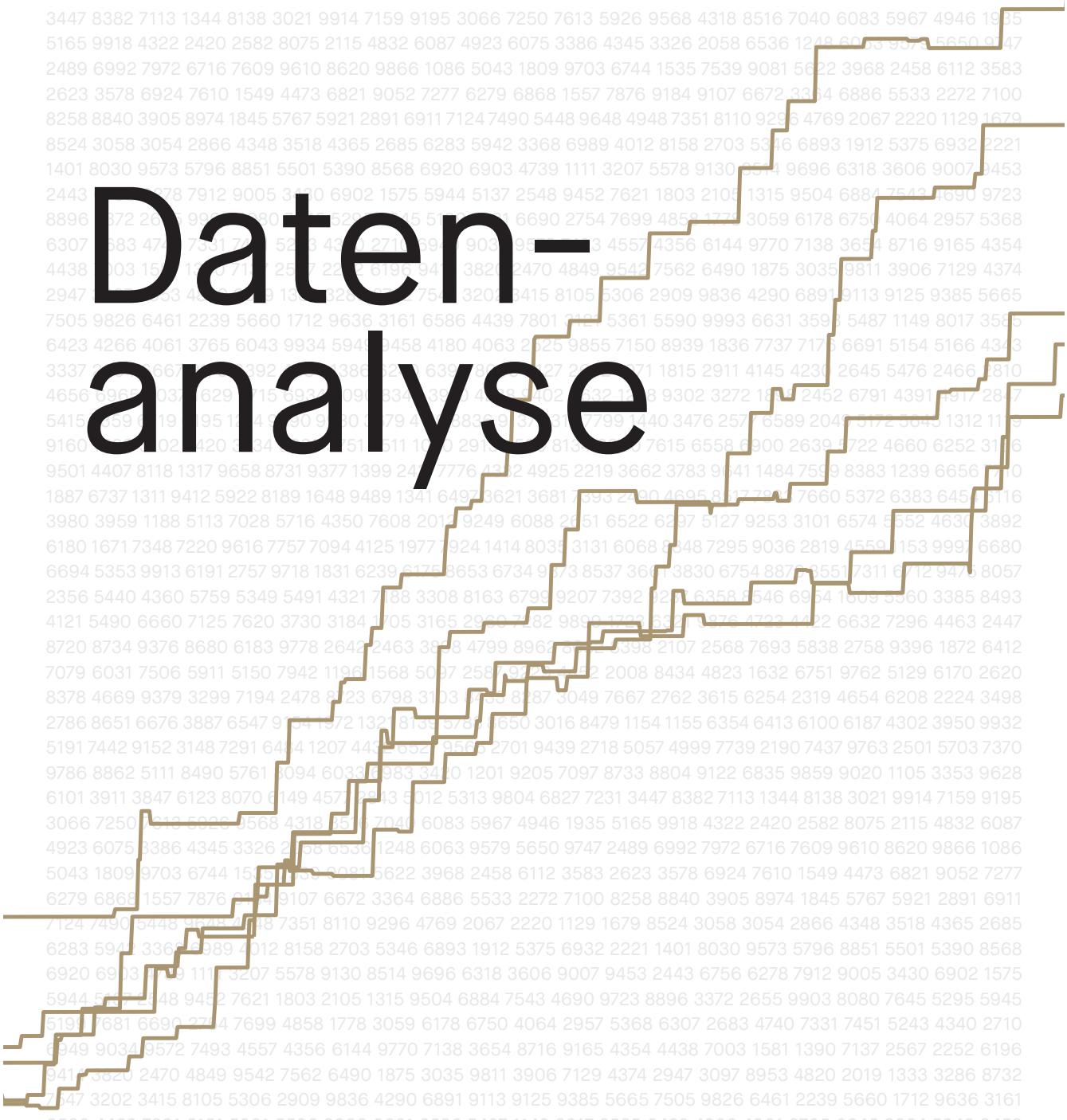
Anregungen zur Umsetzung der «Datenanalyse»
im eidg. Rahmenlehrplan 2012 für die Berufsmaturität

Daten-analyse



Beate Sick, Oliver Dürr
IDP, ZHAW 2013

1	**Grundbegriffe der Datenanalyse**	**7**
1.1	Einführung – Worum geht es in der Datenanalyse?	7
1.2	Merkmale und Datentypen	9
1.3	Datengewinnung, Urliste, geordnete Stichprobe	12
1.4	Mögliche Fehler in den Daten	13
1.4.1	Übertragungsfehler	13
1.4.2	Systematische Fehler	15
1.4.3	Zufällige Fehler	16
1.4.4	Grobe oder mutwillige Fehler	16
1.5	Stichprobenerhebung: Wie gut ist eine Stichprobe?	16
2	**Deskriptive Statistik quantitativer Merkmale**	**21**
2.1	Das Histogramm	21
2.2	Wie können die Formen von Histogrammen charakterisiert werden?	22
2.3	Kennzahlen für die Lage: Mittelwert, Median und Modus	24
2.4	Wichtige Quantile: Median und Quartile	29
2.5	Kennzahlen für Streuung	30
2.6	Der Boxplot zur Darstellung quantitativer Merkmale	33
3	**Deskriptive Statistik kategorieller Merkmale**	**35**
3.1	Häufigkeitstabellen, Balken- und Kuchendiagramme, Modus	35
4	**Manipulative Graphen**	**39**
5	**Korrelation, Kausalität und Störfaktoren**	**43**

Der neue Rahmenlehrplan für die Berufsmaturität 2012 gewichtet Kompetenzen in der angewandten Mathematik stärker als bisher. Neu soll Datenanalyse in allen Fachrichtungen der Berufsmaturität eingeführt werden und zu interdisziplinärem Arbeiten anregen. Wieso wurde die Datenanalyse gewählt, um die Interdisziplinarität zu fördern? Die Antwort ist im Grunde einfach: Die Wahl fiel auf dieses Gebiet, weil heute in fast allen Bereichen des beruflichen und öffentlichen Lebens Daten gesammelt oder erhoben werden. Methoden aus der Datenanalyse und Statistik werden eingesetzt, um die Daten geeignet grafisch darzustellen, in Beziehungen zu setzen und um daraus Informationen zu gewinnen.

Die vorliegende Broschüre wurde entwickelt, um Ihnen als Berufsmaturitätslehrkräften einen stufengerechten Einstieg in die Materie zu ermöglichen. Sie enthält Anregungen für die Unterrichtsgestaltung und ist unter Beteiligung verschiedener Schweizer Fachhochschulen entstanden. Die Autoren Prof. Dr. Beate Sick und Dr. Oliver Dürr arbeiten als Dozenten am Institut für Datenanalyse und Prozessdesign (IDP) der School of Engineering.

Ich möchte allen beteiligten Personen herzlich für ihr Engagement danken und hoffe, dass die «Einführung in die Datenanalyse» Sie als Lehrpersonen bei der Einführung des neuen Schulfaches unterstützt.

Prof. Dr. Martina Hirayama
Direktorin ZHAW School of Engineering

1

Grundbegriffe der Datenanalyse

1.1 Einführung – Worum geht es in der Datenanalyse?

Die erste Frage, die sich bei einem neuen Schulfach stellt, ist häufig: «Wozu brauchen wir das?»

Im Fall der Datenanalyse ist die Antwort, dass wir Datenanalyse überall brauchen, wo nicht alles vorbestimmt ist, sondern es Unsicherheiten und Variationen gibt. Wir können auch sagen, es braucht schon deshalb Datenanalyse und Statistik, weil nicht alle Menschen gleich sind. Wären alle Menschen gleich, könnte man sich viel Geld für Meinungsumfragen und Volksentscheide sparen und einfach das eigene Spiegelbild befragen. Auch die Schweizer Pharmaindustrie würde sich freuen – statt teure Studien durchführen zu müssen, könnte sie die Wirkung eines neuen Medikaments in einem billigen Selbsttest am CEO oder einem Laboranten demonstrieren. Die Frage, ob alle Mitarbeiter denselben Lohn erhalten sollten, könnte endgültig mit einem klaren «Ja» beantwortet werden. Eine Welt, in der jeder gleich aussieht und reagiert, dasselbe denkt, will, kann, isst und sagt, wäre sehr einfach, aber wahrscheinlich sehr langweilig. Zum Glück sind die Menschen verschieden, verändern sich mit der Zeit und wir leben in einer spannenden Welt – die Kehrseite ist, wir brauchen Datenanalyse, um sie zu beschreiben und zu verstehen.

Um eine Vermutung oder Behauptung – im Fachjargon auch Forschungshypothese genannt – zu belegen, werden häufig Daten gesammelt. Zum Beispiel kann ein Umfrageinstitut 1000 zufällig gewählte erwachsene Personen in der Schweiz dazu befragen, ob sie bereit wären, eine 5%-ige Strompreiserhöhung zu akzeptieren, wenn dafür innerhalb von zwei Jahren ein Atomausstieg realisiert würde. Dabei dienen die 1000 Erwachsenen als Stichprobe für die gesamte Schweizer Population, die oft zu gross ist, um alle Personen darin zu befragen.

Hier sind wir schon auf die ersten Grundbegriffe der Datenanalyse gestossen, die wir hier kurz definieren wollen:

Grundgesamtheit Bei der Grundgesamtheit oder Population handelt es sich um die Menge der für eine Untersuchung relevanten Einheiten/Personen. Je nach Fragestellung können dies z.B. a) alle 18-jährigen Schülerinnen in Schweizer Berufsschulen, b) alle in der Schweiz wohnhaften Personen, c) alle Flugzeuge vom Typ Airbus A320, d) sämtliche Tiere/Pflanzen einer bestimmten Gattung sein.

Stichprobe Eine Stichprobe ist eine Teilmenge der Grundgesamtheit, die unter bestimmten Gesichtspunkten ausgewählt wurde. Dies kann zufällig, systematisch, willkürlich etc. sein.

Stichprobenumfang Besteht die Stichprobe beispielsweise aus 1000 befragten Personen, so ist der Stichprobenumfang n=1000.

Die erhobenen Daten werden dann typischerweise in zwei Schritten analysiert:

Der erste Schritt beschäftigt sich mit der Organisation der Daten, der grafischen Darstellung und der Zusammenfassung der Daten in wenige Kennzahlen – dies wird auch als beschreibende Statistik oder deskriptive Statistik bezeichnet. Die Themen, die in dieser Broschüre behandelt werden, gehören alle zur beschreibenden Statistik.

Im zweiten Schritt wird mit Methoden aus der schliessenden Statistik untersucht, inwieweit sich die Beobachtungen aus der Stichprobe zu Aussagen bezüglich der gesamten Population verallgemeinern lassen. Diese Methoden werden im Rahmen dieser Broschüre nicht angesprochen.

1.2 Merkmale und Datentypen

In welchen Merkmalen können sich Menschen unterscheiden? Wie können die Datentypen dieser Merkmale in Gruppen eingeteilt werden?

Aktivität Wir wollen dieser Frage in der eigenen Klasse nachgehen und ein paar echte Daten sammeln, die wir später für verschiedene Zwecke wieder verwenden können.

Zu diesem Zweck geben wir an jeden Schüler Karten aus, die nun ausgefüllt werden sollen. Da es bei einigen Fragen etwas zu messen gibt, geben wir günstige Messbänder mit aus. Die Karten werden dann wieder eingesammelt und können im Unterricht immer wieder verwendet werden – z. B. für die Demonstration einer Stichprobenziehung, als Datengrundlage für Visualisierungsmethoden oder zur Untersuchung von Korrelationen.

Erhebung persönliche Informationen Klasse: Semester:

Bitte beantworten Sie die nachfolgenden Fragen zu Ihrer Person. In Klammern ist jeweils das Format angegeben, in welchem die Antwort erfolgen soll. Verwenden Sie bitte das angegebene Format, bzw. bei einer Auswahl sind nur Antworten aus dieser Auswahl erlaubt.

Fragen	Antworten
Alter (in Jahren)	
Geschlecht (m/w)	
Augenfarbe (blau, braun, grün, haselnuss)	
Natürliche Haarfarbe (schwarz, braun, rot, blond)	
Körpergrösse (in cm)	
Armspanne (in cm)	
Handspanne* der grösseren Hand (in cm)	
Fusslänge des grösseren Fusses (in cm)	
Anzahl Geschwister	
Mitglied in einem Sportverein (ja/nein)	
Händigkeit (links/rechts)	

*Ausgehend von der gespreizten Hand, die Distanz in cm zwischen Fingerspitze des kleinen Fingers und des Daumens.

→ Abb. 1

Abbildung 1 Beispiel für eine Karte, mit der echte Daten gesammelt werden. Angelehnt an ein Beispiel aus «Teaching Statistics – a bag of tricks», Gelman and Nolan

Auf den Karten wurden nun verschiedene Merkmale von jeder Person erfasst – z.B. Augenfarbe oder Grösse.

