt, setzt ein Primaner etwa doppelt so viel Zeit für Hausaufgaben ein wie ein Quartaner.
Über 25% der Quartaner setzen weniger Zeit für Hausaufgaben ein als als Primaner.

2.34: Martin untersucht eine neue Züchtung von Bohnen und zählt die Anzahl Bohnen in 33 Hülsen. Die Resultate sind:

5, 8, 10, 4, 2, 12, 6, 5, 7, 7, 5, 5, 5, 13, 9, 3, 4, 4, 7, 8, 9, 5, 5, 4, 3, 6, 6, 6, 6, 9, 8, 7, 6

- a) Berechnen Sie den Median, das untere und obere Quartil.

- b) Berechnen Sie den IQA

- c) Zeichnen Sie einen Boxplot.

2.35: Paul zählt die Bolzen in einigen Kisten und erhält folgende Daten:

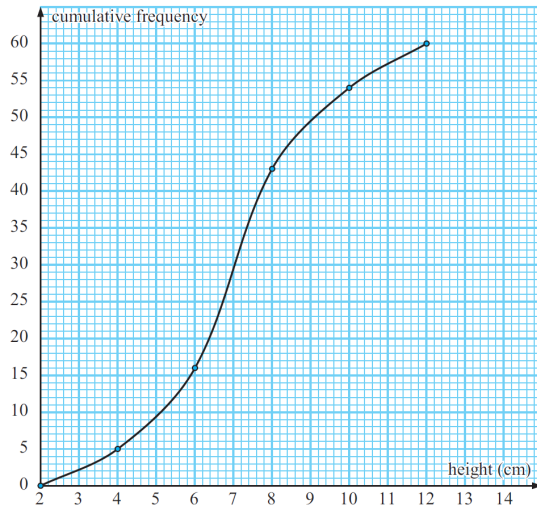
Anzahl Bolzen	Häufigkeit
33	1
34	5
35	7
36	13
37	12
38	8
39	0
40	1

- a) Berechnen Sie die five-number summary.

- b) Berechnen Sie die Spannweite und den IQA.

- c) Zeichnen Sie einen Boxplot.

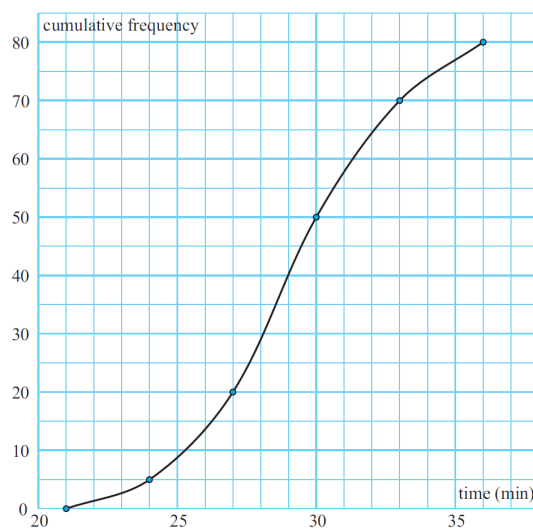
2.36: Ein Gärtner misst die Höhen von 60 Setzlingen und fasst die Daten in einer Häufigkeitsverteilungskurve zusammen:



- Wie viele Setzlinge haben eine Höhe von 5 cm oder weniger?
- Welchen Prozentsatz der Setzlinge sind grösser als 8 cm?
- Was ist der Median der Höhe?
- Was ist der IQA für diese Höhen?
- Berechnen Sie das 90 Perzentil und erklären Sie was diese Zahl bedeutet.

2.37: Die folgende Häufigkeitsverteilungskurve stellt die Zeiten von 80 Teilnehmenden an einem Rennen dar.

Cross-country race times



- Berechnen Sie das untere Quartil
- Berechnen Sie den Median

c) Berechnen Sie das obere Quartil

d) Berechnen Sie den IQA

e) Berechnen Sie das 40 Perzentil.

2.2.4 Varianz und Standardabweichung

Das Problem wenn wir die Spannweite oder den IQA als Mass für die Streuung verwenden ist, dass beide nur jeweils zwei Werte in die Berechnung einbeziehen. Einige Datensätze könnten so das wahre Aussehen verbergen wenn wir nur die Spannweite und den IQA verwenden würden. Wir brauchen also noch eine bessere Möglichkeit die Streuung zu berechnen.

Betrachten wir den Datensatz mit n Werten: $x_1, x_2, x_3, x_4, \dots, x_n$, mit Durchschnitt $\bar{x}$.

$x_i - \bar{x}$ misst wie weit weg der Wert x_i vom Mittelwert liegt, daher könnte man denken, dass die Summe dieser Differenzen $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})$ uns ein gutes Mass für die Streuung gibt. Leider ist dieser Wert immer Null.

Dafür verwenden wir die Formel

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

und nennen diesen Wert die **Varianz** der n Datenwerte.

Erläuterungen zur Berechnung der Varianz:

1. $(x_i - \bar{x})^2$ ist auch ein Mass dafür wie weit der Wert x_i vom Mittelwert $\bar{x}$ entfernt ist. Aber durch das quadrieren erreichen wir, dass jeder dieser Werte positiv ist. Daher ist die Summe auch nicht mehr gleich Null.
2. Wenn $\sum_{i=1}^n (x_i - \bar{x})^2$ klein ist, ist es ein Zeichen, dass die meisten Werte nahe an $\bar{x}$ liegen.
3. Durch das dividieren mit n erhalten wir ein Mass dafür, wie weit im Durchschnitt jeder einzelne Wert vom Mittelwert abweicht.

Für einen Datensatz mit n Werten wird

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

als **Standardabweichung** bezeichnet.

Erläuterung zur Berechnung der Standardabweichung:

Die Wurzel bei der Standardabweichung wird benötigt um eine Sinnvolle Einheit zu erhalten. Beispielsweise, wenn x_i das Gewicht eines Studenten in kg angibt, dann ist s^2 in kg^2 gemessen. Daher ist die Standardabweichung häufiger verwendet als die Varianz.

