ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

Informatik
& Security

fh
st.pölten

# Deskriptive Statistik

Dipl.-Ing. Hubert Schölnast, BSc
Stand: 25. Jänner 2023

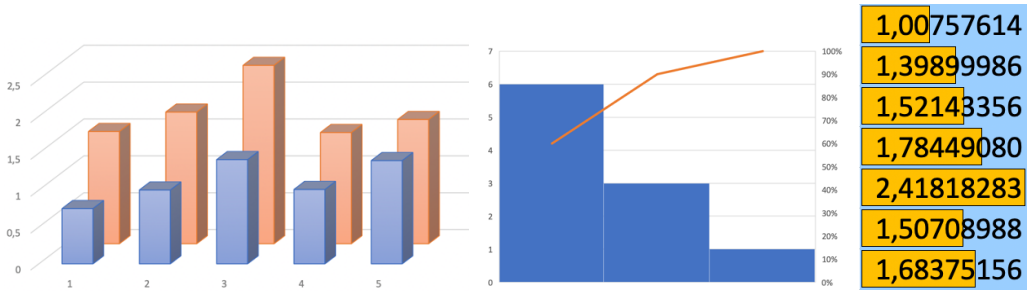
# Inhaltsverzeichnis

# 1 Grundbegriffe

## 1.1 Statistik[1]

Die drei Teilbereiche der Statistik sind:

- **Deskriptive[2] Statistik** (auch beschreibende oder empirische[3] Statistik genannt)
  Dieser Teilbereich behandelt die Beschreibung und übersichtliche Darstellung von Daten. Das umfasst sowohl die graphische Darstellung als auch die Berechnung von Kennzahlen. Die Datenvalidierung, also das Erkennen und Beheben von Fehlern in den Daten gehört ebenfalls zur beschreibenden Statistik, wird in diesem Dokument aber nicht näher beschrieben.




- **Explorative[4] Statistik** (auch: hypothesengenerierende Statistik, analytische Statistik, untersuchende Statistik oder Data-Mining). Dies ist eine Weiterführung und Vertiefung der deskriptiven Statistik, aber auch eine Vorstufe der induktiven Statistik. In der explorativen Statistik wird nach Zusammenhängen innerhalb der Daten gesucht, und es werden, innere Strukturen und Besonderheiten ermittelt.

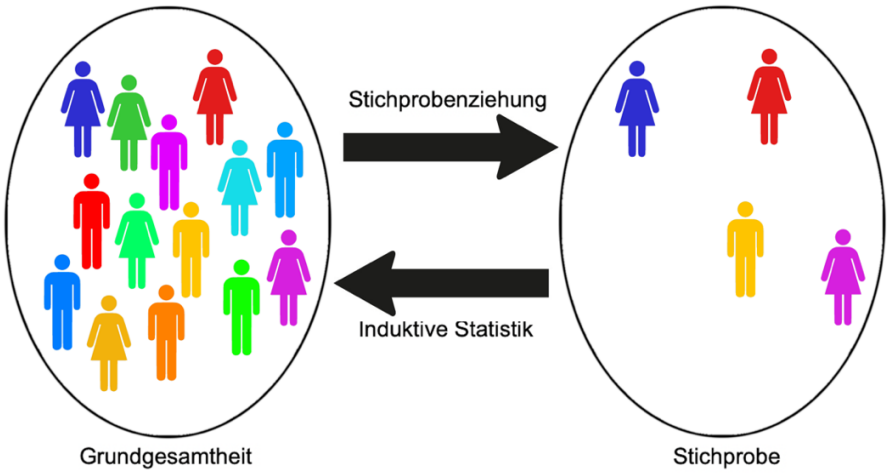
[1] Das italienische Wort *statista* bedeutet »Staatsmann, Politiker«. Es hat seinen Ursprung im neulateinischen *statisticum* »den Staat betreffend«. Das Wort *Statistik* wurde im Jahr 1749 von Gottfried Achenwall in den deutschen Sprachraum eingeführt und bedeutete ursprünglich die »Lehre von den Daten über den Staat«. Etwa 50 Jahre später hat der schottische Ökonom John Sinclair erstmals den Begriff *Statistik* in seiner heutigen Bedeutung »Sammeln und Auswerten von Daten« verwendet.

[2] Über französisch *descriptif* aus spätlateinisch *descriptivus* = beschreibend. (Lateinisch *scribere* = schreiben; *describere* = abschreiben, aufzeichnen, beschreiben)

[3] Griechisch *ἐμπειρικός (empeirikos)* = erfahren, erkennen; *ἐμπειρία (empeiria)* = Erfahrung, Erkenntnis im Sinn von »Einsicht bzw. Schlussfolgerung, die man aus einer Beobachtung gewonnen hat«

[4] »Explorieren« über französisch *explorer* aus Lateinisch *explorare* = erkunden, auskundschaften, ausforschen (Lateinisch *ex* = aus, durch, mittels, ...; *ploro* = schreien. Die ursprüngliche Bedeutung bezog sich vermutlich darauf, dass man ein Jagdgebiet ausgekundschaftet hat, indem man durch lautes Rufen das Wild aufgeschreckt hat.)

- **Induktive[5] Statistik** (auch: schließende Statistik, beurteilende Statistik, Inferenzstatistik oder mathematische Statistik)
  Hierbei werden Methoden aus der Wahrscheinlichkeitsrechnung verwendet, um aus Stichproben Rückschlüsse auf die Grundgesamtheit zu ziehen.




## 1.2 Statistische Einheiten, Erhebungseinheiten, Merkmalsträger

Das sind die Objekte, über die statistische Daten erhoben und verarbeitet werden.
Beispiele:

- Menschen
- Unternehmen (Firmen)
- Tage

## 1.3 Merkmale, statistische Variablen

Das sind die für die Statistik interessanten messbaren Größen der statistischen Einheiten.

Beispiele:

- Geschlecht, Körpergröße
- Anzahl der Zweigstellen, Jahresumsatz
- Sonnenscheindauer, Niederschlagsmenge

[5] »Induzieren« von Lateinisch *inducere* = hineinführen, hinführen. Die Bedeutung »vom Besonderen auf das Allgemeine schließen« geht auf Cicero zurück, der in seiner »Beweisführung durch Angabe ähnlicher Beispiele« den Leser von mehreren Einzelfällen auf den allgemeinen Zusammenhang hinführte.

## 1.4 Ausprägungen, Merkmalsausprägungen

Das sind die konkreten Werte, die von den Merkmalen angenommen werden können.

Beispiele:

- weiblich, 173 cm
- 6, 3,2 Mio €
- 8,3 Stunden, 17,92 l/m$^2$

## 1.5 Grundgesamtheit

Das ist die Menge aller statistischen Einheiten, über die man statistische Aussagen machen will.

Beispiele:

- alle Einwohner von St. Pölten
- alle Kfz-Werkstätten Österreichs
- alle Tage im Jahr 2020

## 1.6 Stichprobe

Das ist jene Teilmenge der Grundgesamtheit, von der tatsächlich auswertbare Daten vorliegen. Die Schlüsse, die sich aus einer Stichprobe ziehen lassen, sind umso valider, je typischer sie für die Grundgesamtheit ist. Diese Anforderung ist meist bei Zufallsstichproben erfüllt. Die Anzahl der Einheiten, aus denen eine Stichprobe besteht, wird Umfang der Stichprobe genannt.

### 1.6.1 Zufallsstichprobe

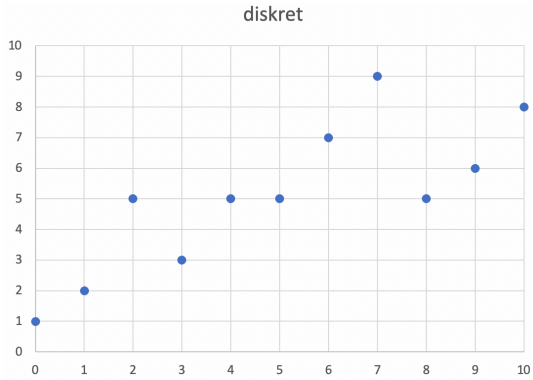
Das ist eine Stichprobe, die dadurch gewonnen wird, dass durch einen Zufallsprozess eine Stichprobe aus einer Grundgesamtheit ausgewählt wird, wobei jede einzelne statistische Einheit eine genau gleich große Wahrscheinlichkeit hat gewählt zu werden. Insbesondere darf diese Wahrscheinlichkeit auf keinen Fall von den zu untersuchenden Merkmalsausprägungen abhängen.

# 2 Merkmalstypen

## 2.1 Diskret vs. stetig

### Diskret[6]

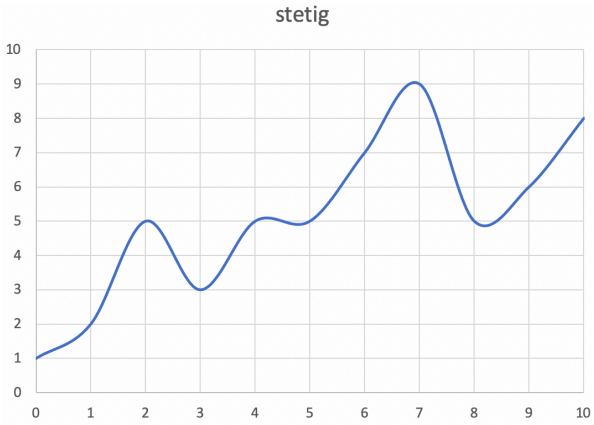
Ein Merkmal heißt diskret, wenn es zwischen zwei möglichen Werten kein Kontinuum an Zwischenwerten gibt. Das ist dann der Fall, wenn es nur eine endliche Menge möglicher Werte gibt, wobei die genaue Anzahl möglicher Werte gar nicht explizit bekannt sein muss. Aber auch Mengen mit unendlich vielen Elementen können diskret sein.




Diskret: Zwischen zwei beliebigen Werten befinden sich nur endlich viele (also nicht unendlich viele) andere Werte

- Zum Beispiel sind Gesellschaftsformen von Unternehmen diskrete Werte: AG, GmbH, OG, KG, ...
- Auch die Zahl der Kinder, die eine Frau geboren hat, ist ein diskreter Wert, obwohl die Obergrenze für den Wertebereich nicht genau bestimmt werden kann.

### Stetig[7]




Stetig: Zwischen zwei beliebigen Werten befinden sich unendlich viele andere Werte

Ein Merkmal nennt man stetig, wenn es bei zwei beliebig herausgegriffenen Werten immer denkbar ist, dass es Merkmalsträger gibt, die Werte dazwischen annehmen.

Als Beispiel sollen die Höhen von Häusern dienen: Wenn das Haus am Sonnenweg 13 genau 15,34 m hoch ist, und das Haus am Holzweg 4 einen cm höher, also 15,35 m hoch ist, dann ist es denkbar, dass es irgendwo auf diesem Planeten auch Häuser mit den Höhen 15,342, 15,345 oder 15,347634 m gibt.

---

[6] Lateinisch *discretus* = getrennt, unterschieden, bzw. *discernere* = absondern, unterscheiden, trennen

[7] Der Begriff ist von den stetigen Funktionen in der Mathematik entlehnt: Eine Funktion ist stetig, wenn der Graph einer Funktion keine Sprünge macht, wenn man ihn also ohne Abheben des Stiftes vom Papier zeichnen kann. Das setzt voraus, dass sich die Werte, die man als Argument einsetzen kann, kontinuierlich ändern können (dasselbe gilt dann natürlich auch für die Funktionswerte). Und diese kontinuierliche (also stufenlose) Veränderbarkeit gibt dem Begriff auch in der Statistik seine Bedeutung.

### Quasi-stetig

Das Beispiel mit den Höhen der Häuser zeigt, dass Stetigkeit zwar in vielen Fällen in der Theorie existieren mag, dass man die Werte in der Praxis aber oftmals nicht so genau misst. Entweder weil die Messmethode gar nicht genau genug ist oder weil man auf kleine Unterschiede gar keinen Wert legt, und stattdessen mit gerundeten Werten arbeitet. Wenn es bei der Feinheit der Unterteilung eine Untergrenze gibt (z.B., wenn man die Höhe von Häusern nur auf ganze Zentimeter genau erfasst), steht einem eigentlich nur ein diskreter Wertevorrat zur Verfügung, nämlich nur ganzzahlige Vielfache von 1 cm im Bereich zwischen 0 und 100.000 cm (das entspricht 1 km, höhere Häuser gibt es derzeit nicht). Das sind maximal 100.000 verschiedene Häuserhöhen, also ein endlicher und damit diskreter Wertebereich.

Wenn, wie im vorliegenden Beispiel, die Feinheit, mit der die Werte erfasst werden, aber einen deutlich kleineren Wert hat als die Bandbreite jenes Bereichs, den die Werte annehmen können, spricht man von quasi-stetigen Merkmalen, und man spricht auch davon, dass die Merkmale in Klassen eingeteilt sind: Alle Häuser, die zwischen 15,335 und 15,345 m hoch sind, werden zur Klasse der Häuser mit der Höhe 15,34 gezählt.

### Klassen

Die Einteilung in Klassen wird aber noch wichtiger, wenn man Werte hat, die man eigentlich recht genau kennt, sie aber in eine absichtlich klein gewählte Zahl von Klassen einteilt. Beispielsweise kann man Gehälter, die man auf Cent genau kennt, in die drei Klassen »niedrig«, »mittel« und »hoch« einteilen, wenn das für die Auswertung der Daten von Vorteil ist.


In so einem Fall spricht man auch nicht mehr von quasi-stetigen Merkmalsausprägungen, sondern ganz klar von diskreten Merkmalen.

Es ist übrigens gar nicht notwendig, dass die tatsächlichen Werte (wenn man sie genau genug messen würde) innerhalb einer Klasse wirklich unterschiedlich sein können. Das Einkommen eines arbeitenden Menschen beträgt immer ein ganzzahliges Vielfaches eines Cents (oder einer anderen Währungseinheit). Das Merkmal Einkommen ist daher streng genommen von vornherein ein diskretes Merkmal, denn es ist überhaupt nicht möglich, dass jemand ein Gehalt hat, das z.B. zwischen den Werten 1453,46 € und 1453,47 € liegt.

Weil aber bei den meisten Einkommen ein Cent mehr oder weniger keinen spürbaren Unterschied ausmacht, bzw. weil die typischen Einkommen erheblich größer als 1 Cent sind, betrachtet man solche Merkmale trotzdem zumindest als quasi-stetig, meist sogar als stetig.

**Gerundete Werte**

Streng genommen hat man es nie mit wirklich stetigen Merkmalen zu tun, weil das eine unendlich große Genauigkeit bei der Datenerfassung erfordern würde, die aus praktischen Gründen, und oftmals sogar aus ganz prinzipiellen Ursachen nicht möglich ist. Man muss bei jeder Messung runden. Man hat in der Praxis daher entweder ganz klassisch diskrete Werte oder quasistetige Werte. Stetige Werte sind ein mathematisches Ideal, das in der Praxis nicht vorkommt. Aber diese Idealvorstellung macht die mathematische Behandlung der Daten erst möglich, und man kann sich daran umso besser annähern, je genauer die Daten sind, also je mehr Stellen man nach dem Runden übriglässt.

## 2.2 Messskalen

Merkmale unterscheiden sich nicht nur dadurch, ob es Zwischenwerte gibt, sondern noch viel mehr dadurch, welche Operationen man mit ihnen durchführen kann. Familiennamen kann man beispielsweise nicht addieren, Körpergrößen schon. Bei Telefonnummern und Postleitzahlen wäre eine Addition zwar theoretisch möglich, macht aber keinen Sinn.

Es gibt auch Merkmale, die man nicht sortieren kann, dazu gehören z.B. Farben. Man kann nur die Namen der Farben sortieren, diese Namen hängen aber von der gewählten Sprache ab. Daher hängt die Reihenfolge, die herauskommt, nicht von den Farben selbst ab. Auch wenn man Farben nach der Helligkeit, ihren Blauanteil oder ähnlichen Merkmalen sortiert, sortiert man nicht nach der Farbe selbst, sondern nach einem Wert, den man willkürlich von der Farbe ableitet.

Man unterscheidet diese vier Messskalen:

### 2.2.1 Nominalskaliert[8,9]

Wenn die Ausprägungen des Merkmals Namen sind, spricht man von nominalskalierten Merkmalen. Mit Namen lassen sich keine Rechenoperationen durchführen (Summe, Differenz, usw.), und es gibt keine natürliche Reihenfolge, in der man die Merkmale anordnen kann. Möglich ist natürlich eine alphabetische Sortierung, diese hängt aber von einer Sprache ab und kann in einer anderen Sprache zu einem anderen Ergebnis führen.

Dieses Sortier-Kriterium ist wichtig, wenn es um die Frage geht, ob man z.B. einen Median ermitteln kann (siehe weiter unten). Das ist bei nominalskalierten Merkmalen nicht möglich, bei allen anderen aber schon.

---

[8] »Nominal« von lateinisch *nomen* = der Name

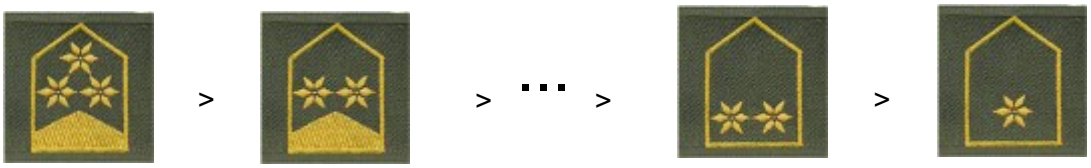
[9] »Skala«: Über italienisch *scala* aus lateinisch *scala* = Stufe, Leiter, Treppe (Lateinisch *scandere* = steigen)

Beispiele für nominalskalierte Merkmale sind alle Arten von Namen, KFZ-Kennzeichen, Farben usw. Aber auch Werte, die auf den ersten Blick wie Zahlen aussehen, können Ausprägungen nominalskalierter Merkmale sein, z.B. die oben schon erwähnten Postleitzahlen und Telefonnummern. Das liegt daran, das Nummern im mathematischen Sinn keine Zahlen sind, sondern systematische Namen. Das gilt manchmal sogar dann, wenn diese Merkmale das Wort »Zahl« im Namen tragen, wie das bei den Postleitzahlen der Fall ist. Postleitzahlen sind nur Namen für Zustellgebiete. Telefonnummern sind Namen für Telefonanschlüsse.

### 2.2.2 **Ordinalskaliert**[10]

Ein Merkmal ist ordinalskaliert, wenn seine Ausprägungen einer natürlichen Ordnung gehorchen, das heißt, wenn man die Werte sinnvollerweise nach »größer«, »besser« oder einem ähnlichen Kriterium sortieren kann, ohne dass die Unterschiede zwischen zwei Werten eine besondere Bedeutung hätten.

Militärische Dienstgrade gehören in diese Klasse: Ein Oberst ist ranghöher als ein Oberstleutnant, und beide sind ranghöher als ein Oberleutnant. Ein Leutnant ist rangniedriger als alle zuvor genannten. Aber die Tatsache, dass zwischen Oberst und Oberleutnant zufällig gleich viele Dienstgrade liegen wie zwischen Oberstleutnant und Leutnant, hat keinerlei Bedeutung.