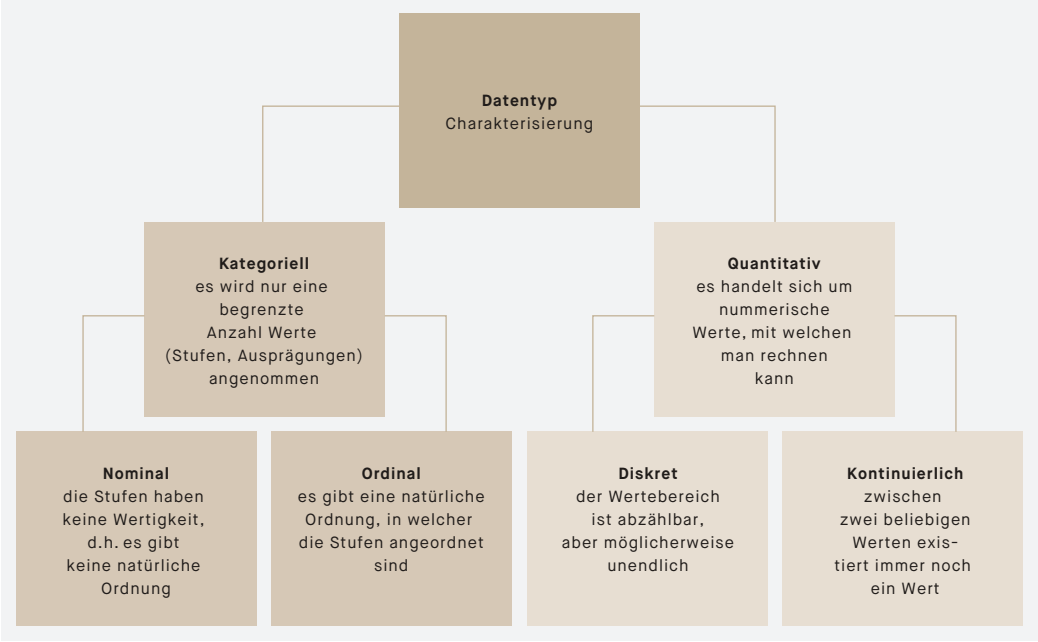
Merkmale bzw. Variablen Merkmale, die an einer Person oder allgemein an einer Untersuchungseinheit (kann auch ein Unternehmen oder eine Maschine sein), beobachtet, erfragt oder gemessen werden, werden häufig auch als Variablen bezeichnet. Der Wert, den eine Variable annehmen kann, wird als Ausprägung bezeichnet.

Ausprägung Eine mögliche Ausprägung für das Merkmal Augenfarbe ist z.B. «blau» und eine mögliche Ausprägung der Variablen Körpergrösse ist z.B. 171 cm.

Betrachten wir die Variablen auf den Karten (Geschlecht, Anzahl Geschwister, Grösse etc.), so wird schnell klar, dass es verschiedene Typen von Variablen bzw. Daten gibt. Sie unterscheiden sich schon darin, ob es sich um Zahlen handelt, mit denen man rechnen kann, oder nicht.

Quantitative Merkmale Wir sprechen von quantitativen oder auch nummerischen Merkmalen, wenn wir damit rechnen können. Das heisst, deren Ausprägung muss zwingend als Zahlen angegeben werden, wie z.B. 60 kg für das Gewicht, 176 cm als Körpergrösse oder 2 als Anzahl Tore, die von einem Fussballspieler in einer Saison geschossen wurden etc. Bei den quantitativen Datentypen kann dann weiter unterschieden werden, ob sie diskret oder kontinuierlich sind. Im diskreten Fall können nur bestimmte Werte, z.B. ganze Zahlen, angenommen werden – es gibt einfach keine halben Geschwister. Kontinuierliche Variablen können dagegen unendlich viele verschiedene Werte annehmen – Personen können 175 cm oder 176 cm gross sein, oder aber auch 175,3 cm, d.h. zwischen je zwei Zahlen gibt es immer noch eine Zahl, die dazwischenliegt.

Qualitative Merkmale Merkmale wie z.B. das Geschlecht einer Person können als Ausprägung nur bestimmte Kategorien annehmen – im Fall der Variable Geschlecht die Kategorien «männlich» oder «weiblich» – mit Kategorien kann man aber nicht rechnen. In solchen Fällen sprechen wir von qualitativen oder kategoriellen Variablen.
Beim Medaillenerfolg gibt es auch nur die Kategorien «gold», «silber», «bronze» und «keine». Das heisst es gibt eine klare Ordnung in den Ausprägungen – eine solche kategorielle Variable wird als ordinale Variable bezeichnet; beim Geschlecht gibt es keine derart klare Ordnung der Ausprägungen, hier handelt es sich um einen nominale Variable.



→ Abb. 2

Abbildung 2 Charakterisierung der Datentypen, zwischen denen wir unterscheiden.

Es führt manchmal zu Verwirrungen, wenn kategorielle Variablen durch Zahlen kodiert werden. Zum Beispiel kann man von allen Spielern einer Fussballmannschaft stoppen, wie lange sie für einen 100-m-Lauf brauchen (wir nehmen jetzt mal an, dass alle Spieler unterschiedliche Zeiten haben) und dann die Plätze 1 bis 11 vergeben. Die Plätze sind durch Zahlen 1, 2, …, 11 kodiert, aber es ist nicht möglich, mit den Platznummern sinnvoll zu rechnen. Man weiss zwar, dass der Spieler auf dem Platz 1 schneller war als der Spieler auf dem Platz 2 und dieser wiederum schneller war als der Spieler auf Platz 3, aber es könnte sein, dass zwischen dem Platz 1 und 2 nur eine halbe Sekunde Zeitdifferenz liegt und zwischen dem Platz 2 und 3 fast 5 Sekunden. Die Differenz der Platznummer macht keinen Sinn, aber sehr wohl die Rangfolge – daher ist die Platzierung eine kategorielle, ordinale Variable, auch wenn sie durch Zahlen kodiert ist. Die Differenzen der Laufzeiten sind dagegen aussagekräftige Zahlen – mit der Laufzeit kann man rechnen und sie ist eine quantitative, kontinuierliche Grösse.

Als Übung werden die gesammelten Merkmale den verschiedenen Datentypen zugeordnet.

1.3 Datengewinnung, Urliste, geordnete Stichprobe

Ein Arzt möchte untersuchen, ob und wie stark sich die Armspannen von weiblichen und männlichen 15-Jährigen unterscheiden. Als Stichprobe bekommt er von einer Schule 23 Karten (siehe Abb. 1, S. 9), die von 15-jährigen Schülern ausgefüllt wurden und auf denen unter anderem die Armspanne abgefragt wurde.

Der Arzthelfer überträgt in der willkürlichen Reihenfolge, in der die Karten im Stapel liegen, nacheinander die Angaben der Karten in eine Tabelle – z.B. in Excel. Für jede Karte bzw. jeden Schüler gibt es eine eigene Zeile – die Zeilennummern gehen von 1 bis n, wobei n=23 dem Stichprobenumfang entspricht. Für jedes Merkmal bzw. jede Variable gibt es eine eigene Spalte. Im Fachjargon wird jede Zeile auch als eine Beobachtung bezeichnet, da jede Zeile zu einem Schüler gehört. Werden an einem Schüler viele (multi) Variablen vermessen, dann spricht man von einer multivariaten Beobachtung, wird nur ein Merkmal vermessen, so spricht man von einer univariaten Beobachtung.

Durch die Spalte mit den Armlängen ist die Urliste der beobachteten Armlängen in dieser Stichprobe gegeben. Nun kann die Spalte mit den Armlängen in aufsteigender Reihenfolge geordnet werden, wobei eine geordnete Stichprobe entsteht – die Zeilennummer entspricht nun der Ordnungszahl der Beobachtung – man spricht statt von Ordnungszahlen auch von Indexwerten (vgl. S. 27). Mittelt man jeweils Ordnungszahlen, die zu den gleichen quantitativen Werten gehören, so erhält man die Ränge der Beobachtungen – diese haben den Vorteil, dass gleich grosse Werte auch gleiche Ränge haben. Zum Beispiel haben, wie es der Zufall will, Anne, Mike und Michaela die gleiche Armlänge von 90 cm. In der geordneten Tabelle haben sie die Nummern 10, 11 und 12. Die Ränge berechnen sich dann als zu 11 für alle drei Personen.

Um den Zusammenhang mit den anderen Merkmalen derselben Person nicht zu verlieren, sollten die ganzen Zeilen umgeordnet werden. Man kann nicht nur Zahlen ordnen, sondern auch qualitative Merkmale, wie zum Beispiel die Variable Geschlecht, die als mögliche Ausprägungen «männlich» und «weiblich» enthalten kann – dabei wird meist

nach dem Alphabet geordnet, sodass erst alle Zeilen mit männlichen Schülern aufgeführt werden und dann alle Zeilen mit weiblichen Schülern.

1.4 Mögliche Fehler in den Daten

Nachdem Daten erhoben wurden, ist die Versuchung gross, daraus sofort die Grösse zu berechnen, für die man sich interessiert – hier z.B. die mittlere Armspanne der 15-jährigen Mädchen und die mittlere Armspanne der 15-jährigen Jungs. Aber wie so oft im Leben ist es auch hier meist der bessere Weg, nicht jeder Versuchung zu erliegen, sondern mit Bedacht das gesteckte Ziel zu verfolgen.

Dazu gehört es beispielsweise herauszufinden, ob es Fehler in den Daten gibt. In den meisten Fällen kann man davon ausgehen, dass es Fehler gibt, und es ist nur noch die Frage, wo diese liegen und wie wir mit den Fehlern umgehen können. Fehler können auf verschiedene Arten entstehen. Im Folgenden zählen wir ein paar häufige Fehlertypen auf:

1.4.1 Übertragungsfehler

Beim Abschreiben von Werten können sogenannte Übertragungsfehler gemacht werden. Einer der berühmtesten Übertragungsfehler passierte ca. 1890 bei einer Tabelle zum Eisengehalt verschiedener Nahrungsmittel. Im Wissen-Teil der Zeitung «Zeit» wird das Thema aufgegriffen[1], als der folgenden Leserfrage nachgegangen wurde:

«Spinat ist gesund, weil er besonders viel Eisen enthält. Stimmt's?»

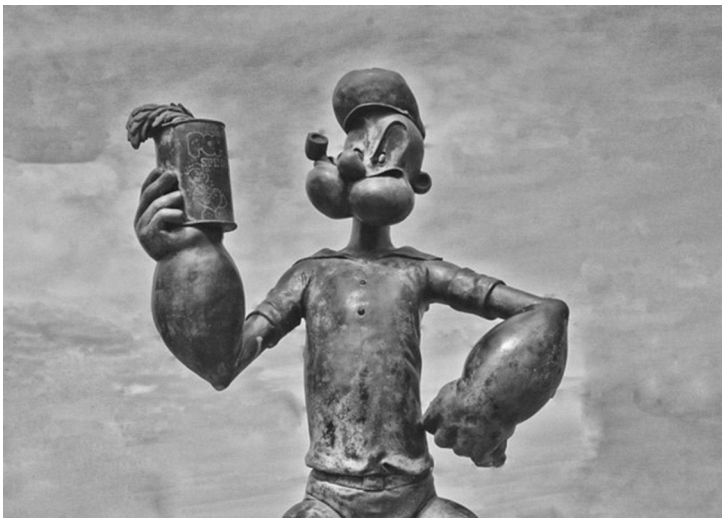
Die Antwort der Zeit-Redakteure lautete wie folgt:

«Stimmt nicht. Der Eisengehalt von frischem Spinat ist mit 2,6 Milligramm auf 100 Gramm eher gering. Wer bei seiner Ernährung Wert auf blutbildendes Eisen legt, der sollte sich eher an Leberwurst (5,9 mg), Schokolade (6,7 mg) und Pistazien (7,3 mg) halten.

Selbst von dem bisschen Eisen, das im Spinat tatsächlich enthalten ist, darf sich der Mensch kaum

[1] **URL** www.zeit.de/stimmts/1997/1997_41_stimmts

Abbildung 3 Durch Spinat erhält Popeye übermenschliche Kräfte



→ Abb. 3

Nutzen versprechen: ‹Der Verdauungstrakt kann nicht viel davon aufnehmen›, erklärt die amerikanische Ernährungsforscherin Judi Morrill. ‹Spinat enthält nämlich auch sehr viel Oxalsäure, die das Eisen bindet.›

Der Ursprung der Legende liegt in den neunziger Jahren des vorigen Jahrhunderts. Damals soll ein Lebensmittelanalytiker bei der Untersuchung von Spinat das Komma versehentlich um eine Stelle nach rechts gerückt und dem Gemüse somit den zehnfachen Eisengehalt attestiert haben.

Das jedenfalls behauptet zum Beispiel der englische Krebsspezialist T. J. Hamblin in einem Artikel, der 1982 im angesehenen British Medical Journal erschien – allerdings kann auch Hamblin die Originalquelle nicht angeben. Ist vielleicht die Geschichte vom Ursprung der Legende selbst eine Legende? Jedenfalls galt Spinat fortan als besonders gesund, zum Leidwesen von Millionen Kindern. Denen hat die grüne Pampe auch nach 1929 nicht besser geschmeckt. In diesem Jahr erfand der Zeichner Elzie Segar die Comicfigur Popeye, dem der Genuss einer Dose Spinat immer zu übermenschlichen Kräften verhalf. Angeblich soll der gezeichnete Seemann den Spinatkonsum in den USA um ein Drittel gesteigert haben – das behauptet jedenfalls die Inschrift auf dem Popeye-Denkmal in der texanischen Spinatmetropole Crystal City.»

1.4.2 Systematische Fehler

Manchmal passiert es, dass Messungen systematisch falsch gemacht werden. Das kann schon in der Küche passieren, wenn man die Zutaten für einen Kuchen mithilfe einer Küchenwaage abwiegt, deren Nullpunkt verschiebbar ist (siehe Abb. 4).

Stellt man den Nullpunkt ohne Schüssel ein und wiegt dann die Zutaten eines Kuchens in der Schüssel ab, dann kann es passieren, dass man von allen Zutaten systematisch zu wenig nimmt, und zwar genau um das Gewicht der leeren Schüssel.

Abbildung 4 Eine Küchenwaage, bei der der Nullpunkt durch Drehen der Auflagefläche eingestellt werden kann.

→ Abb. 4

Als zweites Beispiel kann man an die abgefragten Grössen auf der Karte (vgl. Abb. 1, S. 9) denken. Im Fall der Armspannen ist es möglich, dass einige Schüler nur die Armlänge angeben statt den Abstand von Fingerspitzen zu Fingerspitzen bei ausgebreiteten Armen – die Zahlen auf den Karten dieser Schüler sind systematisch zu klein, da die Armlänge weniger als halb so gross ist wie die interessierende Armspanne. Das Histogramm von Armlängen in Abb. 8 auf S. 22 deutet darauf hin, dass genau dies bei einigen Schülern falsch gelaufen ist, wenn wir davon ausgehen, dass es vermutlich nicht stimmt, dass drei der vermessenen 15-Jährigen nur knapp 1 m gross sind.