Die Standardabweichung ist kein resistentes Mass der Streuung. Das gilt weil die Standardabweichung alle Datenwerte mit einbezieht und auf dem Mittelwert beruht. Auch geben weit entfernte Werte oder Ausreisser aufgrund des 2 grosse Werte. Die Standardabweichung ist ein hilfreiches Mass wenn die Verteilung einigermaßen Symmetrisch ist. Ausserordentlich hilfreich ist die Standardabweichung wenn die Daten Normalverteilt sind. Das betrachten wir aber später.

Der IQA und die Perzentile sind besser geeignet wenn die Daten nicht symmetrisch verteilt sind.

Beispiel 10:

Ein Lebensmittelgeschäft möchte Äpfel von zwei verschiedenen Lieferanten kaufen. Sie nehmen je sechs zufällige Stichproben à 50 Äpfel, um diese auf Makel zu untersuchen.

Die Auswertung ergab folgendes für die Anzahl Äpfel mit Makeln:

Lieferant A: 5 17 15 3 9 11

Lieferant B: 10 13 12 11 12 11

Finden Sie den Mittelwert und die Standardabweichung für die Stichproben. Was sagen uns diese Werte?

Lösung:

Lieferant A:

x	$x - \bar{x}$	$(x - \bar{x})^2$
5	-5	25
17	7	49
15	5	25
3	-7	49
9	-1	1
11	1	1
60	Total	150

$$\bar{x} = \frac{60}{6}$$

$$s_n = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{150}{6}} = 5$$

Lieferant B:

x	$x - \bar{x}$	$(x - \bar{x})^2$
10	-1.5	2.25
13	1.5	2.25
12	0.5	0.25
11	-0.5	0.25
12	0.5	0.25
11	-0.5	0.25
69	Total	5.5

$$\bar{x} = \frac{69}{6}$$

$$s_n = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{5.5}{6}} = 0.957$$

Der Lieferant B liefert Äpfel mit etwas mehr Makel, mit kleinerer Varianz als Lieferant A.

2.3 Stichproben aus einer Population

Populationen sind häufig sehr gross und das Sammeln aller Daten ist daher sehr zeitaufwändig und kostenintensiv.

Daher wählt man oft eine **zufällige Stichprobe** mit der Hoffnung, dass die gesamte Population durch die Stichprobe charakterisiert wird. Um sicherzugehen muss die Stichprobe genügend gross sein und ohne Verfälschung gewählt werden.

Von der Stichprobe möchten wir einige Aussagen über die gesamte Population machen können, wie Mittelwert, Standardabweichung usw.

Um zu unterscheiden ob wir von einer Stichprobe oder der gesamten Population sprechen verwenden wir verschiedene Notationen für den Mittelwert, die Varianz und Standardabweichung. Die Schreibweisen sind in folgender Tabelle zusammengefasst.

	Stichprobe	Population
Mittelwert	$\bar{x}$	μ
Varianz	s_n^2	σ^2
Standardabweichung	s_n	σ

Wenn wir die statistischen Kennzahlen einer Stichprobe haben, können wir daraus bestmögliche Annäherungen für die Population machen. Die folgenden Resultate nehmen wir ohne Beweis zur Kenntnis:

Hat eine Stichprobe den Umfang n und wird daraus auf die Population geschlossen gilt:

- Der Durchschnitt der Stichprobe $\bar{x}$ ist ein biasfreier Wert um μ zu schätzen
- $s_{n-1}^2 = \left(\frac{n}{n-1}\right) s_n^2$ ist ein biasfreier Wert um die Varianz σ^2 zu schätzen.

Beachten Sie: Auch wenn s_{n-1}^2 ein biasfreier Wert für σ^2 ist, dürfen wir nicht daraus schliessen dass auch s_{n-1} ein guter Wert für σ ist.

Bemerkungen über Parameter und Statistiken

Ein **Parameter** ist ein Zahlenwert um eine Population zu beschreiben.

Eine **Statistik** ist ein Zahlenwert um eine Stichprobe zu beschreiben.

Zum Beispiel: Wenn wir den Mittelwert des Alters in den Alterssiedlungen in der Schweiz untersuchen, dann ist der gefundene Wert ein Parameter. Wenn wir eine zufällige Stichprobe von 300 Personen in diesen Alterssiedlungen nehmen und daraus den Durchschnitt berechnen, dann haben wir eine Statistik.

Beispiel 11:

Eine zufällige Stichprobe von 48 Schafen wurde aus einer Herde von 2000 Schafen gewählt. Der Durchschnitt des Gewichtes beträgt 48.6 kg und die Varianz 17.5 kg².

1. Berechnen Sie die Standardabweichung der Stichprobe.
2. Finden Sie eine möglichst genaue Näherung für das durchschnittliche Gewicht der gesamten Herde.
3. Berechnen Sie eine möglichst genaue Näherung für die Varianz der gesamten Herde.

1. $s_n = \sqrt{\text{Varianz}} = \sqrt{17.5} \approx 4.18 \text{ kg}$

2. μ wird angenähert durch $\bar{x} = 48.6 \text{ kg}$

3. σ^2 wird angenähert durch $s_{n-1}^2 = \left(\frac{n}{n-1}\right) s_n^2 \approx \frac{48}{47} \cdot 17.5 \approx 17.9 \text{ kg}^2$.

Standardabweichung für Daten in Klassen

Für Daten in Klassen gilt für die Standardabweichung:

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}}$$

wobei x **irgend ein Wert** ist, $\bar{x}$ ist der **Durchschnitt** und f ist die **Häufigkeit** des Auftretens.

Beispiel 12:

Berechnen Sie die Standardabweichung der folgenden Verteilungen:

Wert	1	2	3	4	5
Häufigkeit	1	2	4	2	1

Lösung:

x	f	fx	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
1	1	1	-2	4	4
2	2	4	-1	1	2
3	4	12	0	0	0
4	2	8	1	1	2
5	1	5	2	4	4
Total	10	30			12

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{30}{10} = 3$$

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} = \sqrt{\frac{12}{10}} \approx 1.10$$

Einsatz des Taschenrechners

Bei grossen Datensätzen macht es Sinn, den TR als Hilfsmittel zu verwenden. Am besten wird die Vorgehensweise anhand von Beispielen klar.

