Statistik 1

Wintersemester 2023/24

Prof. Matthias Guggenmos

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

Part I

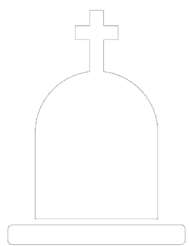
Vorlesung 01: Intro

1 Paradoxia

1.1 Eine rätselhafte Erkrankung

Eine rätselhafte Erkrankung hat seit geraumer Zeit die Bundesrepublik Deutschland heimgesucht. Die Betroffenen zeigen ein paradoxes emotionales Verhalten: sie reagieren allem Anschein nach traurig auf fröhliche Ereignisse und fröhlich auf traurige Ereignisse – als wären zwei Hirndrähte vertauscht worden!

Aufgrund ihres Erscheinungsbilds wird die neuartige Krankheit **Paradoxia** genannt.


1.2 Weiteres Mysterium: blaue Haare

Noch ist wenig bekannt über die Erkrankung. Eine weitere Auffälligkeit ist, dass viele der Betroffenen, eine Neigung haben, ihre Haare blau zu färben. Warum das so ist, ist unklar, auch wenn gemutmaßt wird, dass hierfür ein viral gegangenes Video einer Betroffenen verantwortlich ist.

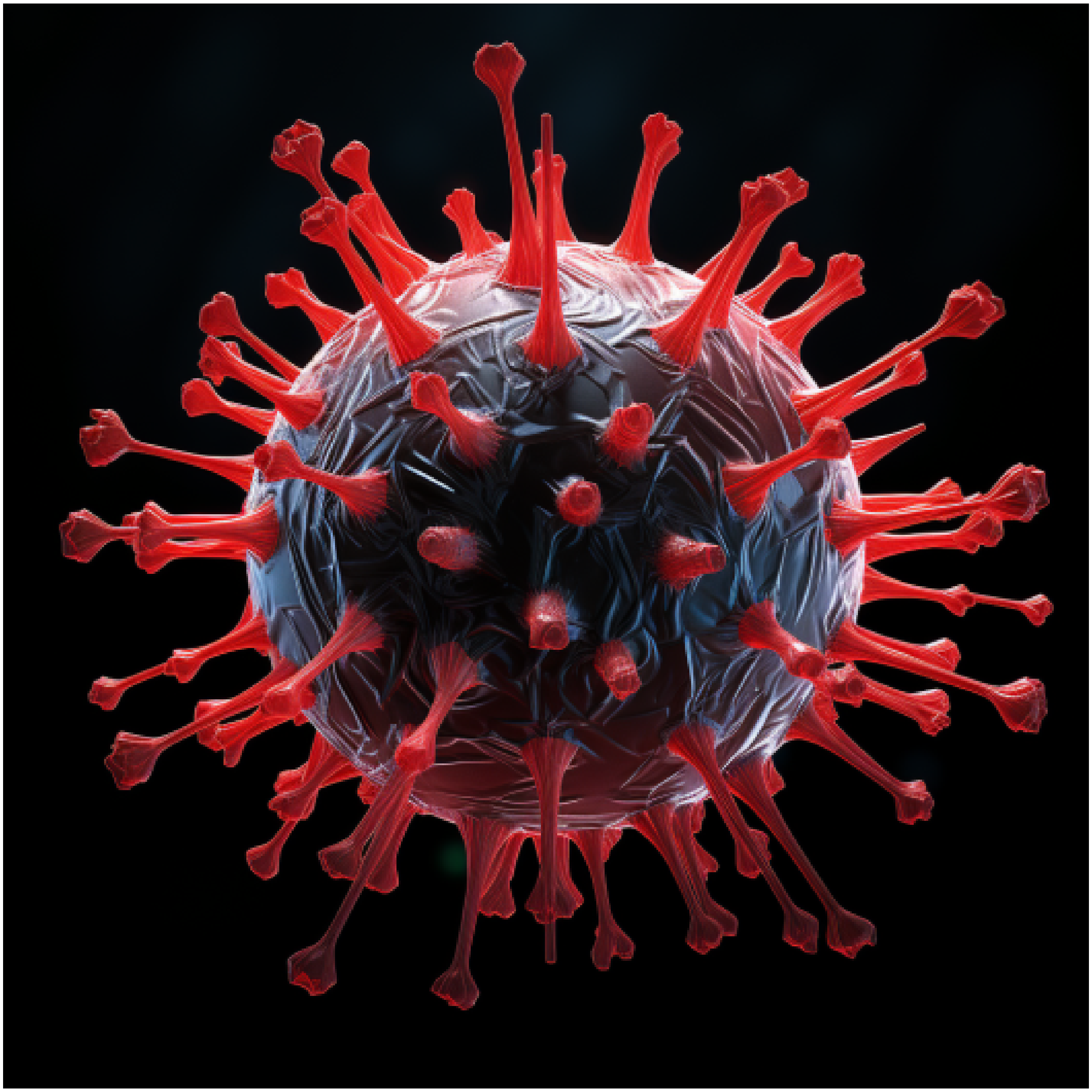
1.3 Task Force

Durch die schnelle Verbreitung der Erkrankung sind das Bundeskanzleramt und die Geheimdienste alarmiert. Schnell wird der Beschluss gefasst, eine Task Force bestehend aus Psycholog:innen, Mediziner:innen und Sozialforscher:innen ins Leben zu rufen, um die Erkrankung besser zu verstehen.