1.4.3 Zufällige Fehler
1.4.4 Grobe oder mutwillige Fehler
1.5 Stichprobenerhebung: Wie gut ist eine Stichprobe?

[2] **URL** www.nzz.ch/aktuell/startseite/eth-forschungschef-ruecktritt-datenfaelschungen-1.3617680)

Abbildung 5 Auch Calvin und Hobbes machen sich einen Spass daraus, Daten zu manipulieren. Quelle: www.holypoll.com

1.4.3 Zufällige Fehler

Zufällige Fehler sind bei allen Messungen unvermeidbar. Lässt man beispielsweise die Armspanne eines Schülers von 10 verschiedenen Mitschülern nachmessen, so ist es völlig normal, dass nicht 10 Mal die gleiche Armspanne gemessen wird. Die Gründe liegen sowohl beim vermessenen Schüler, der vermutlich seine Arme nicht immer gleich stark ausstrecken wird, als auch beim Messprozess, der nicht sehr präzise durchführbar ist. Im Unterschied zu systematischen Fehlern, weichen zufällige Fehler aber nicht systematisch immer in die gleiche Richtung vom «wahren Wert» ab, sondern eben gerade zufällig.

1.4.4 Grobe oder mutwillige Fehler

Bei Daten kann es aber auch grobe oder gar mutwillige Fehler geben, wenn Werte absichtlich falsch angegeben wurden. Vor etlichen Jahren ist sogar der Forschungschef der ETH zurückgetreten, weil in seiner Gruppe Daten manipuliert wurden, auch wenn nicht aufgeklärt werden konnte, wer die Manipulationen vorgenommen hat[2].

Zum Teil werden Daten aber auch aus Jux und Tollerei falsch angegeben. Ein Comic von Calvin und Hobbes illustriert dies (www.holypoll.com), indem Calvin in einem Fragebogen angibt, er sei 43 Jahre alt, würde 500 $ pro Woche für Kaugummi ausgeben und seine bevorzugte Geschmacksrichtung sei Knoblauch/Curry.

Bei der Analyse der Fragebogendaten von Calvin besteht aber durchaus die Chance, dass es auffällt, dass wöchentliche Ausgaben für Kaugummi von 500 $ extrem hoch sind und auch die weiteren Angaben in seinem Fragebogen eher ungewöhnlich sind.

1.5 Stichprobenerhebung: Wie gut ist eine Stichprobe?

Stellen wir uns vor, wir wollen die mittlere Anzahl Kinder pro Familie in der Schweiz ermitteln. Aus praktischen Gründen will man nicht alle in der Schweiz lebenden Familien befragen (Vollerhebung), sondern nur eine Stichprobe aus

ihnen. Zum Beispiel könnte man Ihre Klasse als Stichprobe ansehen und eine Umfrage in der Klasse durchführen.

Wir haben das in einer Berufsmatura-Klasse gemacht und sind auf durchschnittlich etwa drei Kinder pro Familie gekommen (für die Definition des Durchschnittswertes siehe S. 24). Laut dem Schweizer Bundesamt für Statistik gibt es aber im Durchschnitt weniger als 2 Kinder pro Familie in der Schweiz. Wie erklärt sich die Diskrepanz von etwa einem Kind? Machen Kinder aus grossen Familien eher die Berufsmaturität als Kinder aus kleineren Familien? Sind in dieser Klasse mehr Kinder aus kinderreichen Immigrantenfamilien vertreten? Diese und noch viele weitere Thesen könnten den Lernenden einfallen, wenn man sie über diese Diskrepanz nachdenken lässt. Die Lösung dieses Rätsels ist allerdings viel einfacher. Sie wird klar, wenn man sich überlegt, dass in der Schweiz schon ein verheiratetes Paar ohne Kinder als Familie gilt, wodurch die durchschnittliche Anzahl Kinder pro Familie sinkt. In unserer Klasse sind diese Familien aber nicht vertreten, da sie kein Kind in die Berufsschule schicken können, wenn sie keines haben.

In der Statistik spricht man in so einem Fall von einer Stichprobenverzerrung, einem Sampling-Bias oder einfach nur einem Bias. So ein Bias kommt immer dann vor, wenn man die Stichprobe nicht rein zufällig aus der Grundgesamtheit, an der man interessiert ist, zieht. Man erhält so nicht nur statistische Schwankungen, sondern eine systematische Abweichung. In unserem Beispiel ist die Grundgesamtheit die Menge aller in der Schweiz wohnhaften Familien. Mit unserem Beispiel haben wir aber nicht zufällig Familien aus allen Schweizer Familien ausgewählt, sondern nur aus solchen mit mindestens einem Kind. Solche Verzerrungen kommen häufig in der Datenanalyse vor. Hier einige einfache Beispiele, bei denen man leicht einen Bias finden kann:

- Man will das mittlere Einkommen aller Schweizer Bürger herausfinden. Man führt eine telefonische Gehaltsumfrage am frühen Nachmittag durch, bei der man zufällig private Festnetznummern wählt.
- Man möchte die Akzeptanz für mehr Parkraum in der Zürcher Bevölkerung ermitteln und führt am Zürcher Hauptbahnhof eine Befragung der Passanten durch.
- Man interessiert sich für die durchschnittliche Zeit, die ein Schweizer pro Tag im Internet surft, und führt eine Online-Umfrage unter den Besuchern von www.20minuten.ch durch.

Ein weiteres, etwas makabres Beispiel für einen Bias stammt aus dem Zweiten Weltkrieg. So untersuchte der amerikanische Statistiker Abraham Wald an den zurückgekehrten Bombern die Treffer der Flugabwehr. Sein Auftraggeber, die Statistical Research Group, wollte die Stellen an den Bombern finden, die besonders häufig getroffen werden und deshalb eine extra Panzerung benötigten. Eine schematische Illustration der Treffer an allen Flugzeugen ist in Abb. 6 gezeigt:

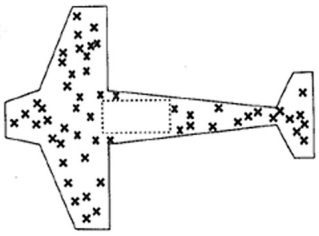


Abbildung 6 Schematische Darstellung der Einschusslöcher bei getroffenen zurückgekehrten Bombern. Quelle: «So lügt man mit Statistik»

→ Abb. 6

Repräsentativität Wir nennen eine Stichprobe repräsentativ, wenn sie ein unverzerrtes Abbild der Grundgesamtheit ist, d.h. alle Merkmale getreu deren relativer Häufigkeit[1] in der Grundgesamtheit aufweist.

Als guter Statistiker liess sich Wald natürlich nicht zu der falschen Annahme verleiten, dass die Stellen, an denen keine Treffer zu finden sind, keiner Verstärkung bedürften. Er kam vielmehr zu dem umgekehrten Schluss, die Stellen zu verstärken, an denen gar keine Treffer zu beobachten sind.[2] Falls Ihnen Walds Lösung nicht plausibel erscheint: Im Englischen werden solche Verzerrungen sehr treffend als «Survival Bias» bezeichnet – also «Überlebens-Bias», da nur die Objekte oder Personen betrachtet werden, die z.B. eine bestimmte Behandlung oder einen bestimmten Einsatz überlebt haben und diese sich häufig systematisch von den restlichen Objekten oder Personen unterscheiden, die nicht überlebt haben. In unserem Beispiel konnten nur Bomber untersucht werden, die nicht an Stellen getroffen wurden, welche unmittelbar zum Absturz führen.

Doch nun zurück zu unserem Ausgangsproblem: die Anzahl der Kinder pro in der Schweiz lebender Familie abzuschätzen. Wie können wir die Untersuchung dennoch

[1] Für die Begriffe Häufigkeit und relative Häufigkeit siehe Kap. 3.1, S. 35

[2] Mangel, M. & Samaniego, F. J. (1984) Abraham Wald's work on aircraft survivability. Journal of the American Statistical Association, 79, 259 – 267.

mit der Klasse als Stichprobe durchführen? Eine Möglichkeit wäre, irgendwie zu versuchen, den Anteil der Familien ohne Kind abzuschätzen und dann unsere Schätzung entsprechend zu korrigieren. Eine weitere Möglichkeit wäre, die Lernenden zu fragen, wie viele Kinder ihr ältester verheirateter Onkel mütterlicherseits hat. Auch hier wird sich bei der Ermittlung der durchschnittlichen Anzahl Kinder pro Familie in der Schweiz vermutlich ein kleiner Bias einschleichen. Kommen Sie darauf?[3] Allerdings ist davon auszugehen, dass dieser Bias viel geringer ausfallen wird.

[3] Wenn man davon ausgeht, dass in Familienverbänden die Anzahl der Kinder korreliert, sind Onkel eher seltener kinderlos als der Rest der Gesellschaft.

2

Deskriptive Statistik quantitativer Merkmale

2.1 Das Histogramm

In einem ersten Schritt der Datenanalyse sollten wir versuchen, ein Gefühl für unsere Daten zu bekommen! Dadurch lernen wir nicht nur unsere Daten besser kennen, sondern es gibt uns auch die Chance, ungewöhnliche Werte oder Ausreisser, die möglicherweise durch Fehler entstanden sind, aufzuspüren. Dazu ist es sehr hilfreich, die Daten darzustellen.

Um die Verteilung der Armspannen darzustellen, gehen wir am besten von der geordneten Stichprobe der Armlängen aus – wir können ihr sofort entnehmen, was der kleinste Wert ist, z.B. 81 (cm) und was der grösste Wert ist, z.B. 195 (cm). Wir teilen den Wertebereich erst mal grob in nicht-überlappende Klassen ein – z.B. in 20-cm-Intervalle: [80, 100), [100, 120), …, [180, 200), wobei die linke eckige Klammer sagt, dass die Zahl des Intervallanfangs noch in diesem Intervall ist und die rechte runde Klammer sagt, dass Zahlen bis zur rechten Zahl drin sind, aber die rechte Zahl selber nicht mehr. Im Intervall [80, 100) liegen also beispielsweise die Zahlen 80.00, 92.4, 99.99, aber die Zahl 100 liegt gerade nicht mehr darin. Wir können nun die Intervalle untereinander schreiben und für jeden Schüler, dessen Armspanne in der entsprechenden Grössenklasse liegt, einen Strich machen:

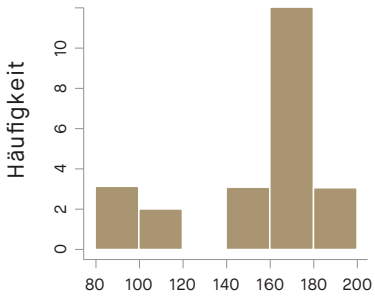
Abbildung 7 Anhand der Anzahl Striche erkennt man, wie viele Armspannen sich in einem Intervall befinden.

Abbildung 8 Die Anzahl Armspannen pro Intervall dargestellt in einem Histogramm.

[80, 100)	|||	[140, 160)	|||
[100, 120)	||	[160, 180)	|||||||||||
[120, 140)		[180, 200)	|||

→ Abb. 7

Wenn wir nur die Silhouette der Striche in Abb. 7 betrachten, erhalten wir ein Histogramm wie in Abb. 8.



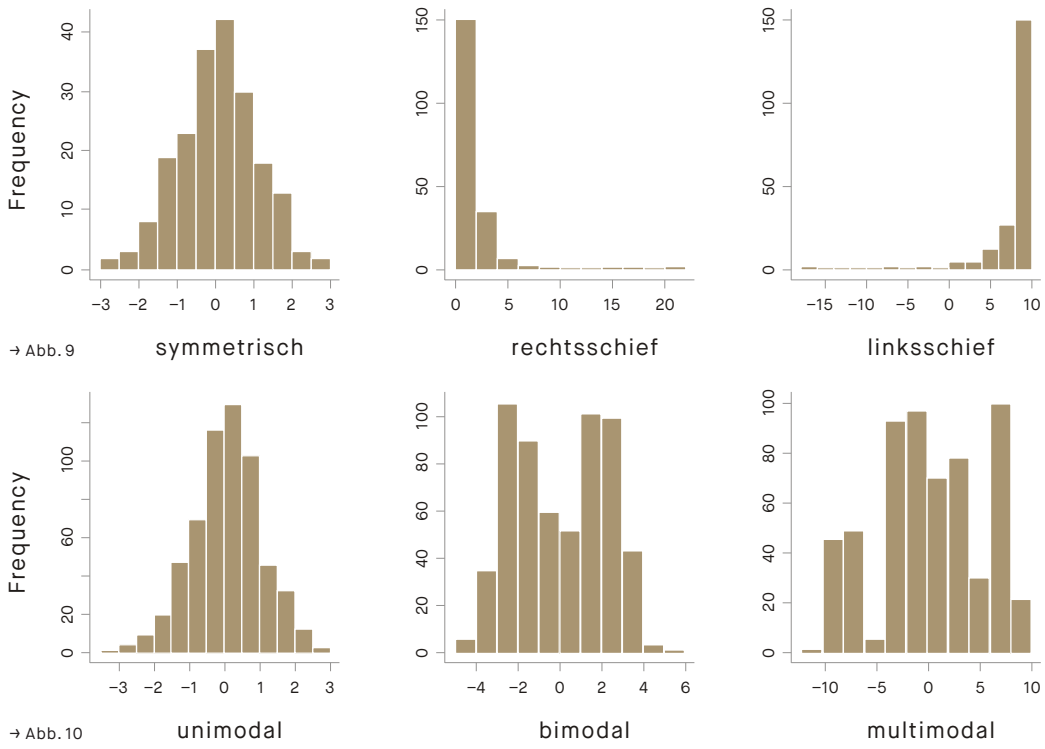
→ Abb. 8

Histogramm Um die Verteilung eines quantitativen Merkmals zu visualisieren, kann basierend auf einer Stichprobe ein Histogramm erstellt werden.
Beim Histogramm werden auf der x-Achse die Intervalle markiert, in die die Werte des Merkmals eingeteilt werden. Für jedes Intervall wird ein Balken gezeichnet, dessen Höhe der Anzahl Beobachtungen in diesem Werteintervall entspricht.