Daten in einer Häufigkeitstabelle

Beispiel 13:

Die Messung der Grösse (in cm) von 103 Schülern wurde in einer Strichliste notiert:

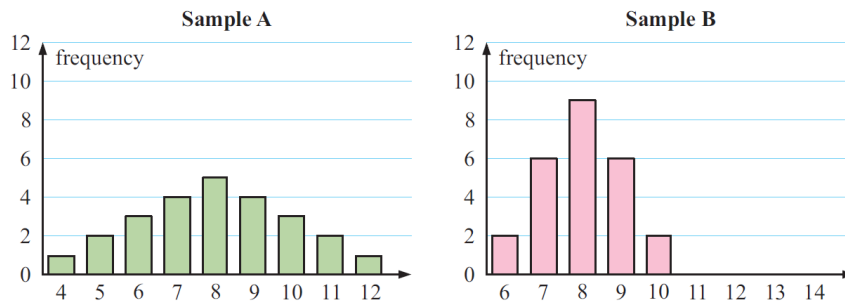
161	166		171		176		181		186	
162	167		172		177		182		187	
163		168		173		178		183		
164		169		174		179		184		189
165		170		175		180		185		190

Berechnen Sie μ , $\bar{x}$, den Median, σ , s , Q_1 , Q_3 usw...

Die Daten werden unter Lists & Spreadsheet in eine Tabelle eingegeben. Anschliessend wird ein neues Calculator-File erstellt und mit **menu** **6** **1** **1** werden die wichtigsten statistischen Werte abgefragt.

Übungen: Varianz und Standardabweichung

2.38: Das Säulendiagramm zeigt zwei Verteilungen:



- Welche der Verteilungen hat die grössere Spannweite?
- Berechnen Sie den Durchschnitt beider Stichproben.
- Berechnen Sie die Standardabweichung für beide Stichproben

2.39: Die erzielten Punkte von Andreas und Thomas in den letzten 8 Spielen sind aufgelistet:

Andreas	Thomas
23	9
17	29
31	41
25	26
25	14
19	44
28	38
32	43

- Berechnen Sie den Mittelwert und die Standardabweichung für beide Spieler.
- Welcher der beiden Spieler ist konstanter?

2.40: Zwei Baseball Coachs vergleichen die Punkte ihrer Teams in den letzten 10 Spielen:

Rockets	Bullets
0	4
10	3
1	4
9	1
11	4
0	11
8	7
5	6
6	12
7	5

- Zeigen Sie, dass beide Teams denselben Durchschnitt und dieselbe Spannweite haben.
- Welches Team spielt weniger konsistent?

- c) Welcher Wert gibt den bessern Hinweis auf Konsistenz? die Spannweite oder die Standardabweichung?

2.41: Ein Hersteller von Süssgetränken beschäftigt einen Statistiker zur Qualitätskontrolle. Er muss sicherstellen, dass 375 mL in jede Flasche gefüllt werden. Die Maschine welche die Flaschen füllt könnte kleine Abweichungen produzieren durch Vibrationen oder andere Faktoren.

- a) Erwarten Sie, dass die Standardabweichung aller abgefüllter Flaschen an einem Tag, bzw. in einer Woche derselbe Wert ergibt?
- b) An jedem Tag werden 125 Flaschen zufällig gewählt. Was würden Sie verwenden um:
den durchschnittliche Füllmenge zu bestimmen
die Variabilität der Füllmenge zu bestimmen?
- c) Was bedeutet eine kleine Standardabweichung in diesem Beispiel?

2.42: Das Gewicht von sieben Footballspielern ist: 79, 64, 59, 71, 68, 68 und 74.

- a) Berechnen Sie den Mittelwert und die Standardabweichung für diese Gruppe.
- b) Erstaunlicherweise haben alle Spieler in den letzten fünf Jahren genau 10 kg zugenommen. Berechnen Sie den neuen Mittelwert und die neue Standardabweichung.

2.43: Das Gewicht von zehn Truthähnen gerundet auf 0.1 kg ist:
0.8, 1.1, 1.2, 0.9, 1.2, 1.2, 0.9, 0.7, 1.0, 1.1

- a) Berechnen Sie den Mittelwert und die Standardabweichung.
- b) Nach einem Monat Spezialdiät verdoppelt sich das Gewicht.
Berechnen Sie erneut den Mittelwert und die Standardabweichung.
- c) Was stellen Sie aus Aufgabe a) und b) im allgemeinen fest?

2.44: Eine Stichprobe von 8 natürlichen Zahlen hat den Mittelwert 5 und die Varianz 5.25.
Die Zahlen sind: 1, 3, 5, 7, 4, 5, p , q . Berechnen Sie p und q wenn $p < q$ gilt.

2.45: Eine Stichprobe von 10 natürlichen Zahlen hat als Mittelwert 6 und als Varianz 3.2.
Die Zahlen sind: 3, 9, 5, 5, 6, 4, a , 6, b , 8. Berechnen Sie a und b wenn $a > b$ gilt.

2.46: Die folgende Tabelle zeigt die Änderung im Cholesterinlevel bei sechs Freiwilligen nach zwei Wochen Diät und Sport:

Freiwilliger	A	B	C	D	E	F
Änderung im Cholesterin	0.8	0.6	0.7	0.8	0.4	2.8

- Berechnen Sie die Standardabweichung.
- Berechnen Sie die Standardabweichung erneut, nachdem Sie den Aussenseiter entfernt haben.
- Was ist der Einfluss eines Extremwertes auf die Standardabweichung?

2.47: In einer zufälligen Stichprobe von 87 Kühen aus einer riesigen Herde hat jedes Tier im durchschnitt das Gewicht 93.8 kg mit einer Varianz von 45.9 kg².