Oberst > Oberstleutnant > ... > Oberleutnant > Leutnant

Auch stetige Merkmale können ordinalskaliert sein, z.B. die subjektive Zufriedenheit mit einem Produkt (»total zufrieden«, »eigentlich eh voll zufrieden aber mit kleinen Abstrichen«, »ziemlich zufrieden«, usw.)

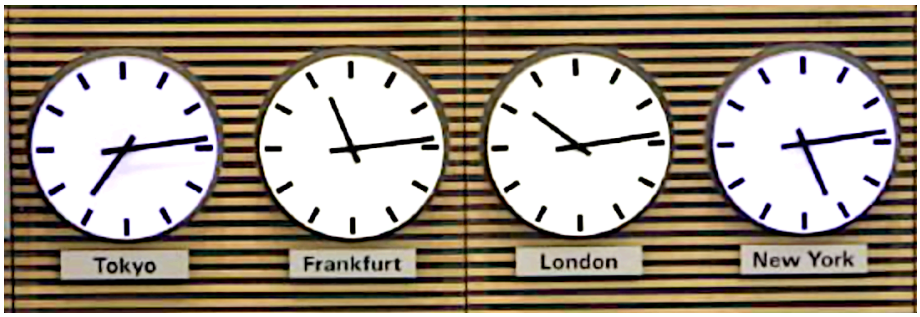
### 2.2.3 **Intervallskaliert**[11]

Ein Merkmal ist intervallskaliert, wenn es nicht nur eine natürliche Reihenfolge gibt, sondern wenn auch die Abstände (Differenzen) unterschiedlicher Ausprägungen einen Sinn tragen, allerdings ohne, dass auch die Verhältnisse (Quotienten) sinnvoll sind.

Beispiele dafür sind Uhrzeiten, Jahreszahlen oder Celsius-Temperaturen. Es ist durchaus sinnvoll, Differenzen solcher Werte zu betrachten (»Elke kam 14 Stunden vor Michael an.« »Heute hat es 5 Grad mehr als gestern«), aber Verhältnisse ergeben keinen Sinn (»~~Timon stand heute dreimal so spät auf wie Leon.~~« »~~Gestern war es nur halb so kalt wie heute.~~«)

[10] lateinisch *ordo* = Rang, Reihenfolge
[11] lateinisch *intervallum* bzw. *inter vallos* = der Zwischenraum zwischen den Schanz-Pfählen einer Verteidigungsanlage.

Das ist immer dann der Fall, wenn der Nullpunkt einer Skala keine natürliche Eigenschaft der Skala ist, sondern willkürlich gewählt wurde, bzw. wenn die Skala auch mit einem anderen Nullpunkt genauso gut funktionieren würde (andere Zeitzone, islamischer statt christlicher Kalender, Fahrenheit statt Celsius usw.)




### 2.2.4 **Verhältnisskaliert**

Ein Merkmal ist verhältnisskaliert, wenn es eine Reihenfolge, bedeutungsvolle Abstände und auch bedeutungsvolle Verhältnisse gibt. Das setzt voraus, dass die Messskala einen Nullpunkt besitzt, der sich ganz natürlich aus der Art des Merkmals ergibt.

Beispiele sind Körpergröße, Alter, Einkommen. Die Aussagen »Michael ist genau doppelt so groß wie sein Sohn«, »Dr. Taschner verdient dreimal so viel wie seine Sekretärin« und »meine Katze wurde doppelt so alt wie mein Hamster« sind durchaus sinnvoll.

### 2.2.5 **Überbegriffe**

**metrische[12] Skala = Kardinalskala[13]**

Weil sowohl die Intervall- als auch die Verhältnisskala ein Maßsystem verwenden (man kann in diesen beiden Skalen etwas messen), fasst man beide unter den Begriffen metrische Skala bzw. Kardinalskala zusammen.

**Kategorialskala[14]**

Der Überbegriff für die beiden anderen Skalen, die nicht geeignet sind, etwas abzumessen (Nominal und Ordinal), lautet Kategorialskala.

---

[12] Altgriechisch *μετρῶ (metro)* = messen.
[13] Lateinisch *cardo* = Türangel, Angelpunkt, also der Punkt, um den sich alles dreht. Daraus entstand die Bedeutung *kardinal* = grundlegend. Im 19. Jhd. wurden in der Mathematik Kardinalzahlen als jene Zahlen eingeführt, die die Größen (Mächtigkeiten) von Mengen wiedergeben.
[14] Altgriechisch *κατηγορία (kategoria)* = Eigenschaft.

## 2.3 Eigenschaften der Messskalen

### 2.3.1 Erlaubte Operationen

| | $A = B$<br>$A \neq B$ | sort | $A < B$<br>$A > B$ | $A - B$ | $\frac{A+B}{2}$ | $A + (X - Y)$ | $A + B$ | $\frac{A}{B}$ | $n \cdot A$ |
|---|---|---|---|---|---|---|---|---|---|
| Nominal | ✅ | ❗ | 🚫 | 🚫 | 🚫 | 🚫 | 🚫 | 🚫 | 🚫 |
| Ordinal | ✅ | ✅ | ✅ | 🚫 | 🚫 | 🚫 | 🚫 | 🚫 | 🚫 |
| Intervall | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | 🚫 | 🚫 | 🚫 |
| Verhältnis | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ |

- **Prüfen auf Gleichheit oder Ungleichheit**
  Ist in jeder Skala möglich

- **Sortieren**
  Bei Nominalskalen kann man durch Sortieren nur Gruppen von Merkmalsträgern mit gleichen Merkmalsausprägungen finden. Sortierte Nominalskalen eignen sich nicht, um den Median, Quantille oder Quartille zu ermitteln. Dies ist aber bei allen anderen Messskalen möglich.

- **Größer/kleiner-Vergleich (besser/schlechter, länger/kürzer usw.)**
  Nicht möglich in Nominalskalen, möglich in allen anderen Skalen

- **Differenz zweier Ausprägungen bilden**
  Nur bei Kardinalskalen (Intervall und Verhältnis), nicht aber bei Kategorialskalen (Nominal- und Ordinalskala).

- **Arithmetisches Mittel zweier Ausprägungen bilden**
  (Auch gewichtetes arithmetisches Mittel, aber keine nichtarithmetischen Mittelwerte.)
  Nur bei Kardinalskalen (Intervall und Verhältnis), nicht aber bei Kategorialskalen (Nominal- und Ordinalskala).

- **Zur Ausprägung eines Merkmals die Differenz zweier Ausprägungen desselben Merkmals addieren**
  Nur bei Kardinalskalen (Intervall und Verhältnis), nicht aber bei Kategorialskalen (Nominal- und Ordinalskala).
  Beispiel:
  Wenn ich einen Topf voll Wasser mit 25 Grad Celsius habe, und ihn um 30 Grad erwärme, hat er dann 55 Grad.
  25 Grad = eine Temperatur (Ausprägung eines Merkmals)
  30 Grad = eine Temperaturdifferenz (Differenz zweier Ausprägungen)
  55 Grad = Summe aus einer Temperatur und einer Temperaturdifferenz

- **Zwei Ausprägungen addieren**
  Nur bei Verhältnisskalen
  Beispiel:
  Hans verdient 2700 €, Maria verdient 2800 €, zusammen haben sie 5500 €.
  Gegenbeispiel Intervallskala:
  Wenn ich einen Topf Waser mit einer Temperatur von 25 Grad habe, und einen anderen Topf mit 30 Grad, haben die beiden Töpfe zusammen nicht 55 Grad. (Hier sind die 30 Grad keine Temperaturdifferenz, sondern eine Temperatur!)

- **Den Quotient zweier Ausprägungen bilden**
  Nur bei Verhältnisskalen. (Das Ergebnis ist eine dimensionslose Zahl)
  Beispiel:
  Ulrich verdient 3400 €, Laura verdient 1700 €, Laura verdient die Hälfte von Ulrich.

- **Eine Ausprägung mit einer dimensionslosen Zahl multiplizieren**
  Nur bei Verhältnisskalen
  Beispiel:
  Heinrich verdient 2000 €. Nach der Erhöhung um 10% (= Multiplikation mit 1,1) verdient er 2200 €.

## 2.4 Qualitativ vs. quantitativ

### Qualitative[15] Merkmale

Merkmale, die eine Zuordnung zu einer Kategorie wiedergeben, sind qualitative Merkmale. Dazu gehören alle nominalskalierten Merkmale. Ordinalskalierte Merkmale werden dann zu den qualitativen Merkmalen gezählt, wenn es nur eine einigermaßen überschaubare Anzahl möglicher Ausprägungen gibt. Die weiter oben beschriebenen militärischen Dienstgrade sind ein Beispiel dafür.

### Quantitative[16] Merkmale

Wenn ein Merkmal eine Intensität oder ein Ausmaß wiedergibt, handelt es sich um ein quantitatives Merkmal. Daher sind alle intervall- und verhältnisskalierten Merkmale quantitativ.

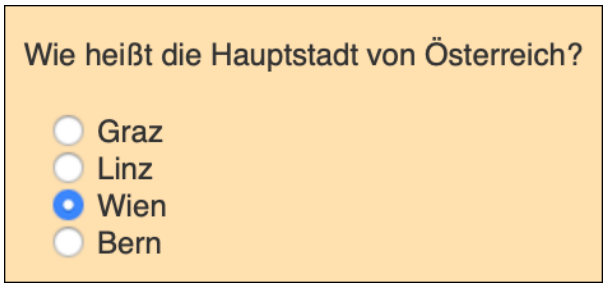
[15] Lateinisch *qualis?* = wie? von welcher Art? bzw. *qualitas* = Beschaffenheit, Eigenschaft, später (Mittelalter) auch Merkmal, Kategorie.
[16] Lateinisch *quantus?* = wie viel? wie groß? bzw. *quantum* = Dosis, Menge

## 2.5 Häufbare und nichthäufbare Merkmale

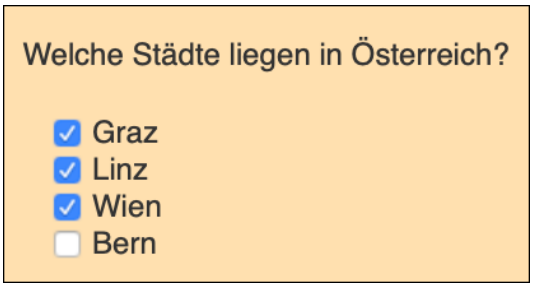
Alle zuvor genannten Beispiele für Merkmale waren nichthäufbar. Das bedeutet, dass jeder Merkmalsträger nur eine Instanz einer Merkmalsausprägung haben kann. Jeder Mensch hat nur eine Körpergröße, jedes Unternehmen hat nur einen Jahresumsatz (für ein bestimmtes Jahr), jeder Soldat hat nur einen Dienstgrad usw.

Aber es gibt auch Merkmale, von denen mehrere verschiedene Ausprägungen zu ein und demselben Merkmalsträger gehören können. Ein Akademiker kann nicht nur einen akademischen Grad haben. Viele von ihnen sind zugleich Bachelor, Master und vielleicht auch noch Doktor.

Menschen haben mitunter mehrere Hobbys, mehrere Wohnsitze oder beherrschen mehrere Sprachen. Solche Merkmale nennt man häufbar.




Nichthäufbares Merkmal:
Hauptstadt eines Landes




Häufbares Merkmal:
Städte eines Landes

Häufbare Merkmale verursachen bei Auswertungen oft Probleme, was aber dadurch relativiert wird, dass häufbare Merkmale so gut wie immer nominalskaliert sind und man damit ohnehin nicht viele Operationen durchführen kann. Nominalskalierte Merkmale kann man nur zählen und deren Häufigkeiten betrachten. Dieselben Operationen kann man mit den einzelnen Instanzen häufbarer Merkmalsausprägungen auch machen.

Die Voraussetzung dafür ist, dass man die einzelnen Werte, die zu einem Merkmalsträger gehören, getrennt speichern und verarbeiten kann. Dafür gibt es z.B. hierarchische Datenbanken (die aber durch gewöhnliche relationale Datenbanken nachgebildet werden können, indem man mehrere Tabellen verwendet).

Da das alles mit Aufwand verbunden ist, dem nicht immer ein ausreichend großer Nutzen gegenübersteht, will die Verwendung häufbarer Merkmale immer wohl überlegt sein.

Als Faustregel gilt daher: Vermeiden Sie bei der Datengewinnung häufbare Merkmale, wann immer es Ihnen möglich ist.

# 3 Univariate Verteilungen

## 3.1 Begriffsdefinition

**univariat, multivariat (bivariat, trivariat, ...)**[17]

Wenn die Daten einer Stichprobe (oder auch der Grundgesamtheit) erhoben werden, kann man, je nach Aufgabenstellung, pro Merkmalsträger entweder nur ein Merkmal erfassen (z.B. bei jeder Person nur die Körpergröße) oder man kann mehrere Merkmale erfassen (bei jeder Person: Vorname, Nachname, Körpergröße, Augenfarbe, Jahreseinkommen, Schuhgröße, Gewicht usw.).

Wenn nur ein Merkmal erfasst oder ausgewertet wird, spricht man von univariaten oder auch von eindimensionalen[18] Daten. Wenn Sie diese Daten in eine Tabelle eintragen, kommen Sie mit nur einer einzigen Spalte aus.

Im anderen Fall (mehrere Merkmale pro statistischer Einheit) liegen multivariaten Daten vor. Man spricht dann auch von mehrdimensionalen Daten, die in der Tabelle mehrere Spalten benötigen.

Gelegentlich findet man auch die Begriffe bivariate (zweidimensionale) und trivariate (dreidimensionale) Daten. Begriffe, die ähnlich gebildet werden und sich auf höhere Anzahlen von Merkmalen pro Merkmalsträger beziehen, sind denkbar, aber kaum üblich.

In diesem Kapitel geht es um univariate Häufigkeitsverteilungen. Sie bilden die Grundlage für multivariate Verteilungen.

## 3.2 Urliste

Wenn die Daten einer Stichprobe erhoben werden, werden die Daten der Reihe nach in einer Liste notiert. Das ist die Urliste. In ihr können Werte mehrfach vorkommen (weil verschiedene Merkmalsträger dieselbe Ausprägung des erfassten Merkmals haben können).

---

[17] *Unus, dou, tres, quattuor* usw. sind lateinische Zahlwörter. Sie bedeuten eins, zwei, drei vier usw.
*Multi-* ist die lateinische Kompositionsform von *multus. Multus* = viel, groß, stark.
*Univariat* bedeutet »von einer Variablen abhängig«. Die Wörter *Variable, variat, variieren* und *variabel* stammen vom lateinischen Verb *variare* (Abwechslung in etwas bringen, verändern, verschieden sein, bunt machen, bunt sein, abwechseln, veränderlich sein) bzw. dem Adjektiv *varius* (mannigfaltig, bunt, abwechselnd, verschiedenartig) ab.

[18] »Dimension«: Von lateinisch *dimensio* = das Vermessen, die Vermessung, das Ausmessen, abgeleitet vom Verb *dimetiri* = abmessen, ausmessen, vermessen, zusammengesetzt aus di- (dis-) (diese Vorsilbe hat viele Bedeutungen, hier drückt sie eine allgemeine Erweiterung des damit verbundenen Begriffs aus) und *metiri* = messen.

Die Reihenfolge, in der die Daten in der Liste stehen, hat bei univariaten Verteilungen keine Bedeutung. Sollte die Reihenfolge dennoch wichtig sein, liegt keine univariate Verteilung mehr vor. Dann kommt nämlich zu jedem Eintrag in der Liste noch eine laufende Nummer, ein Zeitstempel oder ein ähnliches Element hinzu, so dass dann pro Merkmalsträger zwei Merkmale in der Liste stehen, nämlich die laufende Nummer (oder der Zeitstempel) und das „eigentliche“ Merkmal.

**Beispiel:**

Ein Bäcker produziert Semmeln. Das vom Gesetzgeber vorgeschriebene Mindestgewicht dafür beträgt in Österreich 46 g (§ 76 Lebensmittelsicherheits- und Verbraucherschutzgesetz, bzw. Österreichisches Lebensmittelbuch).

Ein Mitarbeiter des Marktamts geht zu dem Bäcker, weist sich als Vertreter der Behörde aus, kauft mehrere Semmeln und wiegt sie später mit einer geeichten Waage ab. Dabei werden diese Gewichte festgestellt:

| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 47 | 49 | 49 | 49 | 48 | 52 | 50 | 46 | 48 | 48 | 51 | 54 | 47 | 47 | 50 | 51 | 52 | 48 | 52 | 48 | 48 | 46 | 49 | 47 | 51 |

Diese Urliste enthält Einträge von 25 Merkmalsträgern, der Stichprobenumfang beträgt also 25.

$$n = 25$$

Ganz offensichtlich sind in der Stichprobe nur Semmeln mit einem gültigen Gewicht enthalten.

Wir haben es hier dem ersten Anschein nach mit diskreten Werten zu tun, plausibler ist aber die Interpretation, dass die Semmeln in Wahrheit Gewichte haben, die nicht immer genau ein Vielfachen von 1 g sind, dass aber entweder die Waage nur ganzzahlige Gewichte anzeigt oder dass beim Notieren der Gewichte auf ganze Gramm gerundet wurde. In beiden Fällen wurden dadurch die Semmeln in ganzzahlige Gewichtsklassen eingeteilt.

Gewichte sind verhältnisskalierte Merkmale. (Die Aussage, dass ein Exemplar doppelt so schwer wie ein anders ist, ist eine sinnvolle Aussage.) Das gilt auch wenn die Gewichte gerundet wurden.

## 3.3 Absolute und relative Häufigkeiten

Es fällt auf, dass einige Gewichte mehrfach vorkommen. Nachdem das Prüfen auf Gleichheit immer möglich ist, kann man bei jeder Art von Merkmal feststellen, ob es in der Urliste gleiche Werte gibt, und man kann auch immer zählen, wie viele Merkmalsträger jeweils gleiche Merkmalsausprägungen haben.

Eine solche Zählung wird einfacher, wenn man die Urliste sortieren kann. Beim Sortieren muss man Größer/Kleiner-Vergleiche durchführen, das ist streng genommen nur bei ordinal-,

intervall- oder verhältnisskalierten Merkmalen möglich. Man kann aber auch nominalskalierte Daten sortieren (z.B. alphabetisch) und erhält dadurch ebenfalls sortierte Urlisten, die das Zählen vereinfachen. Man muss sich dabei nur darüber im Klaren sein, dass die erhaltene Sortierreihenfolge keine Rückschlüsse auf die Zusammenhänge innerhalb der Daten erlaubt. (Insbesondere kann man bei nominalskalierten Daten auch dann keinen Median und ähnliche Maße ermitteln, wenn die Daten nach diesem Merkmal sortiert sind.)