2.2 Wie können die Formen von Histogrammen charakterisiert werden?

Wir wollen nun noch ein paar typische Formen von Histogrammen einführen:

Charakterisierung der Form eines Histogramms Es gibt symmetrische, rechtsschiefe und linksschiefe Verteilungen, bzw. Histogramme. Rechtsschiefe Histogramme haben einen Ausläufer nach rechts – linksschiefe Verteilungen haben den Ausläufer nach links.
Zur Beschreibung der Form eines Histogramms werden ausserdem die Begriffe von uni-, bi- und multimodalen Verteilungen verwendet. Ein unimodale Verteilung hat nur einen Häufungspunkt oder einen Gipfel. Eine bimodale Verteilung hat zwei Häufungspunkte und eine multimodale Verteilung hat mehrere Häufungspunkte.



→ Abb. 9

→ Abb. 10

Mit den oben abgebildeten Grafiken (Abb. 9 und Abb. 10) werden die Begriffe veranschaulicht. Es gibt manchmal auch Verteilungen, die unter keine dieser drei Definitionen fallen. Meist gibt es dafür aber keine spezielle Bezeichnung.

Bei der Entscheidung über Symmetrie oder Schiefe berücksichtige man auch die Anzahl Beobachtungen, mit denen das Histogramm erzeugt wurde. Ist sie sehr klein, so ist man sich «wenig sicher», bzw. das Histogramm kann auch nur durch Zufall schief oder symmetrisch aussehen. Ist die Anzahl Beobachtungen gross (z.B. $n > 50$), so ist man sich bei solchen Entscheidungen deutlich sicherer[4].

In der Abb. 10 wird visualisiert, was unter uni-, bi- oder multimodal zu verstehen ist: Es geht dabei um die Anzahl «Gipfel» bzw. Häufungspunkte, welche eine Verteilung aufweist. Betrachten wir beispielsweise das Histogramm der Armlängen (vgl. Abb. 8 auf Seite 22), so kann man aufgrund dieser Datengrundlage noch nicht entscheiden, ob die Armlängen in der Population aller Schüler dieser Klassen-

Abbildung 9 Typische Formen von Histogrammen sind symmetrische, rechts- und linkschiefe Verteilungen.

Abbildung 10 Die uni-, bi- und multimodalen Verteilungen werden aufgrund der Anzahl Häufungspunkte kategorisiert.

[4] Um ein stabiles Histogramm zu erhalten, sollten in den meisten Klassen eine vernünftige Anzahl Datenpunkte enthalten sein (>5) – als Faustregel gilt, bei n Beobachtungen nicht mehr als $\sqrt{n}$ Klassen zu bilden. Um Stabilität zu garantieren, sollte auch überprüft werden, dass bei gleichbleibender Anzahl Klassen die Wahl der Klassengrenzen keinen deutlichen Einfluss auf die Form des Histogramms hat.

stufe bimodal oder unimodal verteilt ist. Denn Achtung: Auch hier soll man darauf achten, dass genügend Beobachtungen vorliegen, bevor man darüber diskutiert, wie viele Häufungspunkte es in den Daten gibt.

2.3 Kennzahlen für die Lage: Mittelwert, Median und Modus

In den Medien wird oft versucht, durch eine Zahl einen riesigen Zahlenhaufen, also eine grosse Datenmenge, zusammenzufassen. Zunächst der Blick auf ein Beispiel aus der Tagespresse:

Aus dem Tagesanzeiger vom 10.10.2012[5]
«Schweiz (weltweit) an erster Stelle: Bei den privaten Durchschnittsvermögen ist die Schweiz an erster Stelle. Mit 468'186 Dollar auf dem Konto sind die Schweizer unerreicht vermögend, zumindest im Schnitt.»

Vermutlich erscheint Ihnen das durchschnittliche flüssige Vermögen von 468'186 Dollar aus dem Beispiel etwas hoch. Wenn Sie nicht gerade in der Zürcher Bahnhofstrasse oder in Zermatt wohnen (siehe Bias in Stichproben), werden die meisten Ihrer Bekannten nicht 468'186 Dollar auf dem Konto haben.

In obigem Beispiel scheint das angegebene Durchschnittsvermögen nicht wirklich typisch zu sein. Wie kann das sein, wo doch der Durchschnittswert, auch Mittelwert genannt, die bekannteste Zahl für die Lage einer Verteilung ist?

Rufen wir uns erst noch einmal in Erinnerung, wie der Mittelwert, der auch als arithmetisches Mittel bezeichnet wird, berechnet wird: Man bildet die Summe aller Werte und dividiert sie durch die Anzahl n der Beobachtungen.

$$\textit{Der Mittelwert}: \quad \overline{x} = \frac{(x_1 + x_2 + x_3 + ... x_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

[5] **URL** Quelle: www.tagesanzeiger.ch/wirtschaft/konjunktur/46-Millionen-Millionaere/story/12922460

Wir berechnen das arithmetische Mittel $\bar{x}$ anhand eines Zahlenbeispiels. Gegeben seien die Werte 3, 4, 7, 5, 0.5, 6. Wir erhalten dann als Mittelwert:

$$\bar{x} = \frac{3+4+7+5+0.5+6}{6} = \frac{25.5}{6} = 4.25$$

An diesem Beispiel erkennen wir auch, dass der Mittelwert einen Wert annehmen kann, welcher im Zahlenhaufen gar nicht vorkommt – auch schon in diesem Sinn ist der Mittelwert nicht unbedingt ein typischer Wert.

Der Mittelwert hat übrigens auch eine interessante geometrische Bedeutung. Stellt man sich die Daten als Punkte mit gleicher Masse vor und legt sie auf einen Balken, so ist der Mittelwert der Schwerpunkt, wie er in Abb. 11 dargestellt wird.

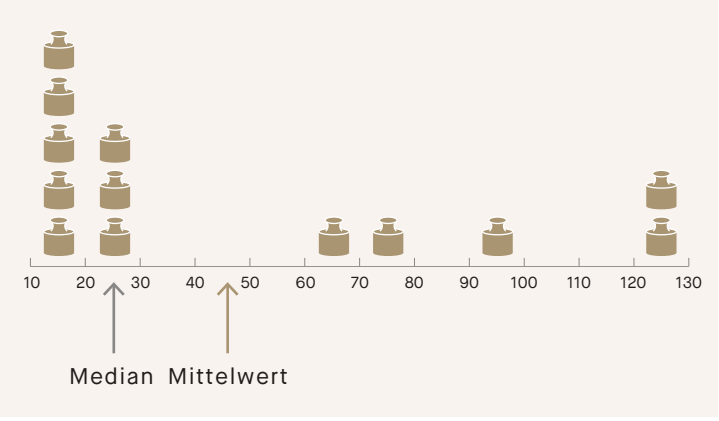


Abbildung 11 Der Mittelwert ist der Schwerpunkt. Unterstützt man in diesem Punkt, so ist die Wippe im Gleichgewicht.

→ Abb. 11

Das Bild mit der Wippe, bei der der Mittelwert den Drehpunkt markiert, zeigt sehr gut, dass der Mittelwert nicht unbedingt da liegen muss, wo die meisten Daten liegen, sondern sehr stark von wenigen sehr grossen Werten beeinflusst werden kann. Da extrem grosse Werte recht weit vom Drehpunkt der gedachten Wippe wegliegen, bekommen sie einen grossen Hebel. Und wir wissen ja schon aus dem Physikunterricht:

«Gewaltig ist des Menschen Kraft, wenn er mit dem Hebel schafft!»

Was hat das Wippen-Bild mit dem Durchschnittsvermögen eines Schweizers aus obigem Beispiel zu tun? Es ist tatsächlich in praktisch allen Ländern der Welt so, dass die Form der Vermögensverteilung auf die Bevölkerung eine ähnliche Form hat wie im Wippen-Bild dargestellt – sie ist rechtsschief: Viele Leute haben recht wenig Vermögen, je grösser das Vermögen wird, desto weniger Personen haben ein solches Vermögen und es gibt wenige Personen mit sehr, sehr hohem Vermögen. In solch einem Fall ist der Mittelwert deutlich grösser als die meisten Werte und so ist es auch im Fall des mittleren Vermögens einer Person aus der Schweiz – die meisten Leute haben deutlich weniger Vermögen als diesen Mittelwert.

Der Mittelwert kann sogar über jeden denkbaren Wert hinaus anwachsen, wenn nur ein einziger Wert in der Stichprobe einem extrem grossen Ausreisser entspricht – z.B. wenn beim Abtippen der Zahl der Finger länger auf der Null-Taste ruhte und so aus Versehen und unbemerkt 6 Nullen an die Zahl angehängt wurden und nun eine Million Mal so gross ist, wie sie eigentlich sein sollte.

Eine Kennzahl heisst robust, wenn sie nicht durch einen einzigen Ausreisser in der Stichprobe über jedes Mass hinaus ändern kann. Der Mittelwert ist also nicht robust.

Was ist dann aber eine gute Zahl, um das typische Vermögen einer Schweizer Person zu beschreiben?
Ein allgemein akzeptiertes Vorgehen ist es, den Vermögenswert zu nehmen, für den gilt, dass 50% aller Personen in der Schweiz weniger als dieses Vermögen haben und 50% aller Personen in der Schweiz ein grösseres Vermögen besitzen. Dieser Wert wird als Median bezeichnet und liegt sozusagen «in der Mitte».

Der Median ist ein robustes Lagemass; der Median bleibt unverändert, auch wenn der grösste Wert der Stichprobe z.B. durch einen Abschreibfehler eine Million Mal grösser wird als jeder andere Wert in der Stichprobe. Dies kann man sich schon anhand zweier Stichproben mit jeweils nur 3 Werten klarmachen: in den beiden Stichproben 3, 6, 10 und 3, 6, 1'000'000 ist jeweils 6 der Median – für den Wert 6 gilt, dass es gleich viele Werte in der Stichprobe gibt, die kleiner sind als 6 und grösser sind als 6.

Zur konkreten Berechnung des Medians müssen wir die Daten nach der Grösse ordnen. Aus der Stichprobe $x_1, x_2, ..., x_n$ wird dabei die geordnete Stichprobe $x_{[1]}, x_{[2]}, ..., x_{[n]}$, wo $x_{[1]}$ der kleinste und $x_{[n]}$ der grösste beobachtete Wert ist.

Besteht unsere Stichprobe z.B. aus den Werten: $x_1=4$, $x_2=2$, $x_3=6$, $x_4=1$, dann sieht die geordnete Stichprobe wie folgt aus: $x_{[1]}=1$, $x_{[2]}=2$, $x_{[3]}=4$, $x_{[4]}=6$

Für den Median gilt nun in der geordneten Stichprobe beim Indexwert

$$\textit{Indexwert des Medians}: \quad \left[\frac{n+1}{2}\right]$$

Sprich, der Median liegt genau in der Mitte zwischen Maximum und Minimum. Wenn eine ungerade Anzahl Messungen vorliegt, so ergeben sich mit der obigen Formel keine Probleme: der Quotient $(n+1)/2$ ist dann ganzzahlig, und wir können die entsprechende Beobachtung aus der geordneten Stichprobe heraussuchen. Bei gerader Anzahl Beobachtungen ist $(n+1)/2$ hingegen nicht ganzzahlig. Wir müssen dann mit den beiden benachbarten Beobachtungen interpolieren[6]:

$$\textit{Median}: \quad \tilde{x} = \begin{cases} x_{\left[\frac{n+1}{2}\right]}, & \textit{falls n ungerade} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right), & \textit{falls n gerade} \end{cases}$$

Wir wollen nun auch den Median aufgrund von einem Zahlenbeispiel berechnen. Dazu verwenden wir erneut die Zahlen 3, 4, 7, 5, 0.5, 6. Es liegt eine gerade Anzahl Beobachtungen vor (6 Stück), der Median befindet sich beim Indexwert [3.5]. Um ihn zu berechnen, müssen wir also zwischen der dritt- und viertkleinsten Beobachtung mitteln:

	3	4	7	5	0.5	6
	x_1	x_2	x_3	x_4	x_5	x_6
geordnet:	0.5	3	4	5	6	7
	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$

$$x_{[3.5]} = \tilde{x} = \frac{1}{2}\left(x_{[3]} + x_{[4]}\right) = \frac{1}{2}\left(4+5\right) = 4.5$$

[6] Wir verwenden hier mit dem Mittelwert die einfachste Interpolation. Es kann auch gewichtet interpoliert werden und die meisten Statistikprogramme bieten mehrere Möglichkeiten für die Interpolation an, wodurch kleine Abweichungen zu dem von Hand berechneten Wert entstehen können. Egal welche Interpolation verwendet wird, der Wert liegt immer irgendwo zwischen den beiden Nachbarwerten und wir wollen hier nicht diskutieren, wann er eher näher am linken als am rechten Nachbarn liegen sollte.

Abbildung 12 Lohnverteilungsbeispiel und die Berechnung Mittelwert, Median und Modus

Eine andere Möglichkeit, einen typischen Wert zu nennen, ist es, den Wert herauszusuchen, der am häufigsten vorkommt – dieser Wert wird Modus oder Modalwert genannt. Der Modus hat den Vorteil, dass er wirklich typisch ist, da er in der Stichprobe ja am häufigsten auftritt, und dass er robust ist, da ein Ausreisser per Definition sicher nicht am häufigsten vorkommt und damit keinen Einfluss auf den Modus hat. Der Nachteil des Modus liegt allerdings auch auf der Hand – es kann sein, dass mehrere Werte gleich häufig vorkommen und dann ist der Modus nicht mehr definiert.