- Berechnen Sie die Standardabweichung der Stichprobe.
- Geben Sie biasfreie Schätzungen für den Mittelwert und die Varianz der gesamten Herde an.

2.48: Das Gewicht (in Gramm) einer zufälligen Stichprobe von Spatzen ist:
87 75 68 69 81 89 73 66 91 77 84 83 77 74 80 76 67

- Berechnen Sie den Mittelwert und die Standardabweichung der Stichprobe.
- Berechnen Sie biasfreie Schätzungen für den Mittelwert und die Varianz der gesamten Spatzenpopulation.

2.49: John fährt jeden Tag zum Strand. An 16 zufällig gewählten Tagen misst er seine Zeit x_i in Minuten.
Er berechnet: $\sum_{i=1}^{16} x_i = 519$ und $\sum_{i=1}^{16} (x_i^2) = 16983$. Berechnen Sie daraus biasfreie Werte für den Mittelwert und die Varianz aller Fahrten zum Strand.

2.50: Gegeben ist eine Stichprobe von Familiengrößen welche zufällig gewählt wurde.

Anzahl Kinder	0	1	2	3	4	5	6	7
Häufigkeit	14	18	13	5	3	2	2	1

- Berechnen Sie für die Stichprobe den Mittelwert und die Standardabweichung.
- Berechnen Sie biasfreie Schätzungen für den Mittelwert und die Varianz für die gesamte Population.

2.51: Gegeben ist eine zufällige Stichprobe mit dem Alter von Squash Spielern in der Junior Nationalmeisterschaft.

Alter	11	12	13	14	15	16	17	18
Häufigkeit	2	1	4	5	6	4	2	1

- a) Berechnen Sie für die Stichprobe den Mittelwert und die Standardabweichung.
- b) Berechnen Sie biasfreie Schätzungen für den Mittelwert und die Varianz für alle Junioren.

2.52: Die Anzahl an Zahnstochern in einer zufälligen Stichprobe von 48 Boxen wurden gezählt:

Anzahl	33	35	36	37	38	39	40
Zahnstocher							
Häufigkeit	1	5	7	13	12	8	2

- a) Berechnen Sie den Mittelwert und die Standardabweichung für die Anzahl Zahnstocher in den Boxen.
- b) Berechnen Sie biasfreie Schätzungen für den Mittelwert und die Varianz aller produzierter Boxen.

2.53: Die Länge von 30 zufällig gewählten 12 Tage alten Kleinkindern wurde gemessen und auf den nächsten cm gerundet:

Länge (cm)	Häufigkeit
40 - 41	1
42 - 43	1
44 - 45	3
46 - 47	7
48 - 49	11
50 - 51	5
52 - 53	2

- a) Berechnen Sie den Mittelwert und die Standardabweichung für die Längen.
- b) Berechnen Sie biasfreie Schätzungen für den Mittelwert und die Varianz aller Kleinkinder.

2.54: Der Wochenlohn von 200 zufällig gewählten Studenten beträgt:

Lohn	Anzahl
360 - 369.99	17
370 - 379.99	38
380 - 389.99	47
390 - 399.99	57
400 - 409.99	18
410 - 419.99	10
420 - 429.99	10
430 - 439.99	3

- a) Berechnen Sie den Mittelwert und die Standardabweichung für den Lohn.
- b) Berechnen Sie biasfreie Schätzungen für den Mittelwert und die Varianz des Lohnes aller Studenten.

2.4 Die Standardabweichung bei normalverteilten Daten

Betrachtet man die Menge an Flüssigkeit in den Dosen eines Süssgetränkeherstellers. Die Verteilung der Flüssigkeitsmenge ist symmetrisch und Glockenförmig. Das liegt an der natürlichen Variation durch die Maschine welche auf eine gewisse Menge eingestellt wurde. Zufall und weitere Faktoren (Vibration, usw.) führen dazu, dass in etwa gleich viele Dosen etwas zu wenig wie zu viel Flüssigkeit enthalten.

Die resultierende glockenförmige Verteilung nennen wir Normalverteilung.

Wenn eine grosse Stichprobe einer Normalverteilten Zufallsvariable genommen wird, welcher Prozentwert der Daten liegt dann zwischen $\bar{x} - s$ und $\bar{x} + s$?

Es kann gezeigt werden, dass für alle messbaren Grössen einer Population welche Normalverteilt ist, unwichtig der Grössen des Mittelwertes oder der Standardabweichung:

- Ca 68% der Messwerte liegt zwischen der einfachen Standardabweichung in beide Richtungen des Mittelwertes
- Ca 95% der Messwerte liegt zwischen zwei Mal der Standardabweichung auf beide Seiten des Mittelwertes
- Ca 99.7% der Messwerte liegt zwischen drei Mal der Standardabweichung auf beide Seiten des Mittelwertes.

Die schöne Funktion, welche Normalverteilung heisst hat als Asymptote die x -Achse, also theoretisch gibt es keine Grenzen in welchen die Daten liegen müssen.

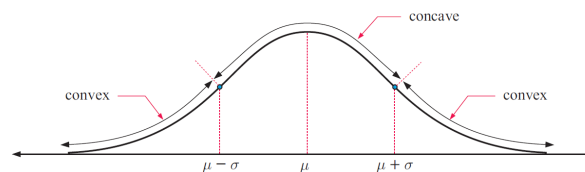


Abb. 2.4.1: Die Normalverteilung

Praktisch hingegen ist es sehr selten einen Wert zu finden welcher eine Distanz von mehr als 3 Standardabweichungen zum Mittelwert hat und noch viel seltener einen Wert zu finden welcher 5 Standardabweichungen weg vom Mittelwert liegt.

Beachten Sie, dass die Position von einer Standardabweichung vom Mittelwert genau die Wendepunkte der Kurve bezeichnen.

2.5 Regression

Bisher wurden nur statistische Erhebungen betrachtet, in denen ein Merkmal gemessen wurde. Oft interessieren aber zwei Merkmale eines Merkmalsträgers.