Die Semmelgewichte sind aber ohnehin verhältnisskaliert, man kann sie also problemlos sortieren, und erhält dabei eine Reihenfolge, die auch für die Merkmalsausprägungen sinnvoll ist.

Wir erhalten diese sortierte Urliste (darunter wurde in jeder Gewichtsklasse von links nach rechts durchgezählt):

| 46 | 46 | 47 | 47 | 47 | 47 | 48 | 48 | 48 | 48 | 48 | 48 | 49 | 49 | 49 | 49 | 50 | 50 | 51 | 51 | 51 | 52 | 52 | 52 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **2** | 1 | 2 | 3 | **4** | 1 | 2 | 3 | 4 | 5 | **6** | 1 | 2 | 3 | **4** | 1 | **2** | 1 | 2 | **3** | 1 | 2 | **3** | **1** |

Und dies sind die absoluten Häufigkeiten der Gewichte in der vorliegenden Stichprobe (Anzahl = absolute Häufigkeit):

| Gewicht | $h_{abs}$ |
|---|---|
| 46 | 2 |
| 47 | 4 |
| 48 | 6 |
| 49 | 4 |
| 50 | 2 |
| 51 | 3 |
| 52 | 3 |
| 54 | 1 |

Die relativen Häufigkeiten erhält man, indem man die absoluten Häufigkeiten durch den Stichprobenumfang teilt:

$$h_{rel} = \frac{h_{abs}}{n}$$

In der Fachliteratur findet man oft das Symbol $h$ für absolute Häufigkeiten und $f$ für relative Häufigkeiten. Zusätzlich sollte man in der Formel auch kennzeichnen, dass sie pro Klasse gilt. Wenn man die Klassen durchnummeriert und für diese Klassennummern den Index $i$ verwendet, schaut die obige Formel nach dieser Konvention dann so aus:

$$f_i = \frac{h_i}{n}$$

Man erhält dann diese relativen Häufigkeiten:

| Gewicht | $h_{abs} = h_i$ | $h_{rel} = f_i$ |
|---|---|---|
| 46 | 2 | 0,08 |
| 47 | 4 | 0,16 |
| 48 | 6 | 0,24 |
| 49 | 4 | 0,16 |
| 50 | 2 | 0,08 |
| 51 | 3 | 0,12 |
| 52 | 3 | 0,12 |
| 54 | 1 | 0,04 |

Die Summe der absoluten Häufigkeiten ergibt immer den Umfang der Stichprobe.

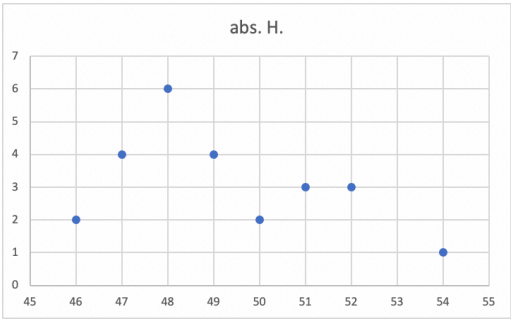
Die Summe der relativen Häufigkeiten ergibt immer 1.

$$\sum_{i=1}^{k} h_i = n \qquad \sum_{i=1}^{k} f_i = 1$$

Dabei soll $k$ für die Anzahl der Gewichtsklassen stehen.

### 3.3.1 Graphische Darstellung der Häufigkeiten

#### Punktdiagramm

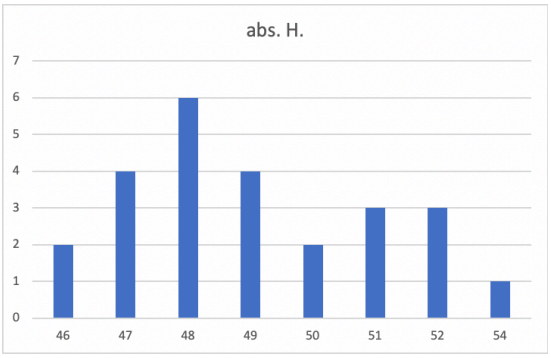


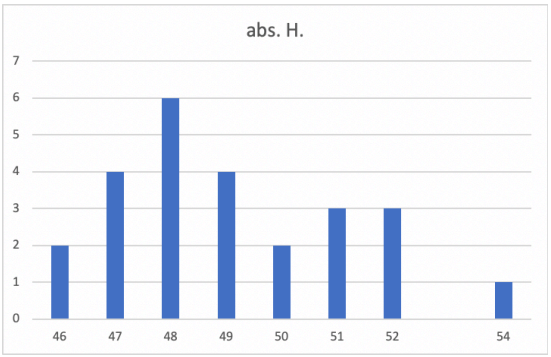

Eine Möglichkeit, Häufigkeiten darzustellen, ist das Punktdiagramm. Um das nebenstehende Diagramm zu erzeugen, markiert man in Excel alle Werte in den Spalten *Gewicht* und $h_{abs}$ inklusive der Überschriften und wählt dann im Menüband *Einfügen* die Diagrammklasse *Punkt (X Y)*, und dort die Diagrammart *Punkt (XY)*.[19]

[19] Diese Bezeichnungen können sich ändern wenn Microsoft beschließt, eine neue Version von Excel herauszubringen.

**Stabdiagramm, Säulendiagramm**

Weil kleine Punkte auf einer großen Diagrammfläche sehr oft verloren wirken, wird zur Darstellung von Häufigkeiten sehr oft ein Stab- oder Säulendiagramm verwendet. Die Erzeugung erfolgt wie zuvor, jedoch wird als Diagrammklasse *Säule* und davon *Gruppierte Säulen* gewählt.




**Achtung Falle!**

Excel betrachtet bei Säulendiagrammen die Werte aus der ersten Spalte nicht als X-Werte, sondern als Namen von Klassen! Excel stellt alle Säulen im selben Abstand voneinander dar. Daher ist der Abstand zwischen den Säulen 52 und 54 gleich groß wie zwischen allen anderen Säulen.

Will man das verhindern, muss man in die Tabelle zusätzliche Leerzeilen einfügen (hier mit gelben Hintergrund dargestellt).

Bei Punktdiagrammen werden die Werte aus der ersten Spalte als X-Werte interpretiert, so dass damit automatisch die richtigen Abstände entstehen.

| Gewicht | abs. H. |
|---|---|
| 46 | 2 |
| 47 | 4 |
| 48 | 6 |
| 49 | 4 |
| 50 | 2 |
| 51 | 3 |
| 52 | 3 |
| | |
| 54 | 1 |

Weil die Möglichkeiten zur graphischen Darstellung so vielfältig sind, gibt es ein separates Dokument, in dem diese Möglichkeiten behandelt werden. Zur Darstellung von Häufigkeiten gibt es einen speziellen Subtyp der Säulendiagramme: das Histogramm. Es wird in dem erwähnten separaten Dokument ausführlich beschrieben.

## 3.4 Lagemaße

Lagemaße sind Kennzahlen univariater Werteverteilung, die Auskunft darüber geben, was ein typischer Wert innerhalb der Wertemenge ist. Lagemaße zeigen an, wo besonders viele Werte liegen (z.B. Modus) oder wo die Mitte der Werteverteilung ist (z.B. Median, Mittelwert), sie charakterisieren daher die Position der Werteverteilung.

### 3.4.1 **Modus, Modalwert**

Der Modus oder Modalwert ist ein Lageparameter, der bei allen Messskalen, also sogar auch bei nominalskalierten Merkmalen verwendet werden kann. Der Modus gibt an, welcher Wert innerhalb einer Stichprobe oder Grundgesamtheit am häufigsten vorkommt.

Der Modus hat aber einige Nachteile:

1. Bei nicht-nominalverteilten Werten (wenn es eine innere Ordnung gibt): Aus der Lage des Modus kann nicht darauf geschlossen werden, wo die Mitte der Verteilung liegt (wie auch immer man den Begriff »Mitte« definieren will). Der Modus kann nämlich auch der größte oder kleinste Wert einer Verteilung mit innerer Ordnung sein.
2. Der Modus ist nicht immer eindeutig. Es kann vorkommen, dass es zwei oder mehr Werte gibt, die beide die größte Häufigkeit innerhalb der Stichprobe oder Grundgesamtheit aufweisen. In bestimmten Verteilungen (Gleichverteilung) ist sogar jeder Wert der Verteilung ein Modus.
3. Je stärker eine Häufigkeitsverteilung einer Gleichverteilung ähnelt, desto beliebigere Werte kann der Modus annehmen.

Beispiele für Modalwerte:

- Das Wort, das in gedruckten deutschsprachigen Texten am häufigsten vorkommt, ist das Wort »der«.
- Das Wort, das in gedruckten englischen Texten im Korpus des Project Gutenberg am häufigsten vorkommt, ist »the«. (Das Wort »you« liegt in diesem Korpus auf Platz 17. »The« wird in gedruckten Texten 9,3-mal so oft verwendet wie »you«.)
- Das Wort, das 2006 in englischsprachigen Fernsehserien am häufigsten ausgesprochen wurde, ist »you«. (Es wurde 1,6-mal so häufig gesagt wie »the«. Auch »I« und »to« wurden häufiger verwendet als »the«.)
- Der häufigste Vorname der Welt ist محمد (Mohammed) bzw. محمّد (Muhammed)
- Der häufigste Nachname der Welt ist 王 (»Wang« bzw. »Wong«). Mehr als 7% aller Chinesen heißen so, der Name ist auch in anderen asiatischen Ländern sehr häufig.
- Der häufigste Nachname in Österreich ist Gruber.
- Der häufigste Nachname in Deutschland ist Müller.
- Die PKW-Modellreihe, von der weltweit die meisten Exemplare verkauft wurden, heißt Toyota Corolla.
- Die am häufigsten verkaufte PKW-Marke heißt VW Käfer.
- Die Lottozahl[20], die zwischen 1.1.2011 und 31.12.2020 am häufigsten gezogen wurde, ist die Zahl 32 (sie wurde in diesen 10 Jahren 182-mal gezogen), obwohl sie in keinem einzigen Jahr dieser Dekade die häufigste Zahl war.

[20] Österreichisches Lotto »6 aus 45«, ohne »Lotto Plus«. Gewinnzahlen inklusive Zusatzzahl. (7 Zahlen pro Ziehung)

### 3.4.2 **Median[21], Zentralwert**

Den Median gibt es nur bei Werten, die über eine innere Ordnung verfügen, also bei ordinal-, intervall- und verhältnisskalierten Werten. Man erhält den Median, indem man die Werte der Stichprobe oder Grundgesamtheit entsprechend dieser inneren Ordnung sortiert, und dann jenen Wert wählt, der genau in der Mitte dieser sortierten Liste steht.

Wenn man die Elemente in der sortierten Liste mit $1$ beginnend aufsteigend durchnummeriert, und $n$ die Anzahl der Elemente (und zugleich die Nummer des letzten Elements) ist, dann ist der Median das Element mit dem Index $\frac{n+1}{2}$.

Diese Definition ist aber nur anwendbar, wenn $n$ (die Anzahl der Elemente in der Wertemenge) ungerade ist. Bei einer geraden Anzahl ergibt $\frac{n+1}{2}$ keine natürliche Zahl, die Indizes sind aber natürliche Zahlen. Man braucht also eine Definition des Medians, der bei einer ungeraden Elementanzahl genau das oben geschilderte Ergebnis liefert, aber auch dann funktioniert, wenn man eine gerade Anzahl an Elementen hat.

Bei verhältnis- und intervallskalierten Werten lässt sich das einfach realisieren: Da berechnet man den Median einer Stichprobe mit einer geraden Elementezahl, indem man das arithmetische Mittel jener beiden Werte berechnet, die der Mitte der Liste am nächsten stehen. Wenn $n$ die Anzahl der Elemente ist, addiert man also die Elemente mit den Indizes $\frac{n}{2}$ sowie $\frac{n}{2}+1$ und teilt diese Summe durch 2.

**Zusammenfassung:**

- Man sortiert zuerst alle Elemente aus der Stichprobe entsprechend der inneren Ordnung.
- Man nummeriert alle Elemente in der sortierten Liste durch. Der Index des ersten Elements ist $1$, der des letzen Elements $n$. $n$ ist zugleich auch die Anzahl der Elemente in der Stichprobe.
- Wenn $n$ ungerade ist, ist der Median das Element mit dem Index $\frac{n+1}{2}$. Der Median ist in diesem Fall also immer eines der Elemente in der Liste.
- Wenn $n$ gerade ist, addiert man die Werte der Elemente mit den Indizes $\frac{n}{2}$ und $\frac{n}{2}+1$ und teilt diese Summe durch 2. Dieser Quotient ist der Median. Es kann durchaus sein, dass das Ergebnis dieser Berechnung ein Wert ist, der selbst gar nicht in der Stichprobe vorkommt.

**Beispiel 1:**

Eine Schülerin hat in ihrem Zeugnis diese Noten stehen:

{»Genügend«, »Sehr gut«, »Sehr gut«, »Gut«, »Sehr gut«, »Genügend«, »Gut«}

Nach der inneren Ordnung sortieren und durchnummerieren:

[21] Über englisch *median* aus lateinisch *medianus* = in der Mitte liegend

{1:»Sehr gut«, 2:»Sehr gut«, 3:»Sehr gut«, **4:»Gut«**, 5:»Gut«, 6:»Genügend«, 7:»Genügend«}

Der letzte Index und somit die Anzahl der Elemente in der Stichprobe ist 7, das ist eine ungerade Zahl. Der Median ist daher das Element mit dem Index $\frac{7+1}{2} = \frac{8}{2} = 4$. An dieser Stelle steht der Wert »Gut«, der Median dieser Stichprobe ist daher: »Gut«.

**Beispiel 2:**

Die Zentralanstalt für Meteorologie und Geodynamik hat am 13. Jänner 2021 um 12:00 Uhr Mittag diese Temperaturen gemessen[22] (linke Liste):

| Station | Temperatur in °C |
|---|---|
| Wien Hohe Warte | 4,0 |
| Wien Mariabrunn | 3,4 |
| Eisenstadt | 4,2 |
| St. Pölten | 1,1 |
| Linz | 1,9 |
| Hörsching Flughafen | 1,1 |
| Salzburg Freisaal | 1,5 |
| Salzburg Flugh. | 0,7 |
| Aigen | -0,1 |
| Innsbruck Flugh. | 1,4 |
| Bregenz | 2,4 |
| Graz Universität | 4,3 |
| Graz Flughafen | 4,2 |
| Lienz | -5,7 |
| Klagenfurt Flugh. | -3,1 |
| Sonnblick | -18,6 |
| Feuerkogel | -9,8 |
| Villacher Alpe | -10,8 |

| Index | Temperatur in °C |
|---|---|
| 1 | -18,6 |
| 2 | -10,8 |
| 3 | -9,8 |
| 4 | -5,7 |
| 5 | -3,1 |
| 6 | -0,1 |
| 7 | 0,7 |
| 8 | 1,1 |
| **9** | **1,1** |
| **10** | **1,4** |
| 11 | 1,5 |
| 12 | 1,9 |
| 13 | 2,4 |
| 14 | 3,4 |
| 15 | 4,0 |
| 16 | 4,2 |
| 17 | 4,2 |
| 18 | 4,3 |

Um den Median der Temperaturen zu ermitteln, werden die Temperaturen nach ihrem Wert sortiert und zugleich durchnummeriert. Das ergibt eine Liste, wie sie rechts dargestellt ist.

Die letzte Nummer in der sortierten Liste ist 18, die Stichprobe enthält also 18 Elemente, das ist eine gerade Zahl. Der Median ist daher der Mittelwert der Elemente mit den Indizes $\frac{18}{2} = 9$ und $\frac{18}{2} + 1 = 9 + 1 = 10$. Die Werte der Elemente 9 und 10 sind $1{,}1$ und $1{,}4$. Die Hälfte der Summe dieser beiden Werte ist der Median der Stichprobe:

$$\tilde{x} = \frac{1{,}1 + 1{,}4}{2} = \frac{2{,}5}{2} = 1{,}25$$

[22] Von dieser Seite: https://www.zamg.ac.at/cms/de/wetter/wetterwerte-analysen

Der Median dieser Stichprobe ist 1,25°C. Dieser Wert repräsentiert zwar die Mitte der Stichprobe, wurde aber an keiner der 18 Messstationen gemessen.

**Beispiel 3, ein Problemfall:**

Sechs Angehörige des österreichischen Bundesheeres haben ihren Dienstgrad angegeben. Die Antworten, bereits nach Dienstgrad sortiert und durchnummeriert, sind:

{1:»Gefreiter«, 2:»Korporal«, 3:»Korporal«, 4:»Wachtmeister«, 5:»Stabswachtmeister«, 6:»Vizeleutnant«}

Die Anzahl der Elemente ist gerade, aber über dieser Menge, die aus ordinalskalierten Elementen besteht, ist die Summenbildung nicht definiert. Ein arithmetisches Mittel kann nicht berechnet werden, daher kann der Median nicht wie oben beschrieben ermittelt werden.

**Die Lösung:**

Um dieses Problem zu lösen, muss man den Begriff »Median« genauer definieren, nämlich so:

> Ein Median ist ein Wert, der zugleich die beiden folgenden Bedingungen erfüllt:
> 1. Der Median ist größer als oder gleich groß wie mindestens die Hälfte aller Elemente.
> 2. Der Median ist kleiner als oder gleich groß wie mindestens die Hälfte aller Elemente.

Wie man sich leicht überzeugen kann, ist diese Definition kompatibel mit den bisherigen Berechnungsvorschriften. Im Beispiel 3 führt diese Definition aber dazu, dass mehrere Werte als Median in Frage kommen:

- Korporal
  Der Wert »Korporal« ist größer oder gleich groß wie die Elemente mit den Indizes 1, 2 und 3, das ist genau die Hälfte aller Elemente. ✅
  Der Wert »Korporal« ist kleiner oder gleich groß wie die Elemente mit den Indizes 2, 3, 4, 5 und 6, das sind sogar mehr die Hälfte aller Elemente. ✅
  Der Wert »Korporal« ist daher ein Median.
- Wachtmeister
  Der Wert »Wachtmeister« ist größer oder gleich groß wie die Elemente mit den Indizes 1, 2, 3 und 4, das ist mehr als die Hälfte aller Elemente. ✅
  Der Wert »Wachtmeister« ist kleiner oder gleich groß wie die Elemente mit den Indizes 4, 5 und 6, das sind genau die Hälfte aller Elemente. ✅
  Der Wert »Wachtmeister« ist daher ein Median.

Es gibt aber noch einen Wert, der als Median qualifiziert ist:

- Zugsführer
  Der Wert »Zugsführer« ist größer oder gleich groß wie die Elemente mit den Indizes 1, 2, 3, das ist genau die Hälfte aller Elemente. ✅
  Der Wert »Zugsführer« ist kleiner oder gleich groß wie die Elemente mit den Indizes 4, 5 und 6, das sind genau die Hälfte aller Elemente. ✅
  Der Wert »Zugsführer« ist daher ein Median.