Die drei wichtigsten Lagemasse:

Mittelwert Das arithmetische Mittel.

$$Mittelwert: \quad \overline{x} = \frac{(x_1 + x_2 + x_3 + ...x_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median Der Wert, der in einer geordneten Stichprobe in der Mitte liegt.

$$Median: \quad \tilde{x} = \begin{cases} x_{\left[\frac{n+1}{2}\right]}, & falls\ n\ ungerade \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right), & falls\ n\ gerade \end{cases}$$

Modus Der Wert, der in der Stichprobe am häufigsten vorkommt.

Wir wollen die Lagemasse Mittelwert, Median und Modus noch mal an einem Beispiel demonstrieren, bei dem ein typischer Lohn von 15 Angestellten eines Kleinbetriebs angegeben werden soll.

Mittelwert

Summe:

	$	125,000
+	$	60,000
+ 2 x	$	24,000
+ 3 x	$	15,000
+	$	12,000
+ 7 x	$	10,000
	$	360,000

$\frac{\$\ 360{,}000}{15} = \$\ 24{,}000$

Median

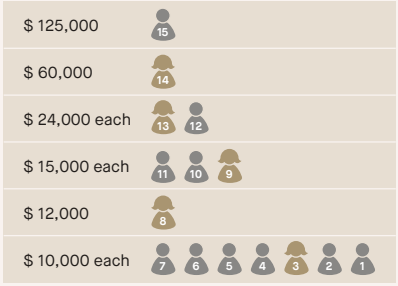
Ordne 15 Angestellte nach dem Lohn. Die 8. Person bekommt einen «mittleren Lohn» – **den Median.**

Hier beträgt der Median $ 12,000

Modus

Der Lohn, welcher am häufigsten ausbezahlt wird.

Hier beträgt der Modus $ 10,000



→ Abb. 12

2.4 Wichtige Quantile: Median und Quartile

Neben dem Median sind noch andere Kennzahlen von Interesse, die als Quantile bezeichnet werden.

Wir haben mit dem Median schon das wichtigste Quantil kennen gelernt.

Der Median teilt die geordneten Daten in zwei Hälften. Betrachten wir nun die linke Seite und nehmen von diesen Daten wiederum den Median, so erhalten wir das erste Quartil Q_1. Mit der rechten Seite machen wir dasselbe und bekommen das Q_3-Quartil. Wir haben so die Daten in 4 gleich grosse Teile geteilt: 25% der Daten sind kleiner als Q_1, 25% der Daten liegen zwischen Q_1 und dem Median (auch Q_2 genannt), 25% zwischen Q_2 und Q_3 und die restlichen 25% aller Daten sind grösser als Q_3.

Um das erste Quartil Q_1 zu bestimmen, nehmen wir also den Indexwert für den Median und mitteln diesen mit dem Indexwert des Minimums, also mit $[1]$. Wie wir bereits weiter oben gesehen haben, befindet sich der Median beim Indexwert $[3.5]$ (vgl. S. 27). Der so errechnete Indexwert ist nur noch in Ausnahmefällen ganzzahlig, d.h. es kann dann nicht der Stichprobenwert mit diesem Index genommen werden, sondern es wird der Mittelwert aus den beiden Nachbarwerten berechnet[7].

[7] Wie schon in einer Fussnote des letzten Kapitels diskutiert, können die Quantilwerte, die nach den hier angegebenen Formeln berechnet werden, leicht von den Werten abweichen, die von Statistikprogrammen wie z.B. Excel berechnet werden, da es neben der Mittelung noch andere Möglichkeiten der Interpolation gibt. Bei jeder Methode liegt der Wert aber irgendwo zwischen denselben beiden Werten der Stichprobe, mal näher am kleineren Wert, mal näher am grösseren Wert.

Wir wollen nun das erste Quartil Q_1 für die Zahlen 3, 4, 7, 5, 0.5, 6 bestimmen.

$$Index\ des\ Q_1: \left[\frac{1+3.5}{2}\right] = [2.25] \qquad Q_1 = \frac{x_{[2]} + x_{[3]}}{2} = \frac{3+4}{2} = 3.5$$

Da der Indexwert nicht ganzzahlig ist, mitteln wir die benachbarten Beobachtungen.

Um das dritte Quartil Q_3 zu bestimmen, gehen wir analog vor, wobei wir verwenden, dass der Index des grössten Stichprobenwerts hier 6 ist und der Index des Medians 3.5 ist.
Wir bestimmen nun das dritte Quartil Q_3 für die sechs Zahlen 3, 4, 7, 5, 0.5, 6.

$$Index\ des\ Q_3: \left[\frac{6+3.5}{2}\right] = [4.75] \qquad Q_3 = \frac{x_{[4]} + x_{[5]}}{2} = \frac{5+6}{2} = 5.5$$

2.5 Kennzahlen für Streuung

Nicht immer ist man nur an mittleren Werten interessiert. Stellen Sie sich vor, Sie sind mit Ihrem Auto oder Ihrer Vespa unterwegs und wissen, dass Sie im Mittel 5 Liter Benzin pro 100 Kilometer verbrauchen. Sie haben noch 5 Liter im Tank und kommen an eine überteuerte Tankstelle. Die nächste billige Tankstelle liegt noch 90 Kilometer entfernt. Um die Tankstelle zu erreichen, dürften Sie maximal einen Verbrauch von 5.55 Litern pro 100 Kilometer haben. Würden Sie weiterfahren? Neben Ihrer Risikobereitschaft kommt es ganz darauf an, wie stark Ihr Benzinverbrauch schwankt. Verbrauchen Sie jedes Mal fast genau 5 Liter pro 100 Kilometer und der Benzinverbrauch schwankt nur ganz gering um den Mittelwert 5 Liter, oder verbrauchen Sie mal 3.0, mal 8.2 und mal 4 Liter? Im ersten Fall stehen Ihre Chancen, die nächste Tankstelle zu erreichen, sehr gut, im zweiten Fall ist es ein Vabanquespiel.

Doch wie können wir die Schwankung oder Streuung der Daten quantifizieren? Notieren wir uns den Benzinverbrauch in Liter bei den letzten fünf 100-km-Fahrten, so könnten wir etwa folgende Zahlen bekommen: 5.1, 4.3, 5.3, 3.4, 6.4. Eine Möglichkeit, die Streuung zu quantifizieren wäre es, einfach den Abstand des maximalen Wertes zum minimalen Wert zu nehmen. Der minimale Verbrauch ist 3.4 Liter, der maximale 6.4 Liter, die Differenz also 3 Liter. Dieser Wert wird in der Statistik als Spannweite bezeichnet (siehe Box S. 32). Die Spannweite wird tatsächlich verwendet. Sie hat allerdings eine unschöne Eigenschaft: Sie hängt systematisch von der Grösse der Stichprobe ab. Würden Sie nochmals fünf Versuche machen, so könnten Sie folgende Werte messen: 4.5, 6.5, 7.6, 6.6 und 5.1 Liter. Das Maximum aller Messungen liegt nun bei 7.6 Litern, das Minimum bei 3.4 Litern. Die Spannweite ergibt sich somit zu 7.6-3.4=4.2 Litern und weitere Messungen würden die Spannweite wahrscheinlich noch steigen lassen. Auch haben Ausreisser und Messfehler einen starken Einfluss auf den Wert der Spannweite, d.h. diese Kennzahl ist nicht robust: Vergessen Sie das Komma, so wird aus der 5.1 eine 51 und die Spannweite ist 51-3.4=47.6 Liter.

Eine Möglichkeit, die systematische Abhängigkeit von der Stichprobengrösse und die Sensibilität gegenüber Ausreissern zu eliminieren ist es, nicht mehr den tiefsten

Wert zu nehmen, sondern einen Wert, für den ein bestimmter Anteil aller Daten noch kleiner ist – also z.B. das erste Quartil, für das 25% aller Werte kleiner gleich dem Q_1 sind. Als oberen Wert nehmen wir entsprechend das dritte Quartil Q_3. Zwischen Q_1 und Q_3 liegen also die mittleren 50% aller Werte. Den Abstand Q_3-Q_1 bezeichnet man als Interquartilsabstand oder nach dem englischen Wort Inter-Quartil-Range auch kurz IQR=Q_3-Q_1. Sollten sich Ausreisser in die Daten eingeschlichen haben, so betreffen diese nur die extremsten Ränder der Verteilung und haben damit keinen Einfluss auf den IQR, d.h. der IQR ist ein sehr robustes Streumass. Im Boxplot (siehe Abb. 14) findet der IQR seine prominenteste Anwendung: Er entspricht der Länge der Box.

Für unser Beispiel mit der Stichprobe, die aus den 6 Werten 3, 4, 7, 5, 0.5, 6 besteht, haben wir die Quartile Q_1=3.5 und Q_3=5.5 schon bestimmt und können damit sofort den Interquartilsabstand bestimmen:

IQR=Q_3-Q_1=5.5-3.5=2

Das wohl bekannteste Mass für die Streuung von Datenpunkten ist die sogenannte Standardabweichung. Die grundlegende Idee bei der Standardabweichung ist, so etwas Ähnliches wie die mittlere Abweichung vom Mittelwert als Streumass zu nehmen.

Die mittlere Abweichung vom Mittelwert wäre aber immer Null[8], da sich die positiven und negativen Abweichungen vom Mittelwert gerade gegenseitig aufheben und damit kein sinnvolles Mass für die Streuung darstellen. Man muss also dafür sorgen, dass alle Abstandswerte positiv sind, bevor darüber gemittelt wird. Konkret werden für die Berechnung der Standardabweichung zuerst für alle Datenpunkte das Quadrat der Abweichung vom Mittelwert bestimmt, dann werden alle quadratischen Abweichungen aufsummiert und die Summe dann durch n-1 geteilt[9], von diesem Wert nimmt man dann die Quadratwurzel:

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

[8] Berechnung des Mittelwerts der Abweichungen vom Mittelwert:
$\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\overline{x}) = \frac{1}{n}\cdot\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\overline{x}\right) = \frac{1}{n}\cdot(n\cdot\overline{x} - n\cdot\overline{x}) = 0$

[9] Um zu mitteln, würde man eigentlich durch n teilen statt durch n-1. Es gibt auch beide Varianten der Berechnung der Standardabweichung und hier kommen wir zu den Feinheiten. Für grosse n macht es sowieso absolut gar keinen Unterschied, ob durch n oder n-1 geteilt wird. Will man mittels der Standardabweichung der Stichprobe aber eine unverzerrte Schätzung der Standardabweichung der Grundgesamtheit erhalten, so nimmt man die Version, bei der durch (n-1) geteilt wird. Will man richtig Eindruck schinden, so fragt man, sobald das Wort Standardabweichung fällt: «unbiased» (d.h. durch n-1 teilen) oder «Maximum Likelihood» (durch n teilen)?

Als Rechenbeispiel wollen wir nun die Standardabweichung der Stichprobe mit den Werten 3, 4, 7, 5, 0.5, 6 bestimmen. Dazu berechnen wir zuerst den Mittelwert:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}\cdot(3+4+7+5+0.5+6) = \frac{25.5}{6} = 4.25$$

und damit die Standardabweichung

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

$$= \sqrt{\frac{1}{5}\cdot\left((3-4.25)^2+(4-4.25)^2+(7-4.25)^2+(5-4.25)^2+(0.5-4.25)^2+(6-4.25)^2\right)}$$

$$= 2.32$$

Die Abb. 13 zeigt, wie man den IQR und die Standardabweichung dieser Stichprobe in Excel[10] berechnet. In der oberen Zeile sind die originalen Daten dargestellt. Den IQR kann man in Excel mit =(QUARTILE(A1:F1;3)-QUARTILE(A1:F1;1)) berechnen: Er steht in der Spalte G. Die Standardabweichung berechnet sich mit dem Ausdruck =STABW(A1:F1) und steht in der Spalte H. In der zweiten Zeile sind die gleichen Daten noch mal aufgetragen, aber es haben sich beim letzten Wert zwei zusätzlichen Nullen eingeschlichen: Es steht als letzter Wert eine 600 anstelle der 6. Die Standardabweichung ändert sich stark von 2.32 auf 243.37, der IQR bleibt als robustes Mass fast ungerührt von dem Ausreisser: Er ändert sich nur von 2.5 auf 3.25.

G1 | =(QUARTILE(A1:F1;3) - QUARTILE(A1:F1;1))

	A	B	C	D	E	F	G	H
1	3	4	7	5	0.5	6	2.5	2.31840462
2	3	4	7	5	0.5	600	3.25	243.366339

→ Abb. 13

Die drei bekanntesten Streumasse

Streumass	Formel	Robustheit
Standardabweichung	$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$	nicht robust
Inter-Quartils-Range	$IQR = Q_3 - Q_1$	robust
Spannweite	$SW = \max(x) - \min(x)$	nicht robust

[10] Man beachte, dass Excel eine etwas andere Definition des Quartils, als unsere verwendet. Liegt in unserer Definition Q zwischen zwei Werten, sagen wir mal 3 und 5, so nehmen wir einfach den Mittelwert, also 4. Man kann sich denken, dass es neben dem Mittelwert viele weitere Möglichkeiten gibt, das Quantil zwischen 3 und 5 zu interpolieren. In allen Definitionen muss der Wert allerdings zwischen 3 und 5 liegen. Excel wählt eine ganz spezielle Version, eine genaue Beschreibung findet sich in www.ischool.utexas.edu/~wyllys/IRLISMaterials/quartiles.pdf.