So werden z. B. bei einer ärztlichen Untersuchung sowohl das Alter als auch der Blutdruck gemessen. Bei einer Stichprobe von 20 Männern ergaben sich folgende Wertepaare.

Alter in Jahren	45	65	63	37	19	18	50	47	56	43	29	49	32	26	28	38	57	41	36	58
Blutdruck in mm Hg	155	192	199	157	139	105	134	135	170	151	137	128	140	115	129	117	179	127	118	158

Gibt es einen Zusammenhang zwischen dem Alter und dem Blutdruck; je älter, desto höher der Blutdruck oder etwa umgekehrt?

Mit der Regression wird das Ziel verfolgt, den mathematischen Zusammenhang zwischen zwei quantitativen Merkmalen zu finden und durch eine Gleichung zu beschreiben.

Wenn wir die Daten in ein Koordinatensystem eintragen, ergibt sich eine Punktwolke mit steigender Tendenz (Abb. 2.5.1). Um einen quantitativen Zusammenhang zwischen Alter und Blutdruck zu bestimmen, legen wir eine Gerade durch die Punktwolke, welche diese am besten approximiert. Die Gleichung dieser Geraden gibt dann einen linearen Zusammenhang zwischen Alter und Blutdruck.

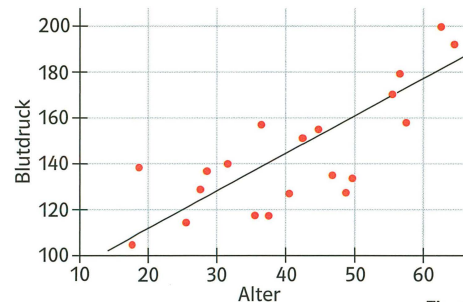


Abb. 2.5.1

Als Kriterium für die Approximation wählt man die Bedingung, dass die Summe der quadratischen Abweichungen minimal ist (Methode der kleinsten Quadrate). Das heißt, man bestimmt die Gerade $f(x) = mx + b$, für die der Ausdruck

$$\sum_{i=1}^n (f(x_i) - y_i)^2$$

am kleinsten wird. Grafisch bedeutet dies, dass die Summe der «Fehlerquadrate» (Abb. 2.5.2) minimal wird.

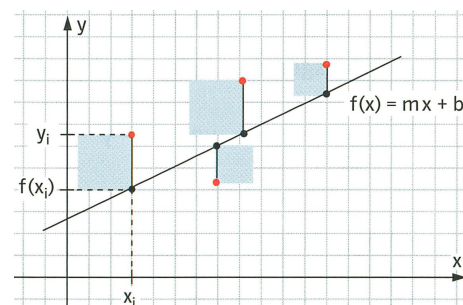


Abb. 2.5.2

Gegeben sind n Paare von Merkmalswerten $(x_1|y_1), (x_2|y_2), \dots, (x_n|y_n)$. Als Regressionsgerade bezeichnet man die Gerade $f(x) = mx + b$, für welche die Fehlerquadratsumme

$$F(m, b) = \sum_{i=1}^n (f(x_i) - y_i)^2 = \sum_{i=1}^n (mx_i + b - y_i)^2 \quad \text{am kleinsten wird.}$$

Die Bestimmung der Regressionsgeraden ist ein Optimierungsproblem das mit Hilfe der Differentialrechnung gelöst werden kann.

Bei n vorgegebenen Merkmalswerten $(x_i|y_i)$ mit den Mittelwerten $\bar{x}$ und $\bar{y}$ berechnen sich die Steigung m und der Achsenabschnitt b der Regressionsgeraden $f(x) = mx + b$ wie folgt:

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{und} \quad b = \bar{y} - m\bar{x}$$

Ein wichtiges Motiv für die Anpassung einer Geraden an eine Punktwolke besteht darin, nicht vorhandene oder zukünftige Werte vorhersagen zu können. Mithilfe der Regressionsgeraden können also Prognosen gemacht werden. Man sollte aber mit solchen Prognosen umso vorsichtiger umgehen, je weiter man in die Zukunft oder in unbekannte Bereiche vorstößt.

Auch für nicht lineare Beziehungen können die Parameter für die «beste» Ausgleichskurve durch Minimieren der mittleren quadratischen Abweichung gefunden werden. In vielen Anwendungen der Praxis reicht jedoch die Bestimmung der Regressionsgeraden.

Beispiel 14:

Die Tabelle zeigt die neu erstellten Wohnungen in der Schweiz seit 2004.

Jahr	2004	2005	2006	2007	2008	2009	2010
Anzahl neu erstellter Wohnungen	36 935	37 958	41 989	42 911	44 192	39 733	43 632

1. Berechnen Sie die Regressionsgerade ($x = \text{Jahr}$, $y = \text{Anzahl neu erstellter Wohnungen}$).
2. Basierend auf dem linearen Modell von Aufgabe a.), wie viele Wohnungen werden im Jahr 2020 erstellt?

-
1. Mittelwerte: $\bar{x} = 2007$; $\bar{y} = 41'050$ Regressionsgerade: $y = mx + b$

$$m = \frac{(2004-2007)(36935-41050)+(2005-2007)(37958-41050)+\dots+(2010-2007)(43632-41050)}{(2004-2007)^2+(2005-2007)^2+\dots+(2010-2007)^2}$$
$$= \frac{25844}{28} = 923$$

$$b = 41050 - 923 \cdot 2007 = -1811411 \quad \rightarrow \quad y = 923x - 1811411$$

2. $y = 923 \cdot 2020 - 1811411 = 53'049$ (Wohnungen)

2.6 Korrelation

Im vorherigen Kapitel wurde mithilfe der Regressionsgeraden ein Zusammenhang zwischen Alter und Blutdruck festgestellt. Es gibt auch ein Mass für die Richtung und Stärke dieses linearen Zusammenhangs. Dieses Mass zeigt, wie stark die Daten vom gewählten linearen Modell abweichen, d.h., wie gut das lineare Modell ist.