Tatsächlich ist es im vorliegenden Fall so, dass es drei verschiedene Mediane gibt, nämlich alle Elemente aus der Menge {Korporal, Zugsführer, Wachtmeister}

Dieselbe Argumentation kann man auch beim Beispiel 2 mit den Temperaturen führen. Alle Werte im beidseitig abgeschlossenen Intervall $[1{,}1;\ 1{,}4]$ sind Mediane.

Sehr oft möchte man aber nicht mehrere (im Fall der Temperaturen sogar unendlich viele) Mediane haben, sondern nur einen. Und genau das leistet der Mittelwert dieses Intervalls, der jedoch nur bei metrischen Skalen definiert ist.

### 3.4.3 **Durchschnitt, Mittelwert, arithmetisches Mittel**

Das arithmetische Mittel ist nur für metrische Skalen definiert (intervall- und verhältnisskalierte Werte). Bei diesen Werten haben Intervalle (also Differenzen zwischen Werten) eine Bedeutung und auch einen durch eine Zahl ausdrückbaren Wert. Man kann sich das arithmetische Mittel wie folgt vorstellen:

Man nehme einen gewichtslosen und ausreichend langen horizontal ausgerichteten Stab, und hänge an diesen Stab so viele gleich große Gewichte, wie es Elemente in der Stichprobe oder Grundgesamtheit gibt. Die Abstände zwischen den Aufhängepunkten der Gewichte entsprechen genau den Intervallen zwischen den Werten. Der Mittelwert ist dann jene Stelle des Stabes, an dem man den Stab samt seinen Gewichten so aufhängen kann, dass er weiterhin genau waagrecht bleibt.


Diese Definition beschreibt den physikalischen Schwerpunkt einer eindimensionalen Gewichtsverteilung, und diese Definition stimmt exakt mit der Definition des arithmetischen Mittels überein. Die Formel zur Berechnung dieses Mittelwertes kann daher aus den

Trägheitsmomenten der Gewichte abgeleitet werden. Darauf wird hier aber verzichtet, denn die sich daraus ergebende Formel ist deutlich einfacher als ihre Herleitung. Sie lautet wie folgt:

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

Diese Formel bedeutet folgendes:

Es gibt $n$ Elemente in der Stichprobe, die man der Reihe nach durchnummerieren kann (vorheriges Sortieren ist nicht notwendig, es würde am Ergebnis nichts verändern). Als allgemeines Symbol für die Indexnummer wird das Symbol $i$ verwendt. $i$ nimmt der Reihe nach die Werte $1, 2, 3$ usw. an, bis maximal $n$. (Der Startwert $1$ steht unter dem Summensymbol $\Sigma$[23], der Endwert $n$ darüber.) Die aufzusummierenden Werte sind $x_1, x_2$ usw., allgemein $x_i$. Diese Summe muss dann noch durch die Anzahl der Elemente (also durch $n$) dividiert werden (was dasselbe ist wie eine Multiplikation mit $\frac{1}{n}$). Das Ergebnis ist der Mittelwert $\bar{x}$ (sprich: »X quer«). Der Querstrich über dem Symbol zeigt an, dass es sich um einen Mittelwert handelt.

**Besondere Eigenschaften des arithmetischen Mittels**

- **Minimum der Abweichungsquadrate**
  Man wähle irgendeinen Wert $\mu$.
  Jedes $x_i$ hat davon den Abstand $\mu - x_i$.
  Diese Abstände kann man quadrieren und von diesen Quadraten die Summe bilden:

  $$S = \sum_{i=1}^{n} (\mu - x_i)^2$$

  Diese Summe hängt bei einer vorgegebenem Stichprobe (also bei vorgegebenen Werten für alle $x_i$) nur von $\mu$ ab. $S$ ist also eine Funktion von $\mu$. Man kann nun jenen Wert für $\mu$ suchen, bei dem $S(\mu)$ ein Minimum annimmt, wo also $S'^{(\mu)} = 0$. Es stellt sich heraus, dass das genau dann der Fall ist, wenn man für $\mu$ den arithmetischen Mittelwert $\bar{x}$ einsetzt.

- **Äquivarianz (lineare Transformierbarkeit)**
  Das arithmetische Mittel ist äquivariant gegenüber Multiplikation und Addition. Das bedeutet folgendes:
  Wenn man alle Werte der Stichprobe mit demselben Faktor $f$ multipliziert, und davon den Mittelwert berechnet, erhält man dasselbe, wie wenn man den Mittelwert der ursprünglichen Werte mit diesem Faktor multipliziert. Dasselbe gilt für das Addieren eines konstanten Summanden $s$ (»AM« steht für »Arithmetisches Mittel«):

  $$\mathrm{AM}(s + f \cdot x_i) = s + f \cdot \mathrm{AM}(x_i)$$

[23] Der griechische Großbuchstabe Σ (»Sigma«) entspricht dem lateinischen Buchstaben S (selber Lautwert) und wird in der Mathematik als Symbol für Summen verwendet.

### 3.4.4 gewichtetes arithmetisches Mittel

In manchen Fällen möchte man einen Mittelwert berechnen, bei dem einige der Werte aus der Stichprobe mehr Einfluss auf den Mittelwert haben sollen als andere. Das ist beispielsweise der Fall, wenn man den Notendurchschnitt eines/einer Studierenden berechnen soll, und dabei erreichen will, dass Fächer mit vielen ECTS den Notendurchschnitt dominieren. Dann verwendet man diese ECTS-Punkte als Gewichte. Zu jeder einzelnen Note $x_i$ gehört also immer auch ein Gewicht $w_i$, wie in diesem Beispiel:

| Nr ($i$) | Fach | Note ($x_i$) | ECTS ($w_i$) |
|---|---|---|---|
| 1 | Geschichte der Zauberei | 4 | 1 |
| 2 | Besenfliegen | 1 | 3 |
| 3 | Kräuterkunde | 2 | 3 |
| 4 | Verteidigung gegen die dunklen Künste | 1 | 5 |
| 5 | Zaubertränke | 3 | 5 |

Die Formel sieht so aus:

$$AM_w = \bar{x}_w = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i}$$

Man multipliziert also jede Note mit ihrem Gewicht und summiert diese Produkte auf:

Zähler = $4 \cdot 1 + 1 \cdot 3 + 2 \cdot 3 + 1 \cdot 5 + 3 \cdot 5 = 4 + 3 + 6 + 5 + 15 = 33$

Im Nenner des Bruchs zählt man nur die Gewichte zusammen:

Nenner = $1 + 3 + 3 + 5 + 5 = 17$

Der Mittelwert ist nun der Quotient:

$$\bar{x} = \frac{33}{17} = 1{,}941176471\cdots$$

Ein Sonderfall tritt ein, wenn alle Gewichte gleich sind und den Wert 1 haben ($w_i = 1$ für alle $i$):

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

Den Faktor 1 im Zähler kann man weglassen (1 ist das neutrale Element der Multiplikation). Im Nenner wird die aus $n$ Summanden bestehende Summe $1 + 1 + 1 + \cdots$ gebildet, das ergibt genau den Wert $n$. Man erhält dann, wie zu erwarten war,

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

also genau die Formel für das ungewichtete arithmetische Mittel.

### 3.4.5 **Quadratisches Mittel**

Anstatt den Mittelwert der unveränderten Werte der Stichprobe zu berechnen, kann man auch zuerst diese Werte quadrieren, von den Quadraten den Mittelwert berechnen, und aus diesem Ergebnis dann die Quadratwurzel ziehen:

$$QM = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Diesen Mittelwert benötigt man häufig in technischen Anwendungen, z.B. bei der Berechnung des Leistungsumsatzes von Wechselstrom an einem ohmschen Widerstand. Man nennt diesen Mittelwert auch »zweites Moment«[24].

Das Quadrieren eines Wertes entspricht der Multiplikation des Wertes mit einer Zahl. Multiplikationen mit Zahlen sind aber nur bei verhältnisskalierten Werten sinnvoll.

Die Quadratwurzel ergibt nur dann reelle Werte, wenn der Wert unter der Wurzel 0 oder positiv ist. Daher macht es keinen Sinn, ein quadratisches Mittel von Werten zu berechnen, die negativ sein können.

Aus den beiden letzten Absätzen folgt, dass das quadratische Mittel nur bei nichtnegativen verhältnisskalierten Werten sinnvoll ist.

Vom quadratischen Mittel gibt es auch eine gewichtete Variante:

$$QM_w = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot w_i}{\sum_{i=1}^{n} w_i}}$$

### 3.4.6 **Kubisches Mittel**

Beim kubischen Mittel, das man auch »drittes absolutes Moment« nennt, werden die Stichprobenwerte zur dritten Potenz erhoben, am Ende wird die dritte Wurzel gezogen:

$$KM = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{n} x_i^3}$$

Es gibt nur wenige Anwendungsbereiche dafür, einer davon ist die Abschätzung der Lebensdauer von Maschinenteilen. In der Statistik kann damit auch der Begriff der Schiefe definiert werden.

[24] Das arithmetische Mittel heißt auch »erstes Moment«. Diese Bezeichnungen haben mit physikalischen Momenten zu tun. Das erste Moment ist das Trägheitsmoment einer rotierenden Masse, das zweite Moment ist das Drehmoment.

### 3.4.7 **geometrisches Mittel**

Beim arithmetischen Mittel wurden die Werte der Stichprobe addiert, die erhaltene Summe wurde mit dem Kehrwert der Anzahl multipliziert. Wenn man in dieser Vorschrift jede Rechenart durch ihr Gegenstück auf der jeweils nächsthöheren Stufe ersetzt, wird aus der Addition eine Multiplikation und aus der Multiplikation eine Potenzierung. Die Formel sieht dann so aus:

$$GM = \left( \prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Aus dem Summenzeichen $\Sigma$ wurde das Produktzeichen $\Pi$[25] und aus der Multiplikation mit $\frac{1}{n}$ wurde eine Potenz. Mit dem Kehrwert der Zahl $n$ zu potenzieren ist aber dasselbe wie das Ziehen der $n$-ten Wurzel, daher sieht auch diese Formel sehr häufig, die aber denselben Rechengang beschreibt:

$$GM = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Das geometrische Mittel ist nur für nichtnegative verhältnisskalierte Werte definiert, aber nur sinnvoll, wenn keiner der Werte 0 ist. (In diesem Fall ist nämlich auch der Mittelwert gleich 0, egal, wie groß alle anderen Werte sind.)

Das geometrische Mittel verwendet man überall dort, wo Verläufe durch geometrische Reihen beschrieben werden, also überall wo exponentielles Wachstum herrscht. Das ist beispielsweise bei der Verzinsung von Vermögen der Fall. Das geometrische Mittel wird auch in der Chemie verwendet, um Konzentrationen von Substanzen zu berechnen, die in Lösungsmittel gelöst sind. Auch bei der Definition des goldenen Schnitts kommt es zum Einsatz.

### 3.4.8 **Harmonisches Mittel**

Jemand geht mit einer Geschwindigkeit von 3 km/h vom Tal bis zum Gipfel eines Hügels, dreht dort um und geht dieselbe Strecke mit 6 km/h zurück. Wenn dieselbe Person dieselbe Wanderung mit konstanter Geschwindigkeit machen würde (bergauf und bergab gleich schnell, und dafür insgesamt gleich viel Zeit brauchen würde, mit welcher Geschwindigkeit müsste diese Person gehen?

Von allen Werten der Stichprobe oder Grundgesamtheit werden die Kehrwerte genommen, diese Kehrwerte werden addiert, die Summe durch die Anzahl der Elemente geteilt, und davon

[25] Der griechische Großbuchstabe Π (»Pi«) entspricht dem lateinischen Buchstaben P (selber Lautwert) und wird in der Mathematik als Symbol für Produkte verwendet.

wird dann noch einmal der Kehrwert genommen. In der kompakten Schreibweise der Mathematik sieht das so aus:

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

Mit den Zahlen aus dem Beispiel:

$$HM = \frac{2}{\frac{1}{3}+\frac{1}{6}} = \frac{2}{\frac{2}{6}+\frac{1}{6}} = \frac{2}{\frac{2+1}{6}} = \frac{2}{\frac{3}{6}} = \frac{2\cdot 6}{3} = \frac{12}{3} = 4$$

Diese Person müsste mit 4 km/h gehen.

### 3.4.9 **Hölder-Mittel, Potenzmittel**

Wenn man die bisherigen Formeln ein wenig umformt, so dass sie einander ähnlicher sehen, erkennt man ein Muster:

| Name | Originalformel | umgewandelte Formel |
|---|---|---|
| Harmonisches Mittel | $HM = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$ | $HM = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{-1}\right)^{\frac{1}{-1}}$ |
| Geometrisches Mittel | $GM = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$ | $GM = \lim_{p\to 0}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$ |
| Arithmetisches Mittel | $AM = \frac{1}{n}\sum_{i=1}^{n} x_i$ | $AM = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^1\right)^{\frac{1}{1}}$ |
| Quadratisches Mittel | $QM = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$ | $QM = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$ |
| Kubisches Mittel | $KM = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{n} x_i^3}$ | $KM = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^3\right)^{\frac{1}{3}}$ |

Im Fall des geometrischen Mittels erfordert diese Umformung etwas mehr mathematisches Knowhow als in den anderen Fällen (Anwenden der Regel von de l'Hospital und der Logarithmusgesetze; Bildung eines Grenzwertes), liefert aber dennoch eine Formel, die sich (bis auf die Grenzwertbildung) ins Muster der anderen Formeln fügt. Diese allgemeine Formel sieht so aus:

$$PM_p = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$$

Diese Formel ist nur auf verhältnisskalierte Werte anwendbar. Lediglich wenn $p = 1$ kann man die Formel auch auf intervallskalierte Werte anwenden, denn dann handelt es sich ja um das arithmetische Mittel. Dieser Fall ist auch der einzige, in dem der Mittelwert äquivariant gegenüber der Addition ist. Das Hölder-Mittel ist aber bei allen Werten für p gegenüber der Multiplikation äquivariant:

$$PM_p(\mathrm{f} \cdot x_i) = \mathrm{f} \cdot PM_p(x_i)$$

Daher haben auch alle hier behandelten Mittelwerte diese Eigenschaft.

Das Hölder-Mittel in seiner allgemeinen Form ist nur für positive Werte in der Stichprobe definiert. Unter dieser Voraussetzung trifft auch folgender Satz zu:

Der Wert des Hölder-Mittels ist umso größer, je größer der Parameter $p$ ist und umso kleiner, je kleiner $p$ ist. Insbesondere gelten auch, wenn der Parameter gegen plus oder minus unendlich strebt, diese beiden Grenzwerte:

| Name | Originalformel | umgewandelte Formel |
|---|---|---|
| Minimum | $Min = \min(x_i)$ | $Min = \lim_{p \to -\infty} \left( \frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}}$ |
| Maximum | $Max = \max(x_i)$ | $Max = \lim_{p \to +\infty} \left( \frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}}$ |

Daraus folgt (mit den hier verwendeten Abkürzungen), unter der Voraussetzung, dass alle $x_i$ positiv sind und nicht alle $x_i$ genau denselben Wert haben:

$$Min < HM < GM < AM < QM < KM < Max$$

### 3.4.10 **andere Mittelwerte**

Es darf nicht unerwähnt bleiben, dass es noch eine Vielzahl weiter Mittelwerte gibt, auf die hier nicht weiter eingegangen wird. Zu erwähnen ist vor allem das logarithmische Mittel, das immer zwischen dem arithmetischen und dem geometrischen Mittel liegt.

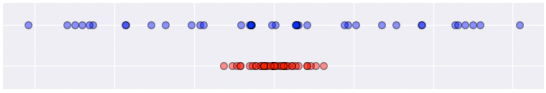
Das Stolarsky-Mittel ist eine Verallgemeinerung des logarithmischen Mittels, das in der Lage ist, alle hier wiedergegebene Mittelwerte nachzubilden und als Spezialfall auch den *Identric Mean* genannten Mittelwert beinhaltet.

Das f-Mittel ist eine Verallgemeinerung des Hölder-Mittels (das Hölder-Mittel verwendet die Exponentialfunktion $x_i^p$ als charakteristische Funktion, beim f-Mittel kann stattdessen jede beliebige streng monoton wachsende Funktion sein).

Auch das Lehmer-Mittel ist ein allgemeiner Mittelwert, der HM, GM und AM berechnen kann, und zusätzlich noch das bereits in der griechischen Antike bekannte kontraharmonische Mittel. Diese Mittelwerte haben aber allesamt kaum praktische Anwendungsbereiche.

## 3.5 Streumaße

Während Lagemaße etwas darüber aussagen, bei welchen Werten man mit größter Wahrscheinlichkeit Werte aus der Stichprobe findet, geben Streumaße darüber Auskunft, wie weit entfernt von dieser Stelle mit Werten zu rechnen ist.



Die beiden abgebildeten Punktwolken haben denselben Mittelwert, aber die Streuung der blauen Wolke ist größer als die der roten Wolke. Das Streuungsmaß der blauen Wolke ist größer als das der roten.

Wie bei den Lagemaßen gibt es auch bei den Streumaßen eine große Auswahl an verschiedenen Werten. Es macht aber nicht Sinn, jedes beliebige Streumaß mit jedem beliebigen Lagemaß zu kombinieren.

Streumaße machen nur dort einen Sinn, wo es sinnvoll ist davon zu reden, dass ein Wert kleiner oder größer als ein anderer ist. Daher gibt es bei nominalskalierten Werten keine Streumaße. Bei nominalskalierten Werten (Namen, Farben usw.) kann man nur den Umfang der Stichprobe oder Grundgesamtheit angeben, und man kann angeben, wie viele unterschiedliche Ausprägungen die Werte innerhalb dieser Menge annehmen. Diese Größen sind aber keine Streumaße im eigentlichen Sinn.

### 3.5.1 Minimum und Maximum

Den kleinsten und den größten Wert einer Werteverteilung kann man bereits bei ordinalskalierten Werten angeben.

**Beispiel:**

Bei einem Offizierstreffen sind Soldaten mit diesen Dienstgraden anwesend:

{Oberleutnant, Major, Oberstleutnant, Oberleutnant, Hauptmann, Major}

- Das Minimum dieser Menge (der niedrigste Dienstgrad) ist »Oberleutnant«.
- Das Maximum dieser Menge (der höchste Dienstgrad) ist »Oberstleutnant«.

Diese beiden Werte (Oberleutnant und Oberstleutnant) geben gemeinsam darüber Auskunft, wie weit gestreut die Dienstgrade bei diesem Treffen sind.

Darüber hinaus kann man bei ordinalskalierten Werten keine vernünftige Aussage über die Streuung machen. Insbesondere ist es nicht möglich, hier eine Spannweite anzugeben.

Im Kapitel 3.4.9 wurden das Minimum und das Maximum mit verschiedenen Lagemaßen verglichen. Es ist daher auch möglich, Minimum und Maximum als Lagemaße zu interpretieren.