Abbildung 13 Berechnung der Quartile in Excel

2.6 Der Boxplot zur Darstellung quantitativer Merkmale

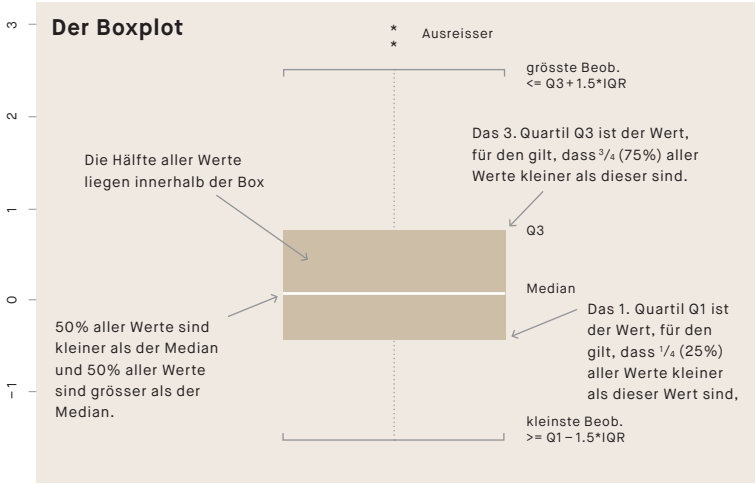


Abbildung 14 Boxplot mit Definitionen

→ Abb. 14

Nun, da wir die Begriffe Median und Quartile eingeführt haben, wollen wir eine weitere grafische Darstellungsmöglichkeit für quantitative Daten vorstellen. Es handelt sich um den Boxplot: Er enthält zwar weniger Details als das Histogramm, ist aber dafür kompakt und übersichtlich.
Lage und Streuung einer Variablen können mit ihm gut erfasst werden. Eine allfällige Schiefe lässt sich dadurch erkennen, dass die Median-Linie nicht in der Mitte der Box liegt. Weniger geeignet ist der Boxplot hingegen für bi- oder multimodale Daten, da allfällige einzelne Häufungspunkte nicht sichtbar werden.

Ein wichtiger Hinweis zum Lesen des Boxplots: die Hälfte aller Daten liegt innerhalb der Box. Die Länge der Box entspricht dem Interquartilsabstand und ist ein robustes Streumass. Der Bereich zwischen den beiden Whiskern (Antennen) markiert den «normalen» Schwankungsbereich der Daten. Alle Datenpunkte, die weiter als 1.5-mal dem Interquartilsabstand von den Rändern der Box entfernt liegen, also ausserhalb der Antennen, werden als Extremwerte bezeichnet.

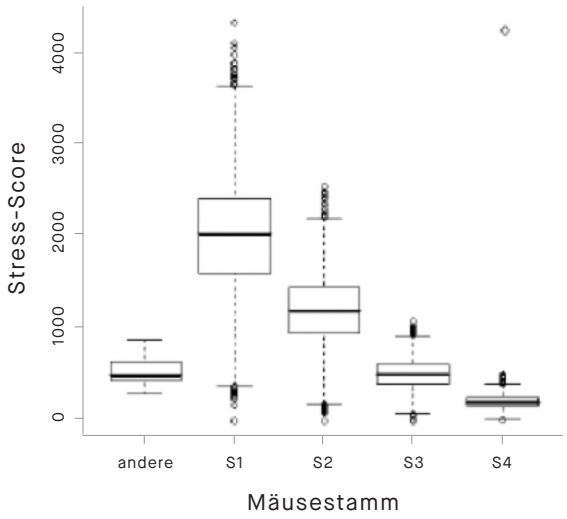
Damit ist ein Boxplot ausgezeichnet geeignet, um Ausreisser zu identifizieren – diese werden immer als Punkte ausserhalb der Antennen des Boxplots sichtbar – je weiter ein Punkt ausserhalb liegt, desto klarer handelt es sich um einen Ausreisser.

Abbildung 15 Gruppierte Boxplots zum Vergleich der Stress-Score-Verteilungen in verschiedenen Mäusestämmen

Der Boxplot hat seine Stärken vor allem dann, wenn die Verteilung eines quantitativen Merkmals für mehrere Untergruppen verglichen werden soll. Dabei handelt es sich aber bereits um deskriptive Statistik mit mehr als einer Variablen.

In einem Experiment kann beispielsweise gemessen werden, wie gestresst eine Maus ist, wenn man sie gemeinsam mit einer Katze in einem Raum einsperrt. Der Stress-Score ist ein quantitatives Merkmal, das einen umso grösseren Wert annimmt, je gestresster die Maus ist. Wir messen für die Mäuse im Experiment aber nicht nur ihren Stress-Score, sondern bestimmen auch, zu welchem von vier bekannten Mäusestämmen sie gehören.

Wir sehen aus nebeneinander dargestellten Boxplots der Stress-Scores in verschiedenen Mäusestämmen (siehe Abb. 15) sofort, dass ein typischer Stress-Score im Stamm S1 viel höher liegt als im Stamm S4 – wir sehen aber auch, dass die Streuung von S1 über S2, S3 bis hin zu S4 immer kleiner wird, dasselbe gilt auch für den Median des Stress-Scores in den 4 Mäusestämmen. Bei genauem Hinsehen fällt aber eine Maus aus dem Stamm S4 auf, die extrem gestresst ist und ein absoluter Ausreisser ihrer Gattung ist. Vielleicht sollte man noch mal überprüfen, ob die vermessene Maus wirklich zum Stamm S4 gehört oder ob bei der Stressmessung etwas schief gelaufen ist.



→ Abb. 15

3

Deskriptive Statistik kategorieller Merkmale

3.1 Häufigkeitstabellen, Balken- und Kuchendiagramme, Modus

Einstiegsbeispiel Wie kommen Schweizer Schüler in die Schule? Zuerst einmal wollen wir einen Überblick darüber bekommen, wie die Schüler in einer bestimmten Klasse in die Schule kommen, d.h. welche Fortbewegungsmittel sie benutzen bzw. ob sie zu Fuss gehen. Am besten, wir fragen jeden Schüler nach seinem Fortbewegungsmittel für den Schulweg – da wir jeden Schüler fragen, führen wir also eine Vollerhebung dieser Klasse durch.

Um das Ergebnis dieser Umfrage übersichtlich z.B. an der Tafel festzuhalten, ist es sehr praktisch, sich erst mal einen Überblick über die möglichen Antwortmöglichkeiten bzw. Kategorien zu verschaffen und dann einfach eine Strichliste zu machen. Diese könnte in einer Klasse mit 25 Schülern wie folgt aussehen:

Velo: ||| zu Fuss: ||||||||||||
ÖV: |||||| anders: ||
Töff: ||

→ Abb. 16

Abbildung 16 Strichliste zur Darstellung einer Häufigkeitsverteilung

Abbildung 17 Häufigkeitstabelle der absoluten und relativen Häufigkeiten des gewählten Fortbewegungsmittels für den Schulweg der befragten Klassen.

Wir können aus dieser Strichliste leicht eine Häufigkeitstabelle erstellen, wobei wir aus den absoluten Häufigkeiten h_i der Gruppe i – z.B. der Velofahrer – die relativen Häufigkeiten f_i bestimmen können:

$$rel.\ Häufigkeit:\ f_i = \frac{h_i}{n} = \frac{h_i}{\sum_j h_j}$$

$$rel.\ Häufigkeit\ ÖV:\ f_{ÖV} = \frac{h_{ÖV}}{n} = \frac{6}{25} = 24\%$$

Ausprägung/ Kategorie des Merkmals Fortbewegungsmittel	absolute Häufigkeit h	relative Häufigkeit f in %
Velo	3	12
ÖV	6	24
Töff	2	8
zu Fuss	12	48
anders	2	8

→ Abb. 17

Wie können wir diese Häufigkeitstabelle grafisch darstellen?

Balken- und Kuchendiagramm
Die bekanntesten Darstellungen für die Häufigkeitsverteilung kategorieller Daten sind das Balkendiagramm und das Kuchendiagramm.
Beim Balkendiagramm gibt es für jede mögliche Kategorie einen Balken und die Höhe des Balkens entspricht der absoluten Häufigkeit, mit der diese Kategorie in der Stichprobe vorkommt.
Beim Kuchendiagramm gibt es für jede Kategorie ein Segment, dessen Fläche proportional zur relativen Häufigkeit dieser Kategorie in der Stichprobe ist.

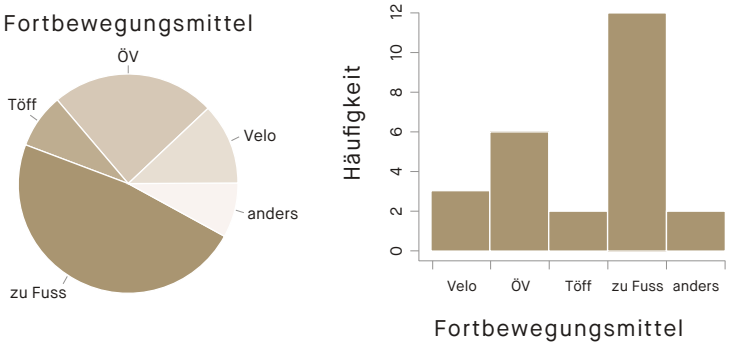


Abbildung 18 Das Kuchendiagramm in der linken Hälfte und das Balkendiagramm in der rechten Hälfte visualisieren die Häufigkeitsverteilung der gewählten Fortbewegungsmittel für den Schulweg in einer befragten Klasse.

→ Abb. 18

Vermitteln das Kuchendiagramm und das Balkendiagramm wirklich dieselbe Information?

Aus beiden Diagrammen in Abb. 18 können wir leicht ablesen, dass die Kategorie «zu Fuss» von allen möglichen Kategorien am häufigsten angegeben wurde.
Aus dem Balkendiagramm können wir ausserdem ablesen, dass 12 Schüler dieser Klasse zu Fuss zur Schule kommen – dem Kuchendiagramm können wir die absolute Häufigkeit einer bestimmten Kategorie nicht entnehmen. Dafür können wir dem Kuchendiagramm entnehmen, dass etwas weniger als die Hälfte der Klasse zu Fuss zur Schule kommt – um solch eine relative Häufigkeit aus dem Balkendiagramm entnehmen zu können, wäre erst etwas Rechnerei nötig:

$$rel.\ Häufigkeit:\ f_i = \frac{h_i}{n} = \frac{h_i}{\sum_j h_j}$$

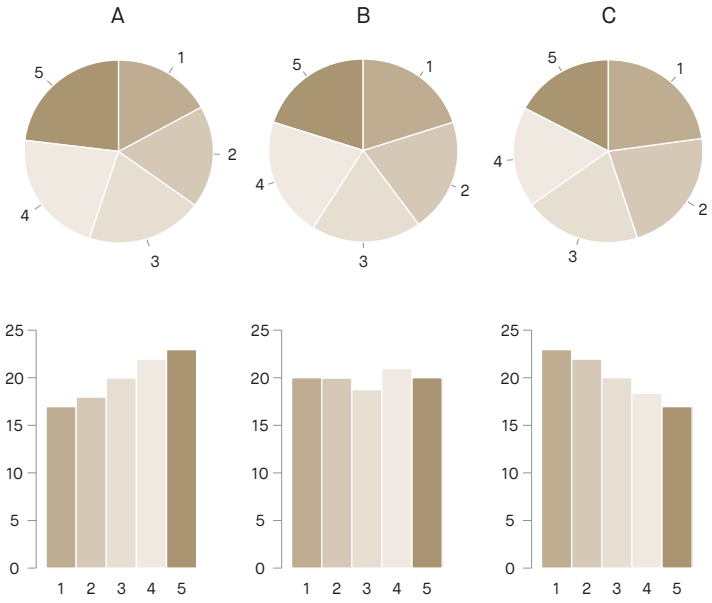
$$rel.\ Häufigkeit\ «zu\ Fuss»:\ f_{ÖV} = \frac{h_{zu\ Fuss}}{\sum_j h_j} = \frac{12}{3+6+2+12+2} = \frac{12}{25} = 48\%$$

Kennzahlen bei kategoriellen Daten Da man mit kategoriellen Daten nicht rechnen kann, gibt es in diesem Sinne auch keine Kennzahlen. Eine Ausnahme ist der Modus: Wenn eine der möglichen Kategorien deutlich häufiger vorkommt als die anderen Kategorien, dann ist bei dieser Kategorie der Modus. Beim Beispiel der Fortbewegungsmittel ist der Modus also bei der Kategorie «zu Fuss» (vgl. Abb. 18).

Abbildung 19 Übereinanderstehende Kuchendiagramme und Balkendiagramme visualisieren jeweils die gleichen Häufigkeitsverteilungen. Mit diesen Darstellungen kann experimentell untersucht werden, welcher Diagrammtyp leichter und genauer abgelesen werden kann. (Bildquelle: Wikipedia)

Was ist nun besser – das Kuchendiagramm oder das Balkendiagramm?

Aktivität Wir können ein kleines Experiment machen und 6 Schüler aus dem Zimmer schicken. Die Schüler werden einzeln hereingeholt und bekommen dann entweder die 3 Balkendiagramme oder die 3 Kuchendiagramme aus der Abb. 19 gezeigt. Sie sollen für jedes Diagramm möglichst rasch die Farben/Kategorien in eine Rangfolge entsprechend der relativen Häufigkeit bzw. Grösse der Kategorie bringen. Ihre abgelesene Rangfolge und die benötigte Zeit werden notiert. Im Normalfall zeigt sich, dass bei den Kuchendiagrammen mehr Unsicherheit herrscht und entsprechend mehr Zeit benötigt wird – es scheint für die meisten Menschen schwieriger zu sein, die Flächen der Segmente in einem Kuchendiagramm zu vergleichen als die Höhen verschiedener Balken im gleichen Diagramm.