Die Berechnung der Stärke und Richtung des Zusammenhangs zwischen zwei Merkmalen beruht auf folgender Idee: Bildet man für ein Wertepaar $(x_i|y_i)$ das Abweichungsprodukt $(x_i - \bar{x})(y_i - \bar{y})$ vom Mittelwert, so ist dieses positiv, wenn x_i und y_i beide grösser oder beide kleiner als der Mittelwert $\bar{x}$ bzw. $\bar{y}$ sind. Der Punkt $(x_i|y_i)$ liegt somit in einem der positiven Bereiche von Abb. 2.6.1. Ist x_i grösser als $\bar{x}$ und y_i kleiner als $\bar{y}$ (oder umgekehrt), dann ist das Abweichungsprodukt negativ und der Punkt $(x_i|y_i)$ liegt in einem der negativen Bereiche.

Bildet man die Summe der einzelnen Abweichungsprodukte $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ man über die Richtung

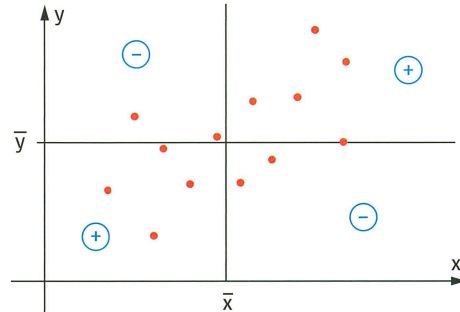


Abb. 2.6.1

des Zusammenhangs Folgendes feststellen:

Eine positive Summe resultiert, wenn die beiden Variablen weitgehend gemeinsam in die gleiche Richtung von ihrem Mittelwert abweichen. Es besteht eine Tendenz der Form «je grösser x , desto grösser y », der Zusammenhang ist positiv. Dagegen ergibt sich eine negative Summe, wenn viele entgegengesetzt gerichtete Abweichungen vom jeweiligen Mittelwert auftreten. Die Merkmale weisen einen negativen Zusammenhang auf: «je grösser x , desto kleiner y ». Sind die Abweichungen mal gleich, mal entgegengesetzt gerichtet, so heben sich die Abweichungsprodukte gegenseitig auf und es resultiert eine Summe nahe null. In diesem Fall besteht kein linearer Zusammenhang zwischen den Variablen x und y .

Um auch eine quantitative Kennzeichnung des Zusammenhangs zweier Merkmale zu erreichen, wird ein standardisiertes Mass eingeführt.

Bei n vorgegebenen Merkmalswerten $(x_i|y_i)$ mit den Mittelwerten $\bar{x}$ und $\bar{y}$ berechnet sich der Korrelationskoeffizient r wie folgt:

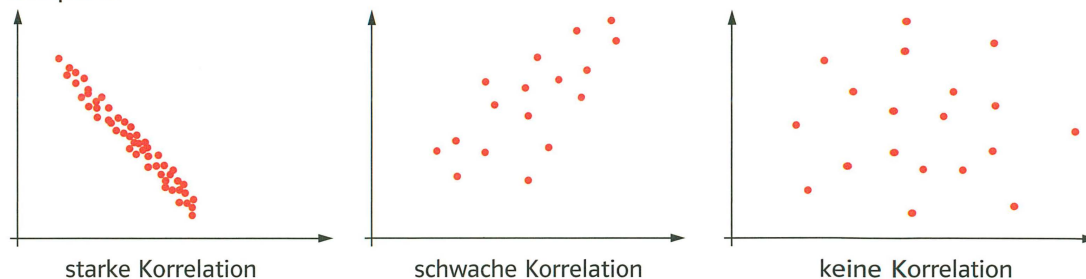
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Der Korrelationskoeffizient r misst die Stärke und Richtung des Zusammenhangs zwischen zwei Merkmalen x und y . Er hat folgende Eigenschaften:

1. Der Korrelationskoeffizient liegt immer zwischen -1 und 1 .
2. Der Korrelationskoeffizient ist positiv ($r > 0$), wenn die Regressionsgerade steigt, und negativ ($r < 0$), wenn sie fällt.
3. Die Korrelation, d. h. der Grad des linearen Zusammenhangs, lässt sich aus dem Korrelationskoeffizienten wie folgt bewerten:

$ r $	1	$[0,7; 1[$	$[0,3; 0,7[$	$]0; 0,3 [$	0
Korrelation	volle	starke	mittlere	schwache	keine

Beispiele:



2.6.1 Korrelation ist nicht Kausalität

Bei starker Korrelation kann ein ursächlicher Zusammenhang zwischen den beiden Merkmalen vorliegen. Dies ist z.B. der Fall, wenn man die Merkmale «Kraft» und «Verlängerung einer Feder» betrachtet.

Eine starke positive Korrelation liegt aber auch zwischen «der mittleren Tageslänge im Monat» und «der mittleren Temperatur im Monat» vor. Hier spricht die starke Korrelation nicht für einen ursächlichen Zusammenhang (die Tageslänge hängt nicht von der Temperatur ab!), sondern beide Merkmale werden in ihren Werten durch eine gemeinsame Ursache, den Sonnenstand, beeinflusst.

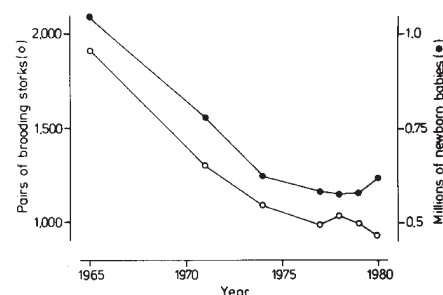
Bei der Betrachtung der Merkmale «Anzahl Störche» und «Anzahl Geburten» von 1965 bis 1975 in Schleswig-Holstein stellte man eine starke positive Korrelation fest. Ob hier eine gemeinsame Ursache vorliegt, kann nicht mathematisch, sondern nur von der Sache her entschieden werden.