### 3.5.2 **Spannweite, Streubreite**

Die Spannweite oder Streubreite ist die Differenz zwischen dem Maximum und dem Minimum. Die Voraussetzung zur Ermittlung dieser Größe ist, dass es sinnvoll sein muss, Differenzen von Werten zu berechnen. Das ist nur bei metrischen Skalen möglich (intervall- und verhältnisskaliert), daher gibt es auch bei diesen Skalen die Möglichkeit, eine Spannweite zu berechnen.

### 3.5.3 **Vorbereitung auf Streumaße, die zum Median gehören**

Um diese Streumaße beschreiben zu können, muss man zuerst den Begriff des Medians verallgemeinern. Die Definition des Medians lautet wie folgt (mit leicht verändertem Wording gegenüber der Definition aus 3.4.2):

> Der Median ist ein Wert, der zugleich die beiden folgenden Bedingungen erfüllt:
> 1. Der Median ist größer als oder gleich groß wie mindestens $\frac{1}{2}$ aller Elemente.
> 2. Der Median ist kleiner als oder gleich groß wie mindestens $1 - \frac{1}{2}$ aller Elemente.

In der ursprünglichen Definition stand nach dem Wort »mindestens« in beiden Zeilen der Begriff »die Hälfte«. Das wurde in der ersten Zeile durch $\frac{1}{2}$ und in der zweiten Zeile durch $1 - \frac{1}{2}$ ersetzt.

Man kann nun anstelle von $\frac{1}{2}$ beispielsweise den Wert $\frac{1}{4}$ einsetzen. Das ergibt die Definition des unteren Quartils[26]:

**unteres Quartil = 25%-Quantil[27]**

> Das untere Quartil ist ein Wert, der zugleich die beiden folgenden Bedingungen erfüllt:
> 1. Das untere Quartil ist größer als oder gleich groß wie mindestens $\frac{1}{4}$ aller Elemente.
> 2. Das untere Quartil ist kleiner als oder gleich groß wie mindestens $\frac{3}{4}$ aller Elemente.

(Hier wurde der Ausdruck $1 - \frac{1}{4}$ bereits durch $\frac{3}{4}$ ersetzt)

Entsprechend gibt es auch ein ...

**oberes Quartil = 75%-Quantil**

> Das obere Quartil ist ein Wert, der zugleich die beiden folgenden Bedingungen erfüllt:
> 1. Das obere Quartil ist größer als oder gleich groß wie mindestens $\frac{3}{4}$ aller Elemente.
> 2. Das obere Quartil ist kleiner als oder gleich groß wie mindestens $\frac{1}{4}$ aller Elemente.

[26] Von lateinisch *quartus* = der vierte, hier im Sinn von »der vierte Teil«, also »ein Viertel«
[27] Lateinisch *quantus?* = wie viel? wie groß? bzw. *quantum* = Dosis, Menge

Die folgende Tabelle veranschaulicht diese Begriffe am Beispiel von Uhrzeiten (intervallskalierte Größe)

| | |
|---|---|
| 8:32 | ← Minimum = 0% Quantil |
| 8:45 | |
| 8:46 | ← unteres Quartil = 25% Quantil |
| 8:46 | |
| 8:50 | ← Median = 50%-Quantil |
| 14:23 | |
| 14:26 | ← oberes Quartil = 75% Quantil |
| 23:40 | |
| 23:40 | ← Maximum = 100% Quantil |

Diese Tabelle lässt auch schon ahnen, wie Quantile definiert sind:

**p%-Quantil**

Das p%-Quantil ist ein Wert, der zugleich die beiden folgenden Bedingungen erfüllt:
1. Das p%-Quantil ist größer als oder gleich groß wie mindestens p% aller Elemente.
2. Das p%-Quantil ist kleiner als oder gleich groß wie mindestens 100-p% aller Elemente.

Zu beachten ist, dass nach dieser Definition auch jeder beliebige Wert, der kleiner als das Minimum ist, ein 0%-Quantil ist, und jeder Wert, der größer als das Maximum ist, ist ein 100%-Quantil. Daher verzichtet man auf die Verwendung von 0%- und 100%-Quantilen zugunsten der ohnehin bekannteren Begriffe *Minimum* und *Maximum*.

### 3.5.4 Interquartilsabstand

Da wir nun mit den nötigen Fachbegriffen vertraut sind, können wir endlich das erste Streumaß definieren, das mit dem Median zusammen verwendet wird. Der Interquartilsabstand ist die Differenz der beiden Quartile:

IQR[28] = oberes Quartil – unteres Quartil

Innerhalb des IQR liegen 50% aller Werte, je 25% der Werte liegen darunter und darüber.

Bei der Berechnung dieses Wertes wird eine Differenz gebildet, dieses Maß kann daher nicht für ordinalskalierte Werte verwendet werden. Das gilt auch für die allgemeinere Version dieses Streumaßes, den Quantilsabstand.

[28] von englisch **I**nter **Q**uartil **R**ange

### 3.5.5 Quantilsabstand

Der Interquartilsabstand ist identisch mit dem 25%-Quantilsabstand:

$$QA_{25\%} = 75\%\text{-Quantil} - 25\%\text{-Quantil}$$

Oder allgemein:

$$QA_p = (100\text{-}p)\%\text{-Quantil} - p\%\text{-Quantil}$$

Innerhalb des $QA_p$ liegen (100-2p)% aller Werte, je p% der Werte liegen darunter und darüber.

### 3.5.6 Mittlere absolute Abweichung vom Median

Wenn man den Median $\tilde{x}$ schon ermittelt hat, kann man von jedem Wert der Stichprobe die Differenz zum Median berechnen, davon den Absolutbetrag nehmen, und von diesen Abständen das arithmetische Mittel berechnen. Das Ergebnis ist die »mittlere absolute Abweichung vom Median« oder englisch »median deviation«:

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|$$

### 3.5.7 Median der absoluten Abweichung vom Median

Eine weitere Möglichkeit zu einem Streumaß zu kommen, das mit dem Median verträglich ist, funktioniert wie die soeben mittlere absolute Abweichung, jedoch wird nun von den Abständen nicht das arithmetische Mittel gebildet, sondern es wird der Median ermittelt.

Auf den ersten Blick wirkt diese Methode, als wäre sie auch auf ordinalskalierte Werte anwendbar, das ist aber nicht der Fall, denn auch bei dieser Methode muss man zuerst einmal die Abstände zwischen Median und den einzelnen Abständen ausrechen, und das erfordert eine Subtraktion, die für ordinalskalierte Werte nicht definiert ist. (Was soll herauskommen, wenn man von einem Vizeleutnant einen Wachtmeister subtrahiert?)

### 3.5.8 Mittlere absolute Abweichung vom Mittelwert

Wenn man die in 3.5.6 beschriebe Formel verwendet und den darin vorkommenden Median $\tilde{x}$ (»X Tilde« oder »X Schlange«) durch das arithmetische Mittel $\bar{x}$ (»X quer«) ersetzt, erhält man die Mittlere absolute Abweichung vom Mittelwert (englisch »mean absolute deviation«):

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

Der einzige praktische Nutzen der mittleren absoluten Abweichung vom Mittelwert ist die didaktische Hinleitung zur Varianz.

### 3.5.9 Varianz = Mittlere quadratische Abweichung vom Mittelwert

Anstelle die Absolutbeträge der Abstände aufzusummieren, kann man auch die Quadrate dieser Abstände zu einer Summe vereinen, die dann wieder durch die Anzahl dividiert werden muss.

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

Diese Größe ist – im Gegensatz zu den zuletzt beschriebenen Streumaßen – ein außerordentlich wichtiges Maß in der deskriptiven Statistik. Die Varianz ist nämlich sehr eng mit dem arithmetischen Mittel verknüpft.

In 3.4.3 wurden bereits besondere Eigenschaften des arithmetischen Mittels beschrieben, und es wurde dort darauf hingewiesen, dass das arithmetischen Mittel die Summe der Abweichungsquadrate minimiert. Diese Summe ist aber nichts anderes als die Varianz, multipliziert mit der Anzahl der Elemente in der Stichprobe. Diese Anzahl ist für eine gegebene Stichprobe aber konstant, daher minimiert das arithmetischen Mittel nicht nur die Summe der Abweichungsquadrate, sondern auch die Varianz.

Die Varianz hat aber einen entscheidenden Nachteil:

Wenn die Werte in der Stichprobe z.B. Temperaturen sind, die in Grad Celsius gemessen wurden, dann hat die Varianz einen Wert, der in Quadratgraden Celsius angegeben wird. Ganz allgemein hat die Varianz immer eine physikalische Dimension, die das Quadrat der Dimension der Werte in der Stichprobe ist. Damit ist die Varianz nicht direkt vergleichbar mit den Werten in der Stichprobe.

**Beispiel:**

20 erwachsene Personen haben sich auf eine Waage gestellt, das Gewicht jeder Person wurde auf ganze kg gerundet. Das sind die Messergebnisse:

{70, 59, 91, 66, 94, 101, 72, 63, 73, 73, 77, 62, 85, 77, 82, 67, 87, 82, 59, 60}

Diese 20 Personen wiegen zusammen 1500 kg, im Schnitt bringt jede Person 75 kg auf die Waage. Das ist das arithmetische Mittel:

$$\bar{x} = 75\ \text{kg}$$

Zieht man von den Werten den Mittelwert ab, erhält man diese Differenzen:

{-5, -16, 16, -9, 19, 26, -3, -12, -2, -2, 2, -13, 10, 2, 7, -8, 12, 7, -16, -15}

Quadrieren jedes Wertes ergibt

{25, 256, 256, 81, 361, 676, 9, 144, 4, 4, 4, 169, 100, 4, 49, 64, 144, 49, 256, 225}

Die Summe all dieser Quadrate ist 2880. Teilt man diese Zahl durch die Anzahl der Werte (also durch 20) erhält man 144, das ist die Varianz:

$$VAR = \sigma^2 = 144\ \text{kg}^2$$

Die Varianz der Gewichte der 20 Personen beträgt 144 Quadratkilogramm.

Die Wichtigkeit der Varianz in der deskriptiven Statistik ergibt sich aber nicht aus dem Wert als Endergebnis einer Berechnung, denn mit der Information, dass das Gewicht einer bestimmten Personengruppe um 144 Quadratkilogramm variiert, fängt niemand wirklich etwas an. Die Varianz ist aber ein wichtiges Zwischenergebnis, das in vielerlei andere Berechnungen einfließt.

Aus der Varianz lässt sich aber mit einem naheliegenden Rechenschritt eine durchaus praktisch relevante Größe berechnen, nämlich die Standardabweichung.

### 3.5.10 **Standardabweichung**

Die Standardabweichung ist die Wurzel aus der Varianz. Ihr Symbol ist ein kleines Sigma ($\sigma$):

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

Damit ist auch klargestellt, warum für die Varianz neben dem Symbol VAR auch das Symbol $\sigma^2$ in Verwendung ist: Die Varianz ist das Quadrat der Standardabweichung. Tatsächlich wird aber immer zuerst die Varianz berechnet, und daraus erst die Standardabweichung.

**Fortsetzung des Beispiels:**

Damit fehlt nur noch ein einfacher Rechenschritt um aus dem unhandlichen Wert der Varianz ein praktisch sehr viel nützlicheres Streumaß zu machen. Man muss nur noch die Wurzel ziehen:

$$\sigma = \sqrt{VAR} = \sqrt{144\ \text{kg}^2} = 12\ \text{kg}$$

Die Standardabweichung in diesen Daten beträgt 12 kg.

Diese Aussage lässt sich auch viel besser verstehen. Sie bedeutet nämlich, dass jeder Mensch der vermessenen Gruppe ein Gewicht hat, das durchschnittlich 12 kg vom Durchschnittsgewicht (welches 75 kg beträgt) entfernt ist.

Man nennt die Standardabweichung daher auch »mittlere Abweichung vom Mittelwert«.

### 3.5.11 **Varianz und Standardabweichung einer Stichprobe**

Die Formeln für Varianz und Standardabweichung, wie sie in 3.5.9 und 3.5.10 beschrieben wurden, gehen von der Annahme aus, dass der Mittelwert $\bar{x}$ den tatsächlichen Mittelwert der Grundgesamtheit wiedergibt. Aber in Wahrheit weichen die Mittelwerte verschiedener Stichproben, die aus derselben Grundgesamtheit gezogen werden, voneinander ab.

**Beispiel:**

Hier werden wieder die Gewichte derselben 20 Personen verwendet, die schon im Kapitel über die Varianz untersucht wurden. Nehmen wir an, diese 20 Personen wären die Grundgesamtheit. Daraus ziehen wir 5 Stichproben, die jeweils aus den Werten von 4 Personen bestehen. Wir berechnen für jede Stichprobe das arithmetische Mittel und mit der in 3.5.10 beschriebenen Formel die Standardabweichung (gerundet auf 3 Nachkommastellen):

| Stichprobe | $\bar{x}$ | $\sigma$ |
|---|---|---|
| $\{70, 59, 91, 66\}$ | 71,50 | 11,927 |
| $\{94, 101, 72, 63\}$ | 82,50 | 15,532 |
| $\{73, 73, 77, 62\}$ | 71,25 | 5,585 |
| $\{85, 77, 82, 67\}$ | 77,75 | 6,833 |
| $\{87, 82, 59, 60\}$ | 72,00 | 12,629 |
| (Grundgesamtheit) | 75,00 | 12,000 |

Wie man gut sehen kann, weichen die Mittelwerte der Stichproben vom Mittelwert der Grundgesamtheit ab. Der Mittelwert der Stichprobenmittelwerte stimmt zwar mit dem Mittelwert der Grundgesamtheit überein (jedoch nur, weil die Stichproben die Grundgesamtheit vollständig partitionieren), aber die einzelnen Mittelwerte weichen davon ab. Diesen »mittleren Fehler der Mittelwerte« kann man wieder mit der Standardabweichung berechnen. Er beträgt in diesem Beispiel ca. 4,453 kg.

Bei der Berechnung der Standardabweichungen wurden aber die »falschen« Mittelwerte verwendet (nicht 75,00 sondern 71,50, 82,50 usw.). Wenn man in die Formel für die Varianz bzw. Standardabweichung anstelle der Mittelwerte der jeweiligen Stichprobe den Mittelwert der Grundgesamtheit (also 75,00) einsetzt, erhält man diese Werte:

| Stichprobe | $\sigma_{\text{Stichprobe}}$ | $\sigma_{\text{korrigiert}}$ |
|---|---|---|
| $\{70, 59, 91, 66\}$ | 11,927 | 12,430 |
| $\{94, 101, 72, 63\}$ | 15,532 | 17,248 |
| $\{73, 73, 77, 62\}$ | 5,585 | 6,727 |
| $\{85, 77, 82, 67\}$ | 6,833 | 7,365 |
| $\{87, 82, 59, 60\}$ | 12,629 | 12,981 |
| (Grundgesamtheit) | 12,000 | 12,000 |

Alle Standardabweichungen werden dadurch größer. Während der Mittelwert der Standardabweichungen vorher $10{,}501$ betrug, liegt er bei den korrigierten Werten bei $11{,}350$ und somit deutlich näher am wahren Wert $(12{,}000)$.

Meistens kennt man aber nicht alle Werte der Grundgesamtheit (das ist ja der Grund, warum man mit Stichproben arbeitet). Man muss also bei der Formel, die nur Werte aus der Stichprobe verarbeitet, eine Korrektur anbringen, um eine Varianz bzw. Standardabweichung zu bekommen, die näher am vermuteten (aber unbekannten) Wert der Grundgesamtheit liegt.

Es lässt sich mathematisch gut begründen, dass der beste Weg darin besteht, die Summe der quadrierten Abstände vom Mittelwert der Stichprobe nicht durch die Anzahl der Elemente in der Stichprobe (also durch $n$) zu dividieren, sondern durch $(\mathrm{n}-1)$. Andere Methoden, die eine noch genauere Annäherung ermöglichen, setzen voraus, dass man weiß, aus wie vielen Elementen die Grundgesamtheit besteht. Diese Zahl ist aber sehr oft unbekannt, und führt nur zu geringfügig anderen Werten.

| | alte Formel (gilt weiterhin für die Grundgesamtheit) | neue Formel (gilt für eine Stichprobe) |
|---|---|---|
| Varianz | $VAR = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ | $VAR = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ |
| Standardabweichung | $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ | $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ |

Verwendet man die neue Formel, erhält man für die Stichproben aus dem Beispiel diese Werte:

| Stichprobe | $\sigma_{\text{Stichprobe}}$ | $\sigma_{\text{korrigiert}}$ | $\sigma_{\text{neue\_Formel}}$ |
|---|---|---|---|
| $\{70, 59, 91, 66\}$ | 11,927 | 12,430 | 13,772 |
| $\{94, 101, 72, 63\}$ | 15,532 | 17,248 | 17,935 |
| $\{73, 73, 77, 62\}$ | 5,585 | 6,727 | 6,448 |
| $\{85, 77, 82, 67\}$ | 6,833 | 7,365 | 7,890 |
| $\{87, 82, 59, 60\}$ | 12,629 | 12,981 | 14,583 |
| (Grundgesamtheit) | 12,000 | 12,000 | |

Der Mittelwert der neuen Standardabweichungen liegt bei $12{,}157$ und liegt damit sogar näher der Wert, den wir mit dem korrigierten (meist aber unbekannten) Mittelwert erhalten haben.

### 3.5.12 **Variationskoeffizient**

Bei Werten, die verhältnisskaliert sind (deren Skala also einen »natürlichen« Nullpunkt besitzt), möchte man manchmal auch Aussagen wie die folgende machen:

- Das Körpergewicht der vermessenen Personen schwankt um 10% um den Mittelwert.
- Das Alter der Studierenden in Klasse A ist sehr einheitlich, denn das Alter eines/einer durchschnittlichen Studierenden weicht nur um ca. 7% vom Mittelwert ab. In Klasse B beträgt diese Abweichung 34%, dort herrscht also eine weitaus größere Altersdiversität.

(Überlegen Sie sich, dass diese Aussagen bei intervallskalierten Werten wie Uhrzeiten, Celsius-Graden usw. nicht besonders sinnvoll sind.)

Dieser Prozentwert, der hier verwendet wird, ist der Quotient aus der Standardabweichung und dem Mittelwert. Nachdem Standardabweichung und Mittelwert immer dieselbe physikalische Dimension haben, ist der Quotient eine dimensionslose Zahl und kann daher in Prozenten ausgedrückt werden. Man nennt diese Zahl »Variationskoeffizient«.

$$v = \frac{\sigma}{\bar{x}}$$

### 3.5.13 **Absolute Durchschnittsdifferenz = Gini[29]-Durchschnittsdifferenz**

Ein weiteres Streumaß kommt ganz ohne vorangehende Mittelwertbildung aus. Es ist ganz einfach der Durchschnitt aller absoluten Differenzen aller Paare, die sich aus der Stichprobe bilden lassen. Dabei wird auch jedes Element mit sich selbst gepaart (was jedes Mal die Differenz 0 ergibt). Der englische Name dafür ist »Gini mean difference« und das dafür verwendete Symbol ist »GMD«:

$$GMD = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|$$

### 3.5.14 **Relative absolute Durchschnittsdifferenz; Gini-Koeffizient**

Wenn die Werte intervallskaliert sind, macht es Sinn, die absolute Durchschnittsdifferenz zum arithmetischen Mittel in Beziehung zu setzen. Man erhält die »relative mean difference« RMD:

$$RMD = \frac{GMD}{\bar{x}}$$

Der Wert, der bei dieser Berechnung herauskommt, ist genau das doppelte des Gini-Koeffizienten, der bei den Konzentrationsmaßen behandelt wird (Kapitel 3.7.2 auf Seite 46), und dort aber auf eine ganz andere Weise hergeleitet wird. Es lässt sich aber zeigen, dass beide Verfahren dasselbe Ergebnis bringen.