→ Abb. 19

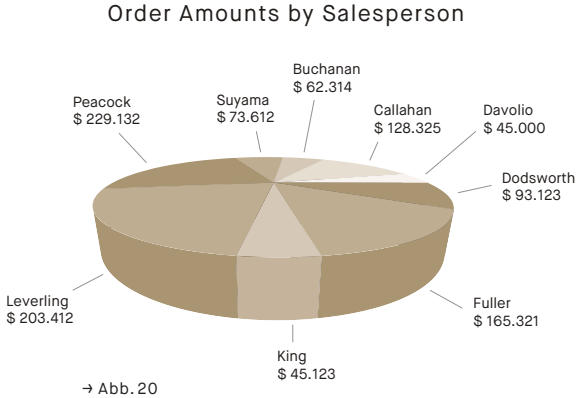
4

Manipulative Graphen

Kann ein einfaches Kuchendiagramm manipulativ sein?

Die Aussage «ein Bild sagt mehr als tausend Worte» ist oft wahr – Bilder sind sehr mächtig!

Auch einfache Graphen wie Kuchendiagramme können uns in Sekundenschnelle einen klaren Eindruck zur Datenlage vermitteln – oft benutzt bei Wahlberichten, wo möglichst effizient visualisiert werden soll, wie sich die abgegebenen Stimmen auf die verschiedenen Parteien verteilen. Es ist erst einmal schwer vorstellbar, aber auch mit einfachen Kuchendiagrammen kann absichtlich oder unabsichtlich manipuliert werden. Wir wollen hier ein paar Beispiele diskutieren.



→ Abb. 20

Abbildung 20 Kuchendiagramm mit 3-D-Effekt Quelle: http://zenassignments.com/creating-charts-with-ms-excel-2010/

Kuchendiagramme können mithilfe von Programmen wie beispielsweise Excel mit wenigen Mausklicks erzeugt werden. In Excel lassen sich Graphen auch leicht aufmotzen, z.B. durch 3-D-Effekte wie beispielsweise im Kuchendiagramm in Abb. 20. In diesem 3-D-Kuchendiagramm sieht es so aus, als ob die beiden schmalen Segmente im hinteren Bereich des Kuchens (Suyama und Buchanan) jeweils einen deutlich kleineren Anteil haben wie das gegenüber ganz vorne liegende Segment (King). Schaut man auf die angeschriebenen Zahlen, wird klar, dass dieser Eindruck täuscht. Wieso kommt es zum falschen Eindruck? Das menschliche Auge vergleicht automatisch die Flächen der Segmente – durch die 3-D-Darstellung wirkt das vorne liegende Segment (King) inklusive sichtbarem Rand wesentlich grösser als das hinten liegende Segment (Buchanan). 3-D-Kuchendiagramme sind also manipulativ, auch wenn dies wahrscheinlich oft gar nicht beabsichtigt ist!

Der Westdeutsche Rundfunk (WDR) hat unter der Rubrik «Wissen» einen sehenswerten Beitrag mit dem Titel «Die schlechtesten Grafiken der Welt» zusammengestellt – aus diesem Beispiel stammen die folgenden beiden Beispiele (www.wdr.de/tv/quarks/sendungsbeitraege/2006/1017/002_zahlen.jsp).

Das Kuchendiagramm in der linken Hälfte der Abb. 21 zeigt eine Grafik, wie sie von der Bildzeitung am 12.3.2006 veröffentlicht wurde. Damit sollte 3 Monate vor der WM die Beliebtheit des damaligen deutschen Bundestrainers Jürgen Klinsmann bei den Bild-Lesern dargestellt werden.

Wodurch hat sich dieses Kuchendiagramm für die Sammlung der schlechtesten Grafiken qualifiziert? Neben der 3-D-Verzerrung fällt bei näherer Betrachtung der angeschriebenen Prozentzahlen auf, dass sich die Prozentzahlen nicht zu 100% aufaddieren, sondern nur zu 62%! Dies ist absolut nicht im Sinne des Kuchendiagramms, bei dem die Flächen der Segmente proportional zu den dargestellten Anteilen sein sollte – d.h. die Fläche des Segments «22%» sollte doppelt so gross sein wie die Fläche des Segments «11%» – im linken Kuchendiagramm sieht das Segment «22%» aber mehr als doppelt so gross aus. Gibt man den fehlenden 38% ein eigenes Segment, so ergibt sich das Kuchendiagramm in der rechten Seite

der Abb. 21. Hier sind nun die Flächen tatsächlich einigermassen proportional zum dargestellten Anteil und die «22%»- Fläche etwa doppelt so gross wie die «11%»-Fläche.

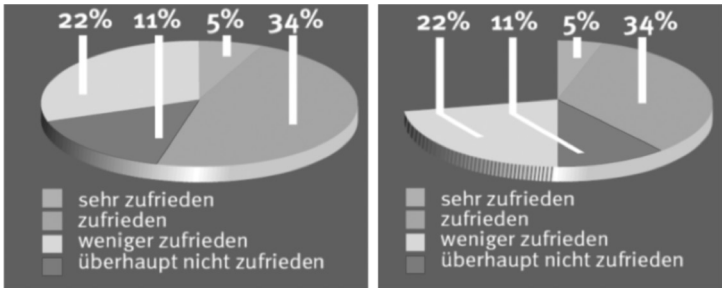


Abbildung 21 Kuchendiagramm zur Visualisierung der Zufriedenheitsangaben von befragten Lesern. Links ist die Darstellung aus der BILD-Zeitung, bei dem sich die Prozentangaben nicht zu 100% addieren; rechts ist das neu erstellte Kuchendiagramm, bei dem sich die Prozentzahlen zu 100% addieren.

→ Abb. 21

Aus derselben Sammlung stammt auch eine Grafik (vgl. Abb. 22), die aus einer Regierungserklärung der deutschen rot-grünen Regierung in der Agenda 2010 publiziert wurde, in der der Anstieg des Kindergeldes ins rechte Licht gerückt werden sollte. Dazu wurde die Kinderwagen-Visualisierung in der linken Hälfte der Abb. 22 benutzt. Der erste Blick sagt, dass das Kindergeld gewaltig gestiegen ist! Aber um wie viel eigentlich?

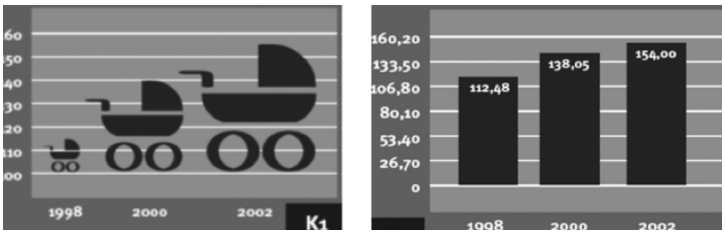


Abbildung 22 Kindergeld-Grafik: links ist die Grafik aus der Agenda 2010, rechts sind die Beträge der Kindergelder als Balkendiagramm dargestellt.

→ Abb. 22

Versuchen wir, das Kindergeld pro Monat und erstem Kind abzulesen, so müssen wir schauen, bis zu welchem Wert ein Kinderwagen in Abb. 22 geht und lesen für 1998 ca. 115 Euro, für 2000 ca. 140 Euro und für 2002 ca. 155 Euro ab. Benutzt man anstelle von Kinderwagen nun die gebräuchlichen Balken, so ergibt sich die rechte Grafik und der Zuwachs sieht bei Weitem weniger spektakulär aus. Der unterschiedliche Eindruck ist dadurch zu erklären, dass die Fläche der Kinderwagen nicht proportional zur Höhe der Kinderwagen ist, aber das Auge automatisch Flächen vergleicht. Dieser

Effekt kommt in der Kinderwagen-Grafik aus zwei Gründen zustande: 1) Die Kinderwagen sind fast kreisförmig und die Fläche eines Kreises wächst nicht proportional zum Durchmesser, sondern viel stärker, nämlich proportional zum Quadrat des Durchmessers. 2) Die Kinderwagen stehen nicht auf der Null-Linie, sondern auf der 100-Euro-Linie, daher ist die Höhe nicht proportional zum Betrag. Es darf spekuliert werden, ob die Regierung diesen Effekt gekannt hat und hier bewusst eingesetzt hat.

Aktivität Finden Sie in Presse und Internet Beispiele für manipulative Grafiken oder erstellen Sie selbst welche!

5

Korrelation, Kausalität und Störfaktoren

Ziel dieses Abschnitts ist es, einen Zusammenhang zwischen zwei quantitativen Grössen zu beschreiben und zu quantifizieren.

In Abb. 23 ist die italienische Euromünze dargestellt. Sie zeigt das Gemälde «Der vitruvianische Mensch» von Leonardo da Vinci. Wie man sieht, ist die Spannweite (der Abstand zwischen den Fingerspitzen bei ausgestreckten Armen) gleich der Körpergrösse. Diese und weitere Proportionen des menschlichen Körpers wurden von Leonardo da Vinci postuliert.

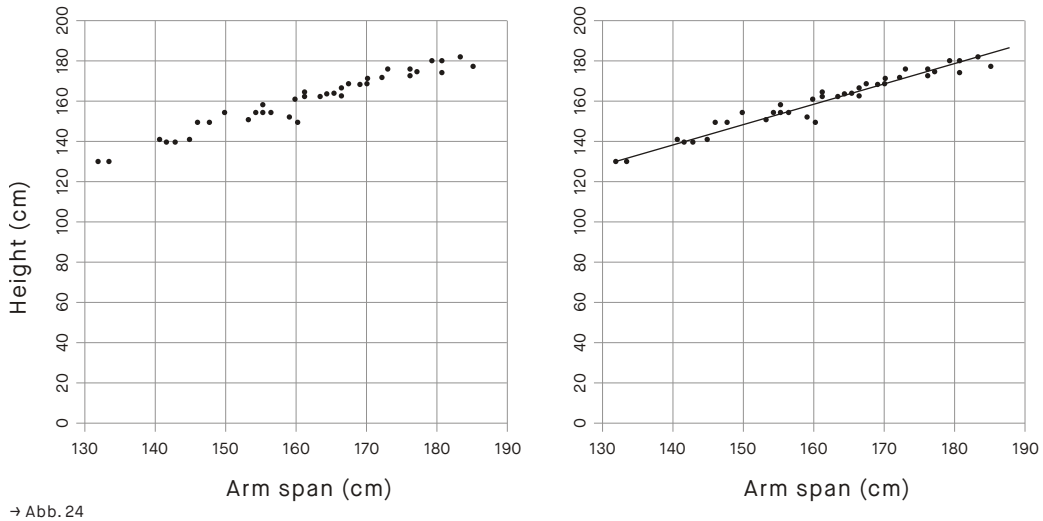
→ Abb. 23

Abbildung 23 Der vitruvianische Mensch nach Leonardo da Vinci. Bild: wiki commons.

[11] **URL** nzmaths.co.nz/resource/da-vincis-ratio

Abbildung 24 Streudiagramm zur Visualisierung des Zusammenhangs zwischen Körpergrösse (Height) und Armspanne (Arm span).

Doch ist es wirklich wahr, dass die Spannweite der Arme von der Körpergrösse abhängt? Die Zeiten, in denen man wie Aristoteles einfach locker behaupten konnte, Männer haben mehr Zähne als Frauen, sind vorbei. Wir wollen unsere Aussagen mit Experimenten belegen. Zum Beispiel könnte man Messungen der Spannweite und Körpergrösse in Ihrer Klasse machen. Solche Messungen z.B. wurden an neuseeländischen Lernenden durchgeführt[11]. In Abb. 24 ist die gemessene Spannweite gegenüber der Körpergrösse aufgetragen. Eine Darstellung zweier quantitativer Variablen nennt man Streudiagramm. Jeder Punkt in dem Diagramm stellt die zwei Messwerte (Spannweite und Grösse) eines Lernenden dar.



→ Abb. 24

Wie man sieht, besteht ein starker linearer Zusammenhang zwischen den beiden Grössen. Zeichnet man zusätzlich, wie in Abb. 24 rechts dargestellt, eine Gerade mit der Steigung 1 und dem y-Achsenabschnitt 0, so sieht man, dass alle Messungen um diese Gerade streuen: je grösser die Spannweite, desto grösser der Mensch. Hier gilt sogar, noch strenger, dass das Verhältnis von Spannweite zu Körpergrösse eins ist, d.h. ist eine Person um 10 cm grösser als eine Vergleichsperson, so hat sie auch eine 10 cm grössere Spannweite als ihre Vergleichsperson.

Korrelation: Zwei quantitative Merkmale sind miteinander korreliert, wenn im Streudiagramm ein linearer Zusammenhang sichtbar wird. Die Korrelation ist umso stärker, je enger die Punkte um eine gedachte Gerade herumstreuen – der Wert der Steigung spielt dabei keine Rolle, solange sie nur ungleich null ist.

Man spricht von einer positiven Korrelation zwischen den Merkmalen x und y, wenn gilt: Je grösser der x-Wert ist, desto grösser ist tendenziell der y-Wert.
Man spricht von einer negativen Korrelation zwischen den Merkmalen x und y, wenn gilt: Je grösser der x-Wert ist, desto kleiner ist tendenziell der y-Wert.

Wie würde ein Diagramm aussehen, falls kein Zusammenhang bestehen würde? In Abb. 25 sind drei Streudiagramme dargestellt. In der rechten Abbildung gilt (wie in dem Beispiel): je grösser x, desto grösser y. Man sagt: «x ist mit y (positiv) korreliert.» In der linken Abbildung gilt: Je grösser x, desto kleiner y. Hier spricht man von einer Antikorrelation. Und in der mittleren Abbildung hängt y gar nicht von x ab. Man könnte auch sagen, man weiss nichts über y, wenn man x kennt und umgekehrt. In diesem Fall gibt es keine Korrelation. Man sagt ausserdem, die Korrelation ist umso grösser, je enger die Datenpunkte um eine Linie mit einer positiven oder negativen Streuung herumstreuen. Man kann die Grösse der Korrelation auch berechnen, doch wir wollen das hier nicht tun, sondern uns auf eine visuelle Beurteilung beschränken.