Der hier eingeführte Korrelationskoeffizient macht nur Aussagen über den linearen Zusammenhang. Beim Fall $r = 0$ besteht kein linearer Zusammenhang, wie in der unregelmässigen Punktwolke von Abb. 2.6.3 links. Es kann aber sein, dass bei $r = 0$ ein nicht linearer Zusammenhang besteht, wie in Abb. 2.6.3 rechts. Würde man in diesem Fall als Modell eine Parabel wählen, dann wäre der zugehörige Korrelationskoeffizient nahe bei 1.

Fehlerhafte Schlüsse können auch gezogen werden, wenn man nicht bedenkt, dass der Korrelationskoeffizient im Allgemeinen zu gross wird, falls der Stichprobenumfang sehr klein ist. Betrachtet man z. B. nur zwei Paare von Merkmalswerten, so erhält man immer einen Korrelationskoeffizienten $r = 1$, weil die Regressionsgerade in diesem Fall immer genau durch die beiden Punkte verläuft.

A new parameter for sex education

SIR—There is concern in West Germany over the falling birth rate. The accompanying graph^{1,2} might suggest a solution that every child knows makes sense.



HELMUT SIES

*Institut für Physiologische Chemie 1,
Universität Düsseldorf,
Moorenstrasse 5, D-4000 Düsseldorf,
FRG*

1. *Fachserie Gebiet und Bevölkerung* (Statistisches Bundesamt, Kohlhammer, Stuttgart, 1984).
2. Bauer, S. & Thielcke, G. *Die Vogelwarte* 31, 183–191 (1982).

Abb. 2.6.2: Die Schein-Korrelation zwischen der Storchpopulation und der Geburtenrate in Deutschland

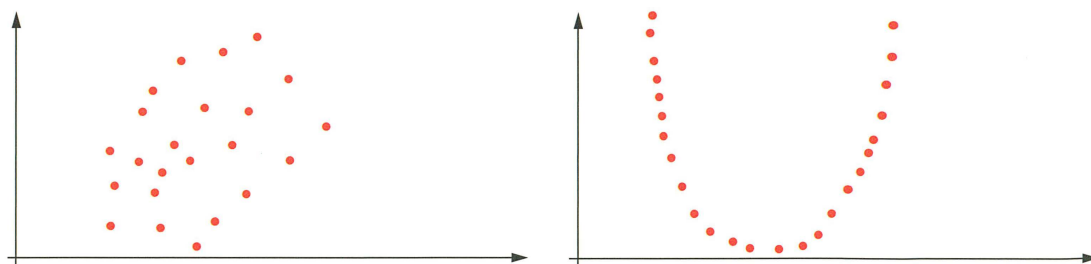


Abb. 2.6.3

2.6.2 Unterschied Regression - Korrelation

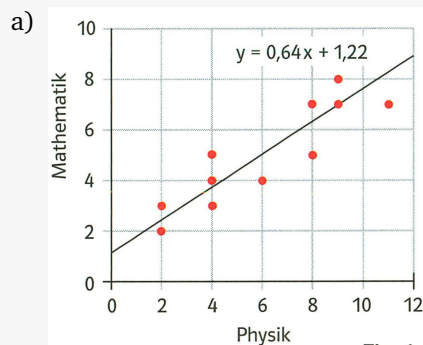
Beide untersuchen den Zusammenhang zwischen zwei Variablen. Bei der Korrelation sind die beiden Variablen gleichberechtigt, bei der Regression ist y von x abhängig. Dadurch erlaubt das Regressionsmodell Vorhersagen, über deren Qualität die Korrelation eine Aussage macht.

Beispiel 15:

An einer Schule wurde der Zusammenhang zwischen Mathematik- und Physikkenntnissen getestet. Die 12 Versuchspersonen erreichten in den jeweiligen Tests folgende Punktzahlen.

Person	1	2	3	4	5	6	7	8	9	10	11	12
Merkmal X (Physik)	2	8	4	9	9	4	8	4	6	2	11	10
Merkmal Y (Mathematik)	3	5	4	7	8	5	7	3	4	2	7	9

- Betrachten Sie für jede Person die Punktzahlen als Koordinatenpaar. Zeichnen Sie das Streudiagramm.
- Bestimmen Sie die Gleichung der Regressionsgeraden und zeichnen Sie diese.
- Welche Punktzahl in Mathematik kann man für eine Person prognostizieren, die in Physik 12 Punkte erreicht?
- Berechnen Sie den Korrelationskoeffizienten und beurteilen Sie die Güte der Anpassung an die Regressionsgerade.



b) Mit TR: $f(x) = 0.64x + 1.22$

c) Mit TR: $f(12) \approx 9$

d) Mit TR: $r = 0.9$
Dies ist eine sehr starke positive Korrelation.

2.7 Lösungen

1.1: a) Grössen können irgendeinen Wert annehmen.

c) Die Klasse 185-190 cm

1.2: a) stetig

c) 10-20 minuten

1.3: a) Säulendiagramm

b) Histogramm. Darstellung mit dem TR.

1.4: a) TR

b) 20

c) 58.33%

d) 1218 bzw. 512

2.1: a) 5.61, 6, 6

b) 16.3, 17, 18

c) 24.8, 24.9, 23.5

2.2: a) A: 6.46, B: 6.85

b) A: 7, B: 7

c) der letzte Wert ist bei A kleiner

d) der Median wird von der Grösse des grössten Wertes nicht beeinflusst

2.3: a) Mittelwert: CHF 29300, Median: CHF 23500, Modus: CHF 23000

b) Modus ist gleich dem kleinsten Wert

c) Nein, da die Daten Schief sind

2.4: a) Mittelwert: 3.19, Median: 0, Modus: 0

b) Der Median ist nahe bei 0

c) Der Modus ist gleich 0

d) Ja, 21 und 42

e) Nein

2.5: a) 44

b) 44

c) 40.2

d) 40.3

2.6 : 116

2.7 : 3144 km

2.8 : 186604

2.9 : x=15

2.10 : a=5

2.11 : 37

2.12 : 14.77

2.13 : 9 und 7

2.14: a) 1

b) 1.43

c) 1

2.15: a) 2.96, 2, 2

c) rechtsschief mit Ausreisser 11

d) Der Mittelwert wird durch die 11 stark beeinflusst

e) Mittelwert

2.16: a) 49, 49, 49.03

b) Nein

2.17: a) 2.61, 2, 2

b) mehr Kinder (werden hier nur Familien mit Kinder betrachtet, im nationalen Vergleich alle, also auch Familien ohne Kinder)