## 3.6 **Formmaße**

Neben den Lagemaßen und den Streumaßen gibt es noch weitere Kenngrößen, mit denen man univariate Verteilungen beschreiben kann. Sie beschreiben die Form der Verteilung und werden daher unter dem Begriff »Formmaße« zusammengefasst. In der Praxis werden sie seltener eingesetzt als die zuvor behandelten Größen, dürfen aber auch nicht ganz außer Acht gelassen werden.

[29] nach dem italienischen Statistiker Corrado Gini (1884-1965)

### 3.6.1 gewöhnliche und zentrale Momente

Werfen wir zunächst einen Blick auf zwei schon bekannte Formeln, wobei die Formel für das arithmetische Mittel ein wenig angepasst wird, damit sich die beiden Formeln ein wenig ähnlicher sehen.

| | Originalformel | angepasste Formel |
|---|---|---|
| arithmetisches Mittel | $\frac{1}{n}\sum_{i=1}^{n} x_i$ | $\frac{1}{n}\sum_{i=1}^{n}(x_i - 0)^1$ |
| Varianz | $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ | $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ |

Die allgemeine Form dieser Formeln sieht so aus:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^k$$

Der Wert $a$ ist ein Bezugswert, der im ersten Schritt der Berechnung von allen Werten der Stichprobe abgezogen wird. (Von welchem Bezugspunkt aus betrachten wir die Daten?) Wenn $a$ gleich 0 ist, spricht man von einem »gewöhnlichen« Moment. Wenn $a$ ein Mittelwert ist (insbesondere, wenn es sich um das arithmetische Mittel handelt), nennt man das Moment »zentral«. Der Parameter $k$ ist eine positive natürliche Zahl, er gibt an, um das wievielte Moment es sich handelt.

Für die beiden hier betrachteten Maße gibt es also neue Bezeichnungen in einem neuen Schema:

- Das arithmetische Mittel ist das erste gewöhnliche Moment einer univariaten Wertemenge.
- Die Varianz ist das zweite zentrale Moment einer univariaten Wertemenge.

Man kann sich leicht überlegen, dass das erste zentrale Moment immer 0 ist, und dass, vom ersten Moment abgesehen, die zentralen Momente eine höhere Aussagekraft haben, weil sie unabhängig von der Lage der Verteilung sind.

Man kann sich nun fragen, was man aus dem dritten oder vierten Moment herauslesen kann, und es stellt sich heraus, dass man daraus Kennzahlen ableiten kann, die etwas über die Form einer Verteilung aussagen. Allerdings ist es dabei sinnvoll, die höheren Momente noch in eine sinnvolle Beziehung zur Streuung der Werte zu setzen. Das wird durch eine Verknüpfung mit der Standardabweichung erreicht.

Der Vollständigkeit halber seien hier noch schnell die Formeln für die höheren zentralen Momente angeführt, sie werden in den folgenden Kapiteln benötigt:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \qquad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

### 3.6.2 **Momente einer Stichprobe**

Die in 3.6.1 genannten Formeln sind die Formeln für die Grundgesamtheit. Wenn man stattdessen eine Stichprobe untersucht, muss der Vorfaktor angepasst werden.

| | Grundgesamtheit | Stichprobe |
|---|---|---|
| arithmetisches Mittel<br>1. gewöhnliches Moment | $\frac{1}{n}\sum_{i=1}^{n}(x_i-0)^1$ | $\frac{1}{n}\sum_{i=1}^{n}(x_i-0)^1$ |
| Varianz<br>2. zentrales Moment | $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ | $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ |
| 3. zentrales Moment | $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$ | $\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i-\bar{x})^3$ |
| 4. zentrales Moment | $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4$ | $\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(x_i-\bar{x})^4$ |

### 3.6.3 **Schiefe (Momentschiefe, Momentenkoeffizient)**

Die Schiefe einer Werteverteilung ist das dritte zentrale Moment dieser Verteilung, geteilt durch die dritte Potenz der Standardabweichung:

$$\tilde{\mu}_3 = \frac{m_3}{\sigma^3}$$

Die Formel für $m_3$ steht weiter oben auf dieser Seite, die Formel für $\sigma$ steht im Kapitel über die Standardabweichung (3.5.10). Setzt man beide Formeln in die Formel für die Schiefe ein, erhält man diesen Ausdruck:

$$\tilde{\mu}_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$$

Das kann man zwar zum folgenden Ausdruck umformen, macht die Formel aber nicht wesentlich einfacher

$$\tilde{\mu}_3 = \sqrt{n}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}$$

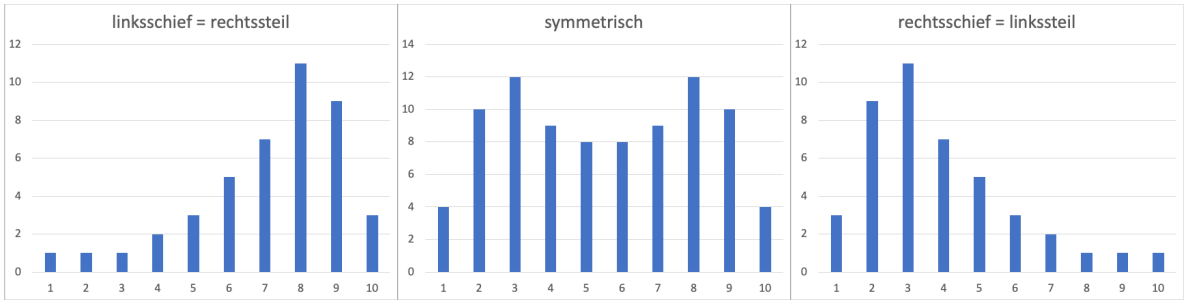
Das ist der Ausdruck für die Schiefe der Grundgesamtheit. Will man die Schiefe einer Stichprobe berechnen, müsste man für $m_3$ die Formel aus 3.6.2 einsetzen, was zu einem noch komplizierteren Ausdruck führt, den sich garantiert niemand mehr auswendig merkt. Wenn Sie diese Formeln irgendwann mal benötigen sollten, können Sie hier, in diesem Dokument nachschauen.

Es ist sicherlich einfacher, sich einfach nur $\tilde{\mu}_3 = \frac{m_3}{\sigma^3}$ zu merken, zumal man meist ohnehin nur am Vorzeichen des Ergebnisses interessiert ist.

Wenn $\tilde{\mu}_3$ negativ ist, bedeutet dass, dass die Verteilung nach »rechts« geneigt ist. »Rechts bezieht sich dabei auf die Ausrichtung der Zahlengeraden, auf der die Werte von links nach rechts immer größer werden.

Ein positiver Wert für $\tilde{\mu}_3$ bedeutet das Gegenteil.

Der Absolutbetrag von $\tilde{\mu}_3$ gibt Auskunft darüber, wie stark die Schiefe ausgeprägt ist.




$\tilde{\mu}_3 < 0$ $\tilde{\mu}_3 = 0$ $\tilde{\mu}_3 > 0$

Drei verschiedene Verteilungen mit unterschiedlicher Schiefe

Je nach dem Vorzeichen der Schiefe gibt es unterschiedliche Bezeichnungen:

| | | | | |
|---|---|---|---|---|
| $\tilde{\mu}_3 < 0$ | negative Schiefe | rechtssteil | linksschief | schief |
| $\tilde{\mu}_3 = 0$ | keine Schiefe | | | symmetrisch |
| $\tilde{\mu}_3 > 0$ | positive Schiefe | linkssteil | rechtsschief | schief |

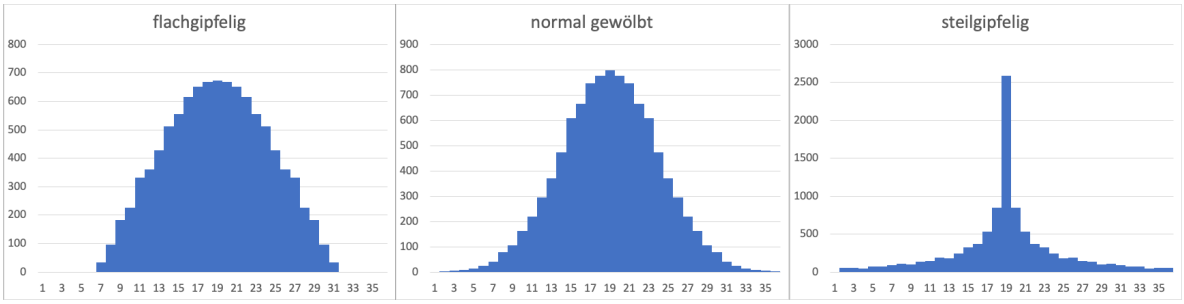
### 3.6.4 **Wölbung (Exzess, Kurtosis)**

Die Wölbung leitet sich vom vierten zentralen Moment ab:

$$\tilde{\mu}_4 = \frac{m_4}{\sigma^4}$$

Dieser Wert ist immer größer als 0 und hat bei einer Wölbung, die jener der Gaußschen Glockenkurve entspricht, genau den Wert 3. Aus diesem Grund hat man einen neuen Wert eingeführt, den man Kurtosis nennt:

$$K = \tilde{\mu}_4 - 3 \;=\; \frac{m_4}{\sigma^4} - 3$$




$\tilde{\mu}_4 < 3$ $\tilde{\mu}_4 = 3$ $\tilde{\mu}_4 > 3$

Drei verschiedene Verteilungen mit unterschiedlicher Wölbung

| | |
|---|---|
| $K < 0 \ (\tilde{\mu}_4 < 3)$ | flachgipfelig |
| $K = 0 \ (\tilde{\mu}_4 = 3)$ | normal gewölbt |
| $K > 0 \ (\tilde{\mu}_4 > 3)$ | steilgipfelig |

Daneben gibt es noch den Begriff »Exzess«, der in manchem Lehrbüchern gleichbedeutend mit »Kurtosis« verwendet wird, in anderen wird der Excess mit $\tilde{\mu}_4$ gleichgesetzt. Dasselbe gilt für den Begriff »Wölbung«.

### 3.6.5 **andere Formmaße**

Neben den Formmaßen, die auf die Momente zurückzuführen sind, gibt es noch andere Maße, die mit dem Median verwandt sind, aber noch seltener zum Einsatz kommen als die eben behandelten Formmaße. Ein Beispiel dafür ist die ...

**Quartilsschiefe:**

Dabei betrachtet man die Abstände des unteren und oberen Quartils vom Median. Wenn beide Abstände gleich sind, ist die Verteilung symmetrisch, ansonsten schief. (Die Begriffe »linksschief«, »rechssteil« usw. sind sinngemäß gleich definiert wie in 3.6.3)

Die Formel für den Zahlenwert der Quartilsschiefe lautet:

$$QS = \frac{(Q_{0,75} - Q_{0,50}) - (Q_{0,50} - Q_{0,25})}{Q_{0,75} - Q_{0,25}}$$

Dabei ist $Q_{0,25}$ das 25%-Quantil (also das untere Quartil) usw.

## 3.7 Konzentrationsmaße

**Beispiel:**

Zwei Unternehmen (»Unishirt« und »Tee Shirt«) produzieren und verkaufen T-Shirts in verschiedenen Farben. Die folgende Tabelle zeigt, wie groß der Umsatz pro Farbe ist (Umsatz in tausend Euro pro Jahr):

| Unternehmen | Unishirt | Tee Shirt |
|---|---|---|
| Farbe | Umsatz | Umsatz |
| weiß | 75 | 110 |
| gelb | 10 | 7 |
| rot | 14 | 11 |
| blau | 12 | 5 |
| grün | 8 | 9 |
| pink | 11 | 12 |
| grau | 70 | 16 |
| schwarz | 40 | 70 |

Beide Unternehmen bieten dieselben 8 Farben an und machen gleich viel Umsatz (240 Millionen Euro pro Jahr). Welches der beiden Unternehmen hat sein Portfolio gleichmäßiger gestreut? Welches setzt stärker auf Spezialisierung?

### 3.7.1 Lorentz-Kurve

Um uns der Antwort zu nähern, wollen wir für beide Unternehmen die Lorentz-Kurve zeichnen. Dazu führen wir für jedes Unternehmen separat die folgenden Schritte durch:

1. Die Umsätze aufsteigend sortieren.
2. Feststellen der Anteile der Merkmalsträger am Produktportfolio. (Wie viele verschiedene Produkte gibt es? Jedes Produkt hat denselben Anteil, nämlich 1 geteilt durch die Anzahl der Produktarten)
3. Feststellen der Anteile der Umsätze am Gesamtumsatz
4. Beide Anteile kumulieren

Das ergibt diese beiden Tabellen:

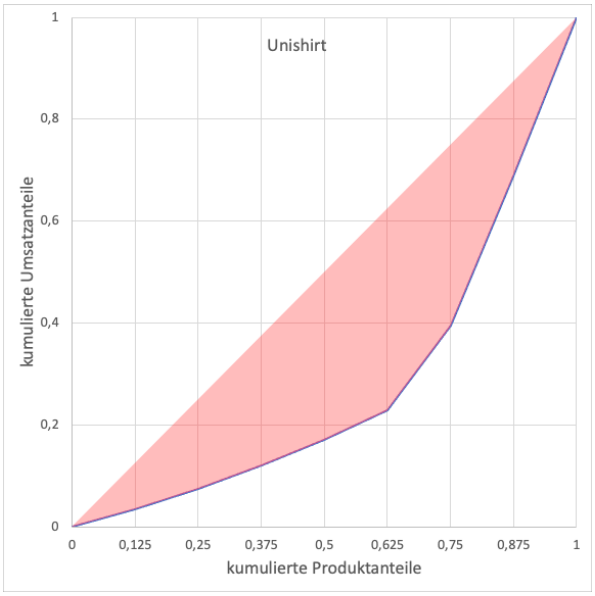
Unishirt

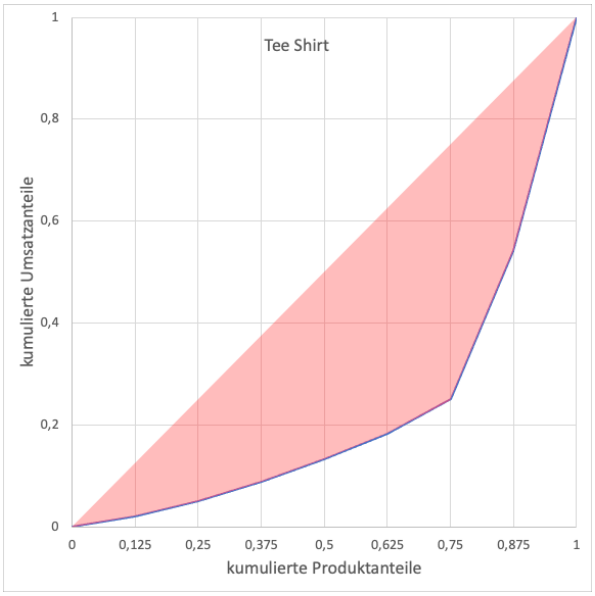
| $i$ | Farbe | Umsatz | Produktanteil | Umsatzanteil | kumulierte Produktanteile $(x_i)$ | kumulierte Umsatzanteile $(y_i)$ |
|---|---|---|---|---|---|---|
| 1 | grün | 8 | 0,125 | 0,033333333 | 0,125 | 0,033333333 |
| 2 | gelb | 10 | 0,125 | 0,041666667 | 0,25 | 0,075 |
| 3 | pink | 11 | 0,125 | 0,045833333 | 0,375 | 0,120833333 |
| 4 | blau | 12 | 0,125 | 0,05 | 0,5 | 0,170833333 |
| 5 | rot | 14 | 0,125 | 0,058333333 | 0,625 | 0,229166667 |
| 6 | schwarz | 40 | 0,125 | 0,166666667 | 0,75 | 0,395833333 |
| 7 | grau | 70 | 0,125 | 0,291666667 | 0,875 | 0,6875 |
| 8 | weiß | 75 | 0,125 | 0,3125 | 1 | 1 |

Tee Shirt

| $i$ | Farbe | Umsatz | Produktanteil | Umsatzanteil | kumulierte Produktanteile $(x_i)$ | kumulierte Umsatzanteile $(y_i)$ |
|---|---|---|---|---|---|---|
| 1 | blau | 5 | 0,125 | 0,020833333 | 0,125 | 0,020833333 |
| 2 | gelb | 7 | 0,125 | 0,029166667 | 0,25 | 0,05 |
| 3 | grün | 9 | 0,125 | 0,0375 | 0,375 | 0,0875 |
| 4 | rot | 11 | 0,125 | 0,045833333 | 0,5 | 0,133333333 |
| 5 | pink | 12 | 0,125 | 0,05 | 0,625 | 0,183333333 |
| 6 | grau | 16 | 0,125 | 0,066666667 | 0,75 | 0,25 |
| 7 | schwarz | 70 | 0,125 | 0,291666667 | 0,875 | 0,541666667 |
| 8 | weiß | 110 | 0,125 | 0,458333333 | 1 | 1 |

Die Daten aus den beiden jeweils letzten Spalten trägt man dann in ein Diagramm ein:



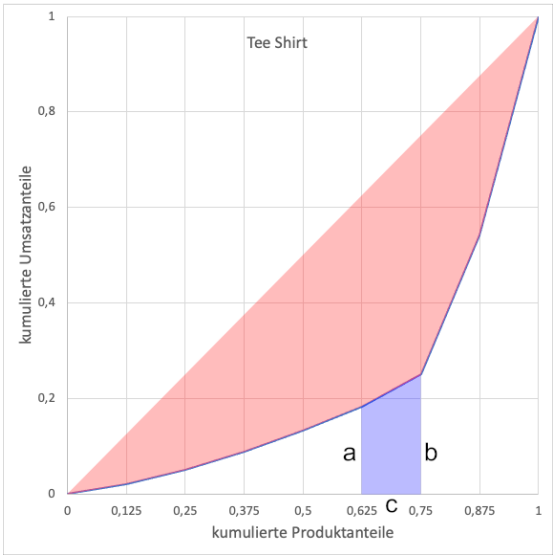



Der blaue Streckenzug am unteren Rand der roten Fläche beginnt bei (0, 0) und geht dann durch alle Punkte aus der Tabelle. Die Konstruktionsvorschrift bewirkt, dass der letzte Punkt immer (1, 1) sein muss. Dieser Streckenzug heißt »Lorentzkurve«.