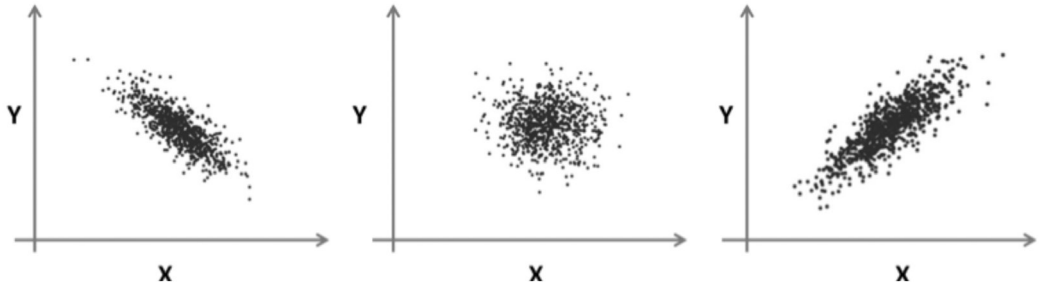


Abbildung 25 Streudiagramm für 3 Arten der Korrelation. Links: negative Korrelation (Anti-Korrelation), Mitte: keine Korrelation und rechts: positive Korrelation.

→ Abb. 25

Korrelation und Kausalität Oft ist man versucht, von einem linearen Zusammenhang, also einer bestehenden Korrelation zwischen zwei Variablen, auf einen kausalen Zusammenhang zu schliessen. Zum Beispiel gibt es eine Korrelation zwischen der Anzahl Fernsehapparate und der Anzahl Verbrechen in Schweizer Städten – je mehr Fernsehapparate

es gibt, umso mehr Verbrechen werden ausgeübt. Es darf daraus aber nicht geschlossen werden, dass Fernsehkonsum kriminell macht. Bei näherem Nachdenken würde man eher zu der Vermutung kommen, dass es in grösseren Städten sowohl mehr Fernsehapparate als auch mehr Verbrechen gibt und somit der Zusammenhang durch eine dritte Grösse, nämlich der Einwohnerzahl der Stadt, verursacht wurde. Eine solche dritte Grösse wird auch Störgrösse genannt und weiter unten noch näher diskutiert.

Es wäre tragisch, wenn wir aufgrund eines in den Daten zu sehenden Zusammenhangs nie auf Kausalität schliessen könnten. Das würde beispielsweise bedeuten, dass die Pharmaindustrie keine Möglichkeit hätte nachzuweisen, dass ihre Medikamente zumindest im Durchschnitt eine positive Wirkung bei Patienten mit bestimmten Krankheiten haben. Es gibt also durchaus Situationen, in denen Kausalität empirisch belegt werden kann, also aus Korrelation auf Kausalität geschlossen werden darf. Dazu müssen die Daten in einem dafür gut geplanten Experiment erhoben wurden, bei dem mögliche weitere Einflussgrössen auf die untersuchten Variablen streng kontrolliert wurden, rausgemittelt werden können, oder aber dafür gesorgt wurde, dass sich ausser den beiden betrachteten Variablen keine andere Variable während des Experiments ändert.

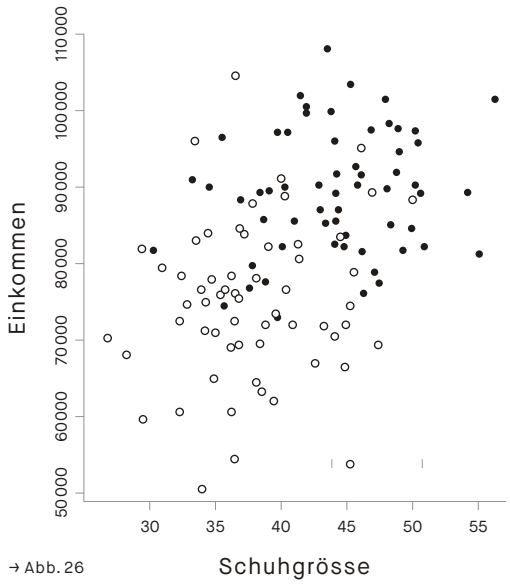


Abbildung 26 Streudiagramm, in dem Einkommen gegen Schuhgrösse abgetragen wurde. Männliche Teilnehmer in ausgefüllten blauen Punkten, weibliche in hellen Donuts.

→ Abb. 26

In vielen Datensammlungen kommen die Daten nicht aus geplanten Experimenten, sondern wurden einfach mal so gemessen oder abgefragt. Wenn dann aufgrund einer Korrelation auf einen kausalen Zusammenhang geschlossen wird, kann dies leicht zu Fehlschlüssen führen.

So zeigt sich zum Beispiel, dass eine erhebliche Korrelation zwischen Schuhgrösse und Einkommen besteht. Je grösser der Schuh, desto mehr verdient die jeweilige Person (siehe Abb. 26).

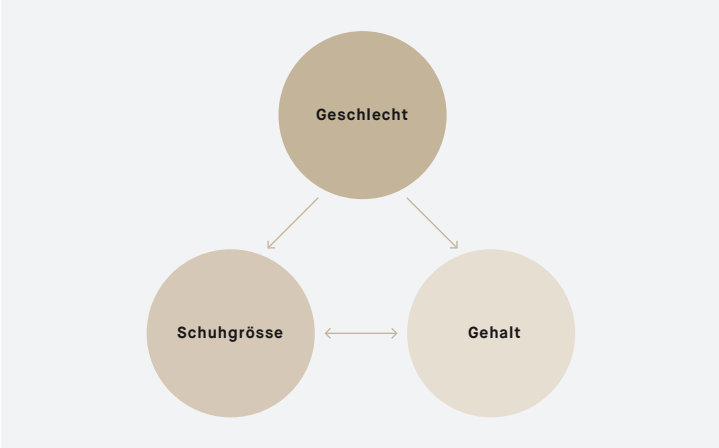


Abbildung 27 Die Korrelation zwischen der Schuhgrösse und dem Gehalt kann durch den Störfaktor Geschlecht erklärt werden.

→ Abb. 27

Wird bei den Gehaltsverhandlungen tatsächlich auf die Schuhgrösse geachtet? Die Antwort ist hoffentlich: nein. Die Lösung liegt vermutlich woanders. Leider verdienen Frauen immer noch weniger als ihre männlichen Kollegen mit ansonsten vergleichbaren Vorzügen und Erfahrungen. Frauen haben aber auch eher kleinere Füsse. Somit bestimmt die Schuhgrösse indirekt über das Geschlecht auch das Gehalt. Hier wäre also das Geschlecht ein indirekter Einflussfaktor für das Gehalt, der als Störfaktor (oder confounding factor) bezeichnet wird.

Es wäre sicher nicht einfach, einen vermuteten Zusammenhang zwischen Schuhgrösse und angebotenem Gehalt in einem Experiment zu untersuchen. Eine mögliche Idee könnte sein, einen Test mit Personalverantwortlichen zu machen. Dafür könnten z.B. 50 Personalverantwortliche eine Stellenausschreibung und eine zugehörige Bewerbung vorgelegt bekommen. Sie sollen dann ein

Gespräch mit dem Bewerber führen und im Anschluss ein angemessenes Gehalt vorschlagen. Wir schicken dann immer wieder die gleiche Person mit demselben Bewerbungsschreiben in der gleichen Aufmachung, jedoch mit verschieden grossen Schuhen zu den verschiedenen Personalverantwortlichen und schauen danach, ob das angebotene Gehalt mit der Grösse der Schuhe korreliert, die während des Vorstellungsgesprächs getragen wurden.

Korrelation und Kausalität Aus einer Korrelation zwischen zwei Variablen x und y (z.B. x: Schuhgrösse, y: Gehalt) alleine kann also nicht auf einen kausalen Zusammenhang geschlossen werden, d.h. es kann nicht gesagt werden, dass eine Änderung der x-Variablen eine Änderung der y-Variablen verursacht.

Nur wenn die Daten in einem entsprechend geplanten Experiment erhoben wurden, kann aus Korrelation auf Kausalität geschlossen werden. Ansonsten kann nie ausgeschlossen werden, dass der beobachtete Zusammenhang zwischen x und y durch eine sogenannte Störgrösse (z.B. Geschlecht) verursacht wird, welche sowohl x als auch y beeinflusst (Frauen haben oft kleinere Füsse und sie verdienen immer noch oft weniger Geld als Männer).

[12] Robert Matthews Teaching Statistics. Volume 22, Number 2, Summer 2000

Ein weiteres schönes Beispiel in diesem Sinne ist der Zusammenhang zwischen der Grösse der Storchpopulation und der Geburtenrate in einem Land. Es ist kaum anzunehmen, dass eine aufgeklärte Person einen kausalen Zusammenhang zwischen der Anzahl der Störche und der Geburten vermutet. Fragen Sie doch mal Ihre Klasse? Robert Matthews von der Aston Universität in Birmingham machte sich die Mühe, die Zahlen der brütenden Weissstörche (Ciconia ciconia) und die Anzahl der Geburten pro Jahr für verschiedene europäische Länder zusammenzutragen[12]. Schauen wir uns die Zahlen an:

	Störche	Kinder (1000/Jahr)
Albanien	100	83
Österreich	300	87
Belgien	1	118
Bulgarien	5000	117
Dänemark	9	59
Frankreich	140	774
Deutschland	3300	901
Griechenland	2500	106
Holland	4	188

	Störche	Kinder (1000/Jahr)
Ungarn	5000	124
Italien	5	551
Polen	30000	610
Portugal	1500	120
Rumänien	5000	367
Spanien	8000	439
Schweiz	150	82
Türkei	25000	1572

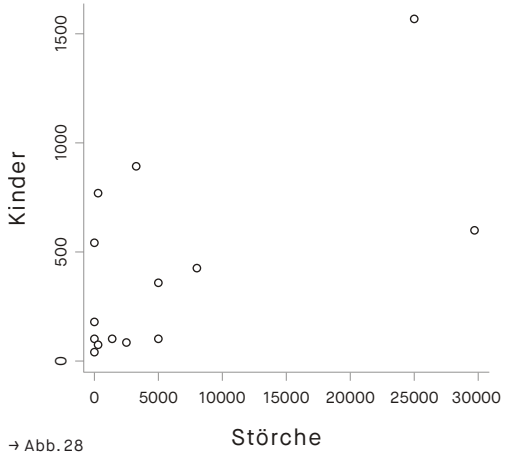


Abbildung 28 Streudiagramm, in dem Anzahl Kinder in 1000 pro Jahr gegen Anzahl Störche im Land abgetragen wurde.

[13] **URL** doris.bfs.de/jspui/handle/urn:nbn:de:0221-20100317939

→ Abb. 28

Die Tabelle und das Streudiagramm in Abb. 28 zeigen deutlich, dass ein starker Zusammenhang zwischen der Anzahl der Störche und der Anzahl der Kinder, die pro Jahr geboren werden, besteht. Doch anders als bei dem Beispiel mit den Schuhen ist es jetzt schwerer, den verantwortlichen Störfaktor (confounding factor) zu finden. Es wird spekuliert, dass die Grösse des Landes einen Einfluss sowohl auf die Geburtenrate als auch auf die Anzahl der Störche hat.

Oft ist es schwer, alle möglichen Störfaktoren zu finden oder auszuschliessen, dass ein Störfaktor vorhanden ist. In letzter Zeit machte z.B. in Deutschland eine Studie des Bundesamtes für Strahlenschutz zu Kinderkrebs in der Umgebung von Kernkraftwerken (KiKK) Furore[13]. Die Studie zeigt eine kleine, aber statistisch signifikante Erhöhung des Leukämierisikos von Kindern unter 5 Jahren, die im Umkreis von 5 km zu einem Kernkraftwerk aufwachsen. Der Schluss, dass die Nähe zum Kernkraftwerk der Grund für das erhöhte Leukämierisiko ist, liegt nahe, ist aber erst mal unzulässig, da man von einer Korrelation nicht auf einen kausalen Zusammenhang schliessen kann. Auch die Tatsache, dass die natürliche radioaktive Strahlung, welche stark schwankt, bis zum Faktor 1000 stärker ist, konnte und kann unseriöse Kernkraftgegner nicht davon abhalten zu behaupten, die KiKK-Studie beweise: Kernkraftwerke seien der Grund für die erhöhte Leukämierate unter den Kindern.
Es gibt sicher gute Gründe, die gegen die Kernkraft sprechen. Ernst zu nehmende Kernkraftgegner sollten sich daher

nicht durch ein falsches Spiel mit der Statistik diskreditieren. Eine Erklärung für die gefundene Korrelation in der KiKK-Studie könnten Störfaktoren sein. So ist zum Beispiel bekannt, dass der sozioökonomische Status der Eltern einen Einfluss auf das Krebsrisiko der Kinder hat – dabei gilt, je höher der Status der Eltern, desto höher das Krebsrisiko. Man erklärt dies dadurch, dass die Immunabwehr von Kindern, die wenig mit Keimen in Kontakt kommen, schwächer ist. Je schwächer die Immunabwehr, desto höher das Krebsrisiko. Neuere Studien, wie die CANUPIS-Studie (Childhood Cancer and Nuclear Power Plants in Switzerland)[14] oder eine gross angelegte britische Studie (siehe z.B. «Die Zeit» vom 23.5.2011)[15], die diesen und andere Störfaktoren einschliessen, konnten keinen statistisch abgesicherten Zusammenhang mehr finden.

[14] **URL** www.krebsliga.ch/de/fachpersonen/forschung/canupis_studie_zu_kinderkrebs_um_schweizer_kernkraftwerke
[15] **URL** www.zeit.de/wissen/gesundheit/2011-05/blutkrebs-akw-studie

ISBN 978-3-905745-71-9