c) rechtsschief

d) Der Mittelwert ist grösser als Median und Modus

2.18: a) 163700 und 147200

b) Verkäufer: Mittelwert, Käufer: Median

2.19: a) 70.9

b) 210

c) 139

2.20 : 10.1cm

2.21: a) 50.8, 49.9

b) Nein, denn gerundet stimmen die Zahlen gut

2.22 : 17.25

2.23 : 6 und 12

2.24: a) 31500, 2800, 33300

b) Mittelwert

2.25 : 31.7

2.26: a) 70

b) 411000

c) 5870

2.27: a) 6; $Q_1 = 4$, $Q_3 = 7$; 7; 3

b) 17.5; $Q_1 = 15$, $Q_3 = 19$; 14; 4

c) 24.9; $Q_1 = 23.5$, $Q_3 = 26.1$; 7.7; 2.6

2.28: a) median = 2.45, $Q_1 = 1.45$, $Q_3 = 3.8$

b) Spannweite=5.2, IQA=2.35

c) ... grösser als 2.45 Minuten; ... weniger als 3.8 Minuten; Minimum: 0; Maximum : 5.2 Minuten; verteilt über 5.2 Minuten.

2.29: a) 7;6;5;7;2

b) 10;7;6;8;2

2.30: a) 35;78;13;53;26

b) 27

2.31: a) 98; 25

b) 70

c) 85

d) 55 und 85

e) 73

f) 30

g) ≈ 68

2.32: a) $\min=3$; $Q_1 = 5$, median = 6, $Q_3 = 8$,
max = 10; range = 7; IQA = 3

b) $\min=0$; $Q_1 = 4$, median = 7, $Q_3 = 8$, max =
9; range = 9; IQA = 4

2.33: a)

Statistic	Quarta	Prima
min value	1	6
Q_1	5	10
median	7.5	14
Q_3	10	16
max value	12	17.5

b) Quarta: 11; 5; Prima: 11.5;6

c) Wahr; Wahr

2.34: a) Median = 6, $Q_1 = 5$, $Q_3 = 8$

b) 3

2.35: a) $\min_x = 33$, $Q_1 = 35$, $Q_2 = 36$,
 $Q_3 = 37$, $\max_x = 40$

b) 7;2

2.36: a) 9

b) $\approx 28.3\%$

c) 7

d) $IQA \approx 2.4\text{cm}$

e) 10 cm, also 90% der Setzlinge sind 10cm oder
kleiner

2.37: a) 27

b) 29

c) 31.3

d) $IQA \approx 4.3\text{ min}$

e) 28 minuten 10 sekunden

2.38: a) Verteilung A

b) A: 8; B: 8

c) A: 2; B: 1.06

2.39: a) Andreas: 25, 4.97; Thomas: 30.5; 12.6

b) Andreas

2.40: b) Vermutlich die Rockets, mit zwei Nul-
lern

Rockets: $s = 3.9$, Bullets: $s = 3.29$

c) Standardabweichung

2.41: a) Wir erwarten kleine abweichungen in
der Standardabweichung jedes Tages, denn die
Faktoren könnten von Tag zu Tag ändern.

b) Den Mittelwert

Die Standardabweichung

c) weniger Variabilität in der Füllmenge (gute
Anlage)

2.42: a) $\mu = 69$, $\sigma = 6.05$

b) $\mu = 79$, $\sigma = 6.05$

2.43: a) $\bar{x} = 1.01\text{ kg}$, $s = 0.17$

b) $\bar{x} = 2.02\text{ kg}$, $s = 0.34$

c) Verdopplung der Werte verdoppelt den Mittel-
wert und die Standardabweichung

2.44 : $p=6$, $q=9$

2.45 : $a=8$, $b=6$

2.46: a) 0.809

b) 0.150

c) Extremwerte beeinflussen die Standardabwei-
chung extrem.

2.47: a) $s_n \approx 6.77\text{ kg}$

b) $\mu \approx 93.8\text{ kg}$, $\sigma^2 \approx 46.4\text{ kg}^2$

2.48: a) $\bar{x} = 77.5\text{ g}$, $s_n \approx 7.44\text{ g}$

b) $\mu = 77.5\text{g}$, $\sigma^2 = 55.4\text{ g}^2$

2.49 : 32.4 min, 9.86 min²

2.50: a) $\bar{x} \approx 1.72\text{ Kinder}$, $s \approx 1.67\text{ Kinder}$

b) $\mu \approx 1.72\text{ Kinder}$, $\sigma^2 \approx 2.83\text{ Kinder}^2$

2.51: a) $\bar{x} \approx 14.5\text{ Jahre}$, $s \approx 1.75\text{ Jahre}$

b) $\mu \approx 14.5\text{ Jahre}$, $\sigma^2 \approx 3.18\text{ Jahre}^2$

2.52: a) $\bar{x} \approx 37.3\text{ Zahnstocher}$, $s \approx 1.45\text{ Zahn-}$
stocher

b) $\mu \approx 37.3\text{ Zahnstocher}$, $\sigma^2 \approx 2.16\text{ Zahnstocher}^2$

2.53: a) $\bar{x} \approx 47.8\text{ cm}$, $s \approx 2.66\text{ cm}$

b) $\mu \approx 47.8\text{ cm}$, $\sigma^2 \approx 7.31\text{ cm}^2$

2.54: a) $\bar{x} \approx 390.3$, $s \approx 15.87$

b) $\mu \approx 390.3$, $\sigma^2 \approx 253.18$