### 3.7.2 **Gini-Koeffizient**

Die rote Fläche ist unten durch den blauen Streckenzug begrenzt und oben durch die Gerade, die durch die beiden Punkte (0, 0) und (1, 1) geht. Das Doppelte des Flächeninhaltes dieser roten Fläche ist der Gini-Koeffizient.

Um den Inhalt dieser Fläche zu berechnen, berechnet man zunächst den Inhalt der Fläche unterhalb der Lorentz-Kurve, also jene Fläche, die in den beiden Abbildungen jeweils durch die Lorentz-Kurve, den unteren und den rechten Rand des Diagramms begrenzt ist. Diese Fläche besteht aus einem Dreieck (ganz links, beim Punkt (0, 0)) und sonst auch lauter Trapezen. Auch das linke Dreieck kann man als Trapez auffassen, dessen eine Seite die Länge 0 hat. Daher kann man alle Teile der gesuchten Fläche mit derselben Formel berechnen.




Die Fläche des blauen Trapezes ist

$$\frac{(a+b)}{2} \cdot c$$

Dabei hat $c$ für alle Trapeze denselben Wert, das ist nämlich der Produktanteil jedes einzelnen Produkts und das ist wiederum $\frac{1}{n}$ wobei $n$ die Anzahl der Produkte ist. Das blaue Trapez hat also die Fläche

$$\frac{(a+b)}{2} \cdot \frac{1}{n} = \frac{(a+b)}{2n}$$

$a$ und $b$ sind benachbarte Umsatzanteile, also $y_{i-1}$ und $y_i$. Die Summe aller Trapezflächen ist daher:

$$A = \frac{1}{2n} \sum_{i=1}^{n} (y_{i-1} + y_i)$$

Es folgen nun mehrere Umformungsschritte

Die Summe aufteilen:

$$A = \frac{1}{2n} \left( \sum_{i=1}^{n} y_{i-1} + \sum_{i=1}^{n} y_i \right)$$

Index der ersten Summe ändern (vorher 1 bis n, nachher 0 bis n-1, dadurch ändert sich beim Summanden der Index von $i-1$ zu $i$.)

$$A = \frac{1}{2n}\left(\sum_{i=0}^{n-1} y_i + \sum_{i=1}^{n} y_i\right)$$

Ersten Summanden der linken Summe aus der Summe herausnehmen. (Achte darauf, von wo bis wo die Indizes nun laufen.)

$$A = \frac{1}{2n}\left(y_0 + \sum_{i=1}^{n-1} y_i + \sum_{i=1}^{n} y_i\right)$$

$y_0$ ist aber die linke Seite des ganz linken Trapezes, das in Wahrheit ein Dreieck ist, also $0$. und das kann man ganz einfach weglassen:

$$A = \frac{1}{2n}\left(\sum_{i=1}^{n-1} y_i + \sum_{i=1}^{n} y_i\right)$$

Damit die linke Summe auch bis $n$ laufen kann, muss man den Summanden $y_n$ in die Summe hineinnehmen (Durch Änderung des Bereichs bis zu dem $i$ läuft), muss ihn aber zugleich außerhalb der Summe einmal abziehen

$$A = \frac{1}{2n}\left(-y_n + \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i\right)$$

$y_n$ ist aber die rechte Seite des ganz rechten Trapezes, das ist immer genau $1$. Also: $y_n = 1$.

$$A = \frac{1}{2n}\left(-1 + \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i\right)$$

Die beiden Summen zusammenfassen

$$A = \frac{1}{2n}\left(-1 + 2\sum_{i=1}^{n} y_i\right)$$

Klammer auflösen

$$A = -\frac{1}{2n} + \frac{1}{n}\sum_{i=1}^{n} y_i$$

Die Fläche unter der Geraden, die durch (0, 0) und (1, 1) geht, ist genau $\frac{1}{2}$. Die rote Fläche ist daher:

$$F = \frac{1}{2} - A$$

$$F = \frac{1}{2} + \frac{1}{2n} - \frac{1}{n}\sum_{i=1}^{n} y_i$$

Der Gini-Koeffizient ist per Definition genau das Doppelte dieser Fläche:

$$C = 2F$$

**Formel für Gini-Koeffizient**

$$C = 1 + \frac{1}{n} - \frac{2}{n}\sum_{i=1}^{n} y_i$$

Die Werte $y_i$ sind die kumulierten Umsatzanteile.

Eine andere Formel, um den Gini-Koeffizienten zu berechnen, wurde bereits in 3.5.14 vorgestellt. Der Gini-Koeffizient ist genau die Hälfte der relativen absoluten Durchschnittsdifferenz.

**Zurück zum Beispiel**

Bei der Firma Unishirt beträgt die Summe der kumulierten Umsatzanteile $\sum_{i=1}^{n} y_i = 2{,}7125$. Bei der Firma Tee Shirt kommt dieser Wert heraus: $\sum_{i=1}^{n} y_i = 2{,}26666667$. $n$ hat in beiden Firmen den Wert $8$ (das ist die Anzahl der Produkte).

Setzt man diese Werte in die Formel ein, erhält man:

Unishirt: $C = 0{,}446875$

Tee Shirt: $C = 0{,}5583333$

**Was sagt nun der Gini-Index aus?**

Je kleiner der Gini-Index ist, desto gleichmäßiger sind die Umsätze auf die einzelnen Produkte verteilt. Ein großer Gini-Index bedeutet, dass es wenige Produkte mit großem Umsatz und viele Umsätze mit wenig Umsatz gibt.

**Zwei extreme Verteilungen:**

**völlige Gleichverteilung**

Nehmen wir an, die Firma EqualShirt schafft es, von jeder der 8 Farben genau gleich viele T-Shirts zu verkaufen. Nehmen wir an, in jeder Farbe 1 Million pro Jahr. Die Sortierung spielt in diesem Spezialfall hier keine Rolle (in allen anderen Fällen schon). Das führt dazu, dass die kumulierten Umsatzanteile 1/8, 2/8, 3/8 usw. betragen, also genau mit den kumulierten Produktanteilen übereinstimmen. Die Lorentzkurve (blauer Streckenzug in den letzten Grafiken) wird dadurch zu einer Geraden, die durch (0, 0) und (1, 1) geht, und die Fläche der roten Kurve schrumpft dadurch zu 0.

Der Gini-Index einer Gleichverteilung ist daher genau 0.

Überlegen Sie, warum der Gini-Index nicht kleiner als 0 sein kann. (Denken Sie über die Rolle der Sortierung nach.)

**Das andere Extrem**

Die Firma BlackShirt produziert zwar T-Shirts in allen 8 Farben, hat bisher aber nur schwarze T-Shirts verkauft. Beim Sortieren rutscht schwarz dadurch ans Ende der Liste. Alle Einträge in der Spalte »kumulierte Umsatzanteile« sind 0, bis auf den letzten, der 1 ist. Die Lorenz-Kurve verläuft also zuerst genau entlang der X-Achse und biegt erst beim allerletzten »Trapez« nach oben ab. Die rote Fläche nimmt daher fast das ganze untere Dreieck ein, seine Fläche liegt also ganz knapp unterhalb von $\frac{1}{2}$ und der Gini-Index (der ja das Doppelte dieser Fläche ist), hat dann einen Wert knapp unter 1.

Die Annäherung an den Wert 1 gelingt umso besser, je mehr Produkte das Unternehmen im Portfolio hat.

Daraus folgt, dass der Gini-Index sich zwar dem Wert 1 beliebig stark annähern kann, er kann aber diesen Wert nie genau erreichen. (Dazu wären unendlich viele Produkte notwendig).

**Einsatzbereich des Gini-Koeffizienten**

Der Gini-Koeffizient kann natürlich bei jeder intervallskalierten Verteilung verwendet werden. In der Praxis kommt er aber meist zum Einsatz, um die Einkommens- oder Vermögensverteilung von Menschen in einer bestimmten Region zu beschreiben. Dazu wird der Gini-Koeffizient häufig als Prozentwert zwischen 0% und 100% ausgedrückt. Auf Wikipedia findet man eine Liste der Länder nach Einkommensverteilung[30] in der für jedes Land der Erde, für das es auswertbare Daten gibt, der Gini-Index für die Einkommen der Menschen in diesem Land angegeben ist.

Österreich liegt in dieser Liste auf dem guten Platz 14 (Gini-Index = 27,9%), Deutschland auf Platz 19 (31,1%), die Schweiz auf Platz 28 (32,3%), USA auf Platz 95 (41,5) (Übrigens zwischen Elfenbeinküste auf Platz 94 und Papua-Neuguinea auf Platz 96).

Platz 1 nimmt Slowenien ein (23,4), der letzten Platz (Platz 145) geht an Südafrika (63,0%)

[30] https://de.wikipedia.org/wiki/Liste_der_Länder_nach_Einkommensverteilung

# 4 Multivariate Verteilungen

Multivariate Verteilungen sind Verteilungen, bei denen von jedem Merkmalsträger mehrere Merkmale erfasst wurden (mindestens zwei). Das ist der Fall, wenn beispielsweise von Personen die Körpergröße und das Gewicht erfasst wurden. Es ist aber keineswegs so, dass das nur für verhältnisskalierte Größen gilt. Auch wenn z.B. der Geburtsort und der Nachname erfasst wurden (also zwei nominalskalierte Werte), liegt bereits eine multivariate Liste vor. Allerdings liegt es natürlich auf der Hand, dass man mit intervall- und verhältnisskalierten Werten mehr und vor allem andere Aussagen machen kann als mit ordinal- oder nominalskalierten Werten.

Ebenfalls liegt auf der Hand, dass man mit einzelnen Dimension (einzelnen »Spalten« wenn die Daten als Tabelle vorliegen) einer multivariablen Verteilung dasselbe machen kann wie mit univariaten Verteilungen. Das wurde ja bereits in den Beispielen zum Gini-Koeffizienten so gehandhabt. Dort waren die (nominalverteilten) Farben der T-Shirts gemeinsam mit den (verhältnisskalierten) Jahresumsätzen gegeben, wir haben uns aber nur um die Umsätze gekümmert und die Farben ignoriert.

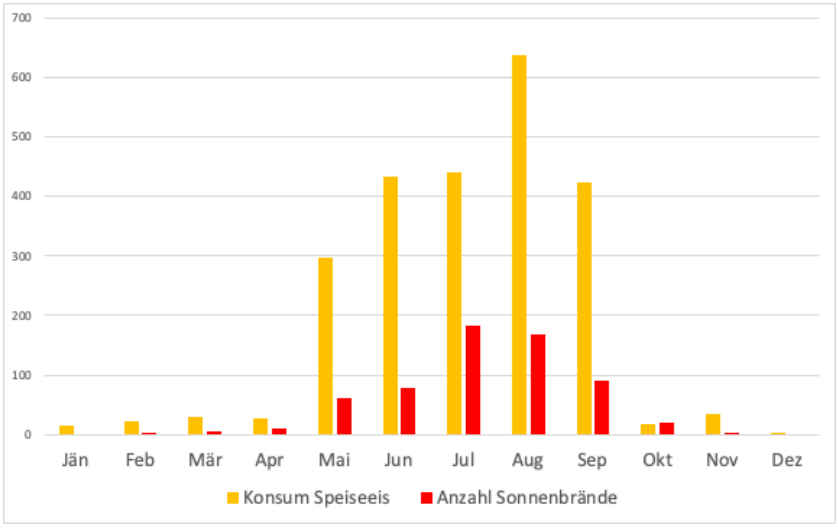
Wenn man aber multivariate Verteilungen im engeren Sinn behandeln will, dann geht es immer um Zusammenhänge zwischen den verschiedenen Werten, die pro Merkmalsträger vorhanden sind.

## 4.1 Korrelation und Kausalzusammenhang

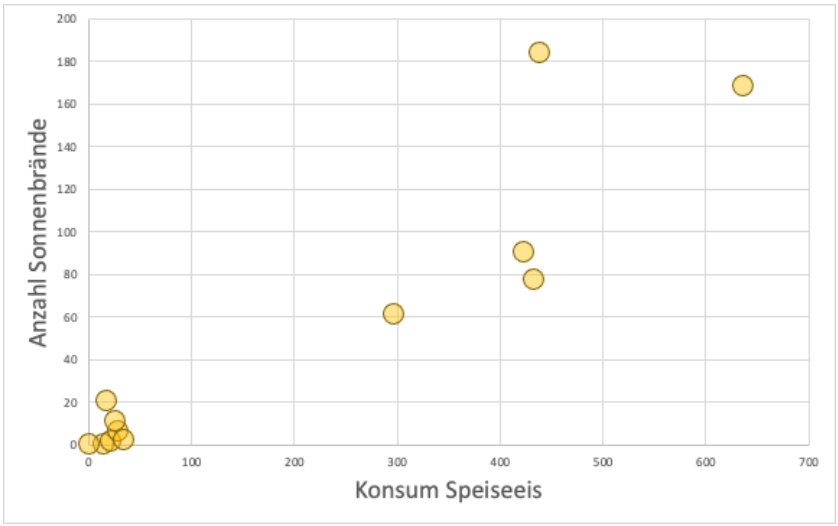
Irgendwo und irgendwann hätte jemand während eines Jahres erheben können, wie viel Speiseeis in jedem Monat verzehrt worden ist, und wie viele Sonnenbrände in derselben Region und im selben Jahr aufgetreten sind. Das Ergebnis könnte so wie in dieser Tabelle aussehen:

| Monat | Konsum Speiseeis | Anzahl Sonnenbrände |
|---|---|---|
| Jänner | 15 | 0 |
| Februar | 23 | 1 |
| März | 29 | 6 |
| April | 26 | 11 |
| Mai | 297 | 61 |
| Juni | 434 | 77 |
| Juli | 440 | 184 |
| August | 637 | 168 |
| September | 424 | 90 |
| Oktober | 18 | 20 |
| November | 35 | 2 |
| Dezember | 1 | 0 |

Trägt man beide Werte gemeinsam in ein gruppiertes Säulendiagramm ein, erkennt man schon eine gewisse Ähnlichkeit beider Verteilungen




Noch besser erkennt man die Zusammenhänge, wenn man stattdessen ein Streudiagramm verwendet:




Man erkennt sehr klar:

Immer dann, wenn nur sehr wenig Speiseeis konsumiert wurde, litt auch kaum jemand an einem Sonnenbrand. Die Häufigkeit von Sonnenbränden steigt deutlich an, wenn die Leute mehr Speiseeis essen.

Die Häufigkeit von Sonnenbränden ist also gut belegbar mit der Menge des konsumierten Speiseeises korreliert.

Dieses Beispiel hat zwar plausible, aber doch frei erfundene Daten verwendet.

In einem anderen Fall wurde aber mit vollem wissenschaftlichem Ernst untersucht, ob der Schokoladenkonsum eines Landes Einfluss auf die Anzahl der Nobelpreisträger hat, die dieses Land hervorgebracht hat.[31]

Dabei kam dieses Diagramm heraus:

Auch hier lässt sich klar belegen, dass die Anzahl der Nobelpreisträger recht gut mit dem Schokoladenkonsum korreliert.

Aber bedeutet diese statistisch belegbare Korrelation auch einen echten Ursache-Wirkung-Zusammenhang? Verursacht der Konsum von Speiseeis Sonnenbrand? Macht Schokolade die Menschen klüger?

Das ist wohl nicht der Fall. Jeder halbwegs vernünftige Mensch wird wissen, dass der erhöhte Konsum von Speiseeis in den Sommermonaten eine Folge der höheren Temperaturen im Freien ist, und dass dieselbe Ursache auch dazu führt, dass sich mehr Menschen spärlich bekleidet der Sonneneinstrahlung aussetzen. In diesem Fall sind die beiden verglichenen Größen also beide die Wirkung einer anderen Größe, die in den Daten nicht erfasst wurde.

Im Beispiel mit den Nobelpreisträgern ist nicht einmal eine gemeinsame Ursache erkennbar. Hier scheint eine rein zufällige Korrelation vorzuliegen.

**Hat Korrelation überhaupt etwas mit Kausalität zu tun?**

Was ist unter dem Eindruck der beiden vorangegangen Beispiel von der Aussage zu halten, dass der vermehrte Konsum von Alkohol die Wahrscheinlichkeit erhöht von einer geraden Linie abzukommen, die man versucht entlang zu gehen?

Diese Frage lässt sich mit Statistik allein nicht beantworten. Dazu braucht man Hypothesen, die als Erklärung für so einen Zusammenhang dienen können. Es ist möglich, dass mehrere unterschiedliche Hypothesen denselben Zusammenhang erklären. Eine Hypothese könnte die Vermutung aufstellen, dass Alkohol das Gravitationsfeld der Erde verzerrt, wodurch es alkoholisierten Menschen natürlich schwerer fällt, geradeaus zu gehen. Eine andere Hypothese könnte davon ausgehen, dass Alkohol die Nervenzellen angreift, sodass sowohl die Wahrnehmung des Gleichgewichts als auch die Fähigkeit darauf zu reagieren beeinträchtigt sind.

Um eine Entscheidung zwischen diesen Hypothesen zu treffen, muss man versuchen Phänomene zu finden, zu denen die Hypothesen unterschiedliche Vorhersagen machen. Beispielsweise würde die Hypothese mit dem verzerrten Gravitationsfeld vorhersagen, dass mit Alkohol gefüllte Flaschen eher umfallen als Flaschen, die mit etwas anderem gefüllt sind, während die Nerven-Hypothese erwarten lässt, dass der Alkoholgehalt keinen Einfluss auf die

[31] »Chocolate Consumption, Cognitive Function, and Nobel Laureates«. Quelle: https://www.nejm.org/doi/full/10.1056/NEJMon1211064

Stabilität von Flaschen hat. Dazu kann man ein Experiment machen und kann nun die aus dem Experiment gewonnen Daten mit statistischen Mitteln auswerten und mit den Vorhersagen der Hypothesen vergleichen. Damit kann man unzutreffende Hypothesen ausschließen.

Einen endgültigen Beweis dafür, dass Hypothesen, die nicht widerlegt wurden, den wahren Zusammenhang erklären, hat man damit allerdings nicht gewonnen. Ein solcher Beweis ist leider nicht möglich. Aber, um Sir Arthur Conan Doyle zu zitieren:

> *»Wenn man das Unmögliche ausgeschlossen hat, muss das, was übrigbleibt, die Wahrheit sein, so unwahrscheinlich sie auch klingen mag.«*[32]

Die Frage, die in der Wissenschaft aber so gut wie nie zufriedenstellend beantwortet werden kann, lautet: »Hat man wirklich alles Unmögliche ausgeschlossen? Oder hat man nicht vielleicht etwas übersehen?«

**Was ist daraus zu lernen?**

Ihnen muss immer klar sein, dass sie als Statistiker nur Korrelationen nachweisen können, und dass sie vermutete kausale Zusammenhänge eventuell ausschließen können. Eine Korrelation taugt nicht als Beweis eines kausalen Zusammenhangs.

## 4.2 Lineare Korrelation

Im vorigen Kapitel haben wir eine Korrelation durch bloßes Betrachten eines Streudiagramms festgestellt. Das wollen wir nun etwas präzisieren und mit Formeln unterstützen. Wir haben beispielsweise gesehen, dass geringer Speiseeiskonsum mit wenigen Fällen von Sonnenbränden einhergehen. Wenn Ihnen jemand sagt, dass in einem nicht näher genannten Monat 100 Einheiten Speiseeis verzehrt wurden, können Sie dann eine Zahl für die wahrscheinlichste Anzahl an Sonnenbränden vorhersagen?

Ist so eine Fragestellung überhaupt sinnvoll, obwohl kein direkter kausaler Zusammenhang besteht? – Ja, so eine Frage ist sinnvoll, denn die Korrelation besteht ja trotzdem, und mehr brauchen wir nicht, um diese Frage beantworten zu können.

Die Frage geht von der Prämisse aus, dass es so etwas wie eine Funktion gibt, die einen Wert für den Speiseeiskonsum als Input entgegennimmt und dann die am ehesten zu erwartende Zahl der Sonnenbrände ausgibt.

[32] Original: *»When you have excluded the impossible, whatever remains, however improbable, must be the truth.«* Doyle hat diese Worte seiner Romanfigur Sherlock Holmes im Roman »The Adventure of the Beryl Coronet« (»Das Abenteuer der Beryl-Krone«) in den Mund gelegt.

### 4.2.1 Ausgleichsgerade durch Nullpunkt und Schwerpunkt

Die einfachste Funktion, die das leisten kann, ist eine Gerade, die durch den Ursprung geht, denn sie kommt mit nur einem Parameter aus, der die Steigung der Geraden angibt:

$$y = mx$$

$x$ ist ein Wert aus der Inputmenge (also eine Eismenge), $m$ ist die Steigung und $y$ ist ein Wert, der zur Outputmenge gehört (eine Sonnenbrand-Anzahl).

Es erscheint auch plausibel, dass diese Gerade durch den Schwerpunkt der Punktwolke verlaufen sollte.

Beim arithmetischen Mittel wurde bereits darauf hingewiesen, dass es den Schwerpunkt einer eindimensionalen Verteilung wiedergibt (siehe 3.4.3). Wenn man alle Punkte der Punktwolke senkrecht nach unten auf die X-Achse projiziert, ist das dasselbe, als würde man die Speiseeismengen als univariate Verteilung betrachten. Daraus folgt, dass die X-Koordinate des Schwerpunkts der Punktwolke genau das arithmetische Mittel der Eismengen ist, und genau dasselbe gilt auch für die andere Achse: Die Y-Koordinate des Schwerpunkts ist genau das arithmetische Mittel der Sonnenbrandzahlen. Der Schwerpunkt der Punktwolke hat daher diese Koordinaten:

$$(\bar{x}, \bar{y})$$

mit

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

Wenn die Gerade durch $(0, 0)$ und $(\bar{x}, \bar{y})$ gehen soll, ist ihre Steigung

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

Setzt man in die Gleichung $y = mx$ nun der Reihe nach die Werte für alle $x_i$ (also für alle Speiseeismengen) ein, werden dabei aber nicht jedes Mal genau die dazu passenden $y_i$-Werte herauskommen. Es gibt also immer eine Differenz zwischen dem wahren Wert aus der Tabelle und dem Wert, den die Ausgleichsgerade liefert:

$$y_i - mx_i$$

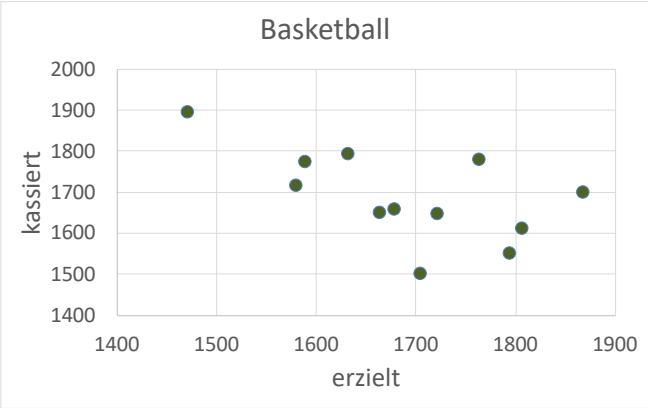
Quadriert man diese Differenzen und summiert sie auf, erhält man wieder eine Quadratsumme, die bei einer ganz bestimmten Steigung ihr Minimum annimmt. Man kann beweisen, dass die Steigung, bei der das der Fall ist, genau jene ist, bei der die Gerade durch den Schwerpunkt der Punktwolke geht. Auf diesen Beweis wird in diesem Dokument aber verzichtet. Dieser Sachverhalt belegt aber, dass unsere Plausibilitätsannahme vom Beginn unserer Überlegungen gerechtfertigt war.

### 4.2.2 **Ausgleichsgerade mit 2 Parametern**

Aber nicht immer macht es Sinn anzunehmen, dass eine Ausgleichsgerade genau durch den Koordinatenursprung gehen muss.

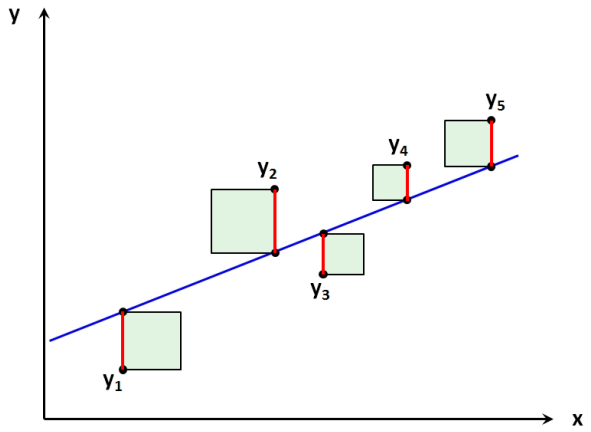
Die folgende Tabelle zeigt, wie viele Punkte Basketballvereine in der zweiten Liga in der gesamten Saison 2019/20 selbst erzielt haben bzw. von den jeweiligen Gegnern einstecken mussten. Daneben die Darstellung dieser Daten als Streudiagramm.

| Verein | eigene Punkte | gegnerische Punkte |
|---|---|---|
| Mattersburg Rocks | 1795 | 1551 |
| Dornbirn Lions | 1868 | 1700 |
| Güssing/Jennersdorf | 1705 | 1502 |
| Fürstenfeld Panthers | 1807 | 1610 |
| Wörthersee Piraten | 1664 | 1651 |
| Mistelbach Mustangs | 1722 | 1647 |
| Basket Flames | 1679 | 1659 |
| BBC Nord Dragonz | 1764 | 1780 |
| BBU Salzburg | 1581 | 1716 |
| KOS Celovec | 1633 | 1793 |
| SWARCO RAIDERS | 1590 | 1775 |
| UDW Alligators | 1472 | 1896 |




Wie man sieht, waren Vereine, die gut offensiv gespielt haben und damit viele Punkte erzielen konnten, auch defensiv gut, so dass deren Gegner nicht so viele Punkte machen konnten. Mit anderen Worten: Je mehr Punkte eine Mannschaft in einer Saison gemacht hat, desto weniger Punkte konnten die Gegner in Summe machen.

Eine Gerade, die diesen Sachverhalt wiedergibt, müsste also von links oben nach recht unten verlaufen und trifft dabei ganz sicher nicht den Nullpunkt. Es ist aber auch hier plausibel anzunehmen, dass sie durch den Schwerpunkt der Punktwolke gehen sollte.




Eine Gerade, die möglichst gut der Verteilung der Punkte folgt, sollte auch die Summe der Quadrate der Abstände der Punkte von der Geraden minimieren.

$$\sum_{i=1}^{n}(g(x_i)-y_i)^2 \quad \rightarrow \quad min$$

Wenn man diesen Abstand immer parallel zur Y-Achse misst, führt diese Forderung automatisch zu diesen Gleichungen:

Allgemeine Geradengleichung:

$$y = a \cdot x + b$$

Dabei ist $b$ der Abschnitt auf der Y-Achse, an der die Gerade diese Achse schneidet und $a$ ist die Steigung.

$$a = \frac{\left(\sum x_i^2\right) \cdot \left(\sum y_i\right) - \left(\sum x_i\right) \cdot \left(\sum x_i y_i\right)}{n \cdot \left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

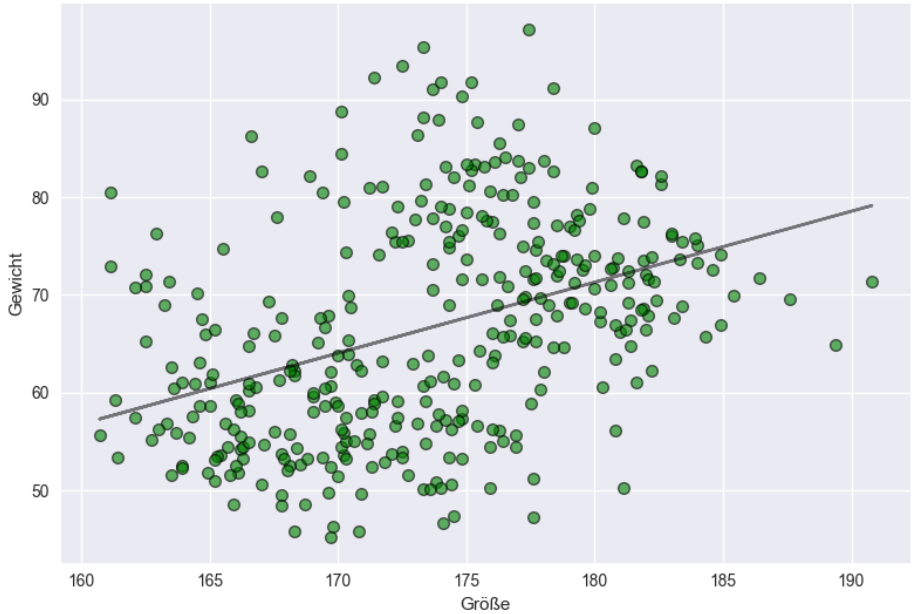
$$b = \frac{n \cdot \left(\sum x_i y_i\right) - \left(\sum x_i\right) \cdot \left(\sum y_i\right)}{n \cdot \left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

Alle Summen laufen von $i = 1$ bis $n$.

### 4.2.3 Orthogonale Regression

Anstatt die Abstände parallel zur Y-Achse zu messen, kann man sie auch im rechten Winkel zur Geraden messen und kann wieder versuchen, die Quadrate dieser Abstände zu minimieren. Dann bekommen Sie auch eine Ausgleichsgerade, allerdings im allgemeinen Fall nicht dieselbe. Die Formeln dafür sind um einiges komplizierter, und in der Praxis ist meist die gewöhnliche Ausgleichsgerade ohnehin genau das, was man haben will.

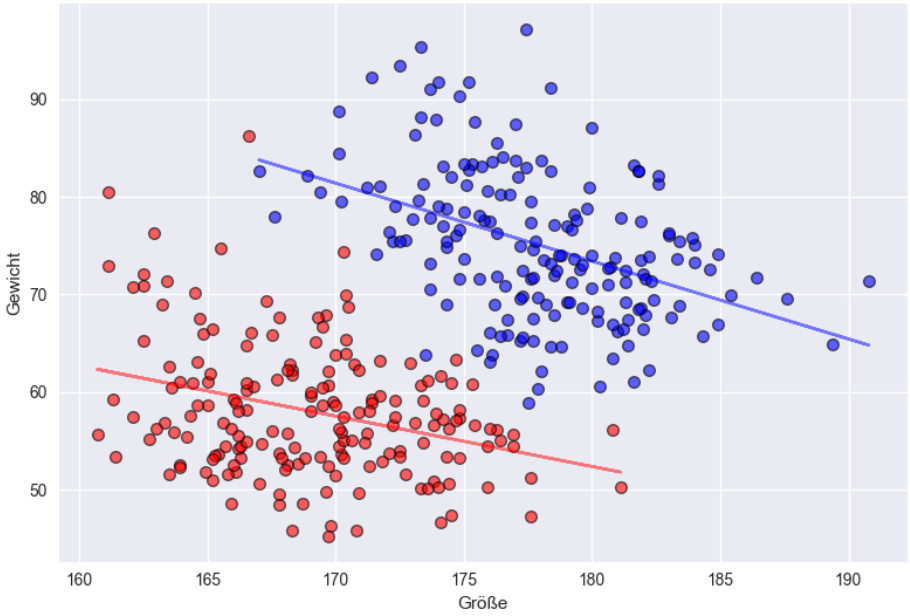
### 4.2.4 Simpson-Paradoxon[33]




[33] Benannt nach dem britischen Statistiker Edward Hugh Simpson (1922-2019)

Das Diagramm mit den grünen Punkten zeigt die Größen und Gewichte von 178 Männern und 178 Frauen und eine Ausgleichsgerade, die den Zusammenhang zwischen Größe und Gewicht in dieser Stichprobe am besten beschreibt.

Die untenstehende Grafik zeigt genau dieselben Daten, nun aber nach Geschlechtern getrennt (rot: Frauen, blau: Männer). Die beiden Ausgleichsgeraden gelten jeweils nur für ein Geschlecht.

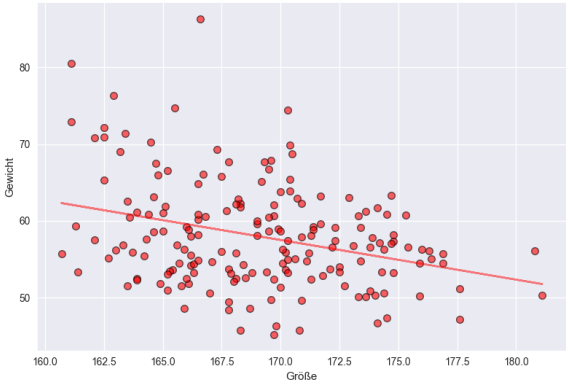


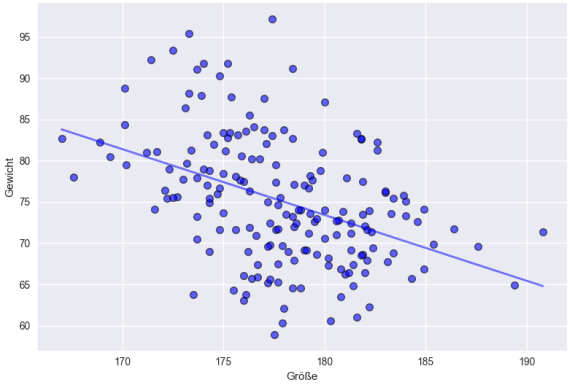

Was kann man nun aus diesen Daten für Schlüsse ziehen? Was gilt für diese Stichprobe von 356 Menschen?

a) Je größer jemand ist, desto schwer ist diese Person.

b) Je größer jemand ist, desto leichter ist diese Person.

Was wäre, wenn man Ihnen die Auswertungen der Frauen und jene der Männer nicht in einem Diagramm gemeinsam zeigt, sondern in zwei getrennten Grafiken?




Dieses Phänomen, dass bestimmte Teilmengen einer Stichprobe auf andere Zusammenhänge hindeuten als die gesamte Stichprobe, nennt man das Simpson-Paradoxon.

Das geschilderte Beispiel entstand, weil aus der Grundgesamtheit systematisch kleine dünne und große dicke Personen ausgeschieden wurden, und nur kleine dicke und große dünne Menschen in die Stichprobe kamen. Es wurde also ein Fehler bei der Stichprobenziehung gemacht (in diesem Fall absichtlich, um das Phänomen zu demonstrieren).

**Ein andere Beispiel:**

An der Universität Berkley führte dieses Paradoxon zu einer Klage, die vor Gericht ausgefochten werden musste. Im Jahr 1973 ergab sich bei der Auswertung der Bewerbungen an der Universität dieses Bild:

| | Bewerber | davon zugelassen |
|---|---|---|
| Männer | 8442 | 44% |
| Frauen | 4321 | 35% |

Wenn man sich als Mann beworben hatte, betrug die Chance, genommen zu werden, 44%. Wenn man eine Frau war, nur 35%. Der Unterschied ist so groß, dass er nicht durch Zufall zu erklären ist, daher wurde die Uni wegen Diskriminierung verklagt. Die Universität konnte aber nachweisen, dass in der Mehrheit aller Departments die Frauen höhere Zulassungsraten hatten als Männer, daher wurde die Klage abgewiesen.

Es zeigte sich nämlich, dass sich Frauen vor allem in jenen Departments beworben hatten, in denen die Chance genommen zu werden schlecht ist, während in den Departments, wo auf einen freien Platz weit weniger Bewerber kommen, sich hauptsächlich Männer beworben haben.

Dies soll anhand eines weiteren konstruierten Beispiels gezeigt werden. Eine bestimmte Institution ist in zwei Departments organisiert. Beide Departments bieten jeweils 100 offene Stellen an. Insgesamt bewerben sich 1005 Personen für die 200 offenen Stellen. Sie verteilen sich wie folgt:

| Department A | Bewerber | zugelassen | Prozent |
|---|---|---|---|
| Männer | 200 | 95 | 47,50% |
| Frauen | 5 | 5 | 100,00% |

In Department A haben sich nur 5 Frauen beworben, sie wurden alle zugelassen. Alle Stellen, für die es keine weiblichen Bewerber gab, wurden mit Männern besetzt. Das Department hat also jede verfügbare Frau genommen. Mehr Frauen zu nehmen, war aufgrund der Bewerberlage nicht möglich.

| Department B | Bewerber | zugelassen | Prozent |
|---|---|---|---|
| Männer | 200 | 5 | 2,50% |
| Frauen | 800 | 95 | 11,88% |

In Department B wurden von 100 ausgeschriebenen Stellen 95 mit Frauen und nur 5 mit Männern besetzt. Wenn eine Frau zum Bewerbungsgespräch kam, hatte sie eine fast fünfmal größere Chance genommen zu werden als ein männlicher Bewerber.

Sowohl in Department A als auch in Department B liegt die Aufnahmequote bei den Frauen signifikant über der Quote der Männer. Das lässt den Schluss zu, dass in beiden Departments Frauen stark bevorzugt und Männer benachteiligt wurden.

Betrachte man aber die ganze Institution, dann erkennt man, dass die Aufnahmequote bei den Männern doppelt so groß ist wie bei den Frauen.

| Gesamt | Bewerber | zugelassen | Prozent |
|---|---|---|---|
| Männer | 400 | 100 | 25,00% |
| Frauen | 805 | 100 | 12,42% |

Es lässt sich mit statistischen Methoden klar zeigen, dass dieser Unterschied nicht mehr durch Zufall erklärt werden kann. Dass man als Mann eine doppelt so große Chance hat, bei dieser Institution genommen zu werden, muss also einen systematischen Grund haben. Das lässt den Schluss zu, dass Männer bei in den Aufnahmeverfahren bevorzugt und Frauen benachteiligt wurden.

Welche Sichtweise ist richtig? Warum?