Sie sind Teil dieser Task Force im **Team Psychologie**.

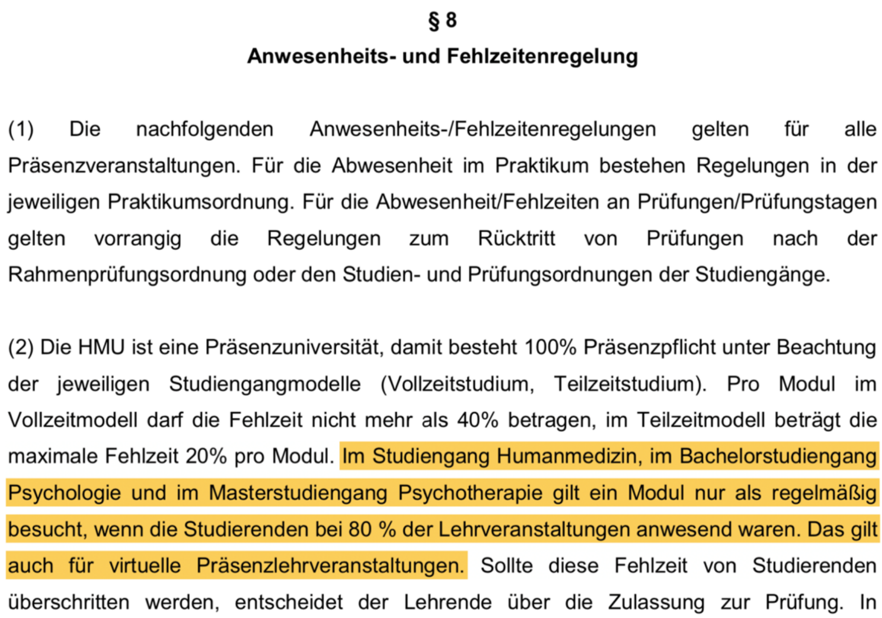
1.4 Eine erste Aufgabe: Hypothesenbildung

Ihre erste Aufgabe besteht darin, Hypothesen zu bilden, wie es zum Auftreten der Erkrankung gekommen sein könnte. Vermutet wird entweder eine **neuartige Viruserkrankung**, die das zentrale Nervensystem befällt, oder ein **sozialpsychologisches Ansteckungsphänomen**.



2 Formalitäten

2.1 Anwesenheits- und Fehlzeitregelung



§ 8

Anwesenheits- und Fehlzeitenregelung

(1) Die nachfolgenden Anwesenheits-/Fehlzeitenregelungen gelten für alle Präsenzveranstaltungen. Für die Abwesenheit im Praktikum bestehen Regelungen in der jeweiligen Praktikumsordnung. Für die Abwesenheit/Fehlzeiten an Prüfungen/Prüfungstagen gelten vorrangig die Regelungen zum Rücktritt von Prüfungen nach der Rahmenprüfungsordnung oder den Studien- und Prüfungsordnungen der Studiengänge.

(2) Die HMU ist eine Präsenzuniversität, damit besteht 100% Präsenzpflicht unter Beachtung der jeweiligen Studiengangmodelle (Vollzeitstudium, Teilzeitstudium). Pro Modul im Vollzeitmodell darf die Fehlzeit nicht mehr als 40% betragen, im Teilzeitmodell beträgt die maximale Fehlzeit 20% pro Modul. Im Studiengang Humanmedizin, im Bachelorstudiengang Psychologie und im Masterstudiengang Psychotherapie gilt ein Modul nur als regelmäßig besucht, wenn die Studierenden bei 80 % der Lehrveranstaltungen anwesend waren. Das gilt auch für virtuelle Präsenzlehrveranstaltungen. Sollte diese Fehlzeit von Studierenden überschritten werden, entscheidet der Lehrende über die Zulassung zur Prüfung. In

Figure 2.1: Bildnachweis[1]

2.2 Nützliche Links

Statistik für Psychologie	https://www.statistikpsychologie.de/blog/	Einige nützliche Blogeinträge zu statistischen Themen der Psychologie
DATAtab	https://datatab.de/tutorial/	Online-Kalkulator mit einigen guten Tutorials
StatistikGuru	https://statistikguru.de/	Fokus auf SPSS und R, aber viele Tutorials in allgemeinem Stil geschrieben

[1] [redacted] Allgemeine

3 Was ist Statistik?

3.1 Was ist Statistik?

> Das Wort Statistik stammt von lateinisch *statisticum* „den Staat betreffend“ und italienisch *statista* Staatsmann oder Politiker [...]

Quelle: Wikipedia

In diesem Sinne beschrieb der Ausdruck Statistik ursprünglich die Lehre von den Daten über den Staat. Heute ist der Begriff wesentlich breiter gefasst:

Statistik ist ein angewandter Zweig der Mathematik, der sich mit der **Erhebung**, **Analyse** und **Interpretation** empirischer Daten befasst.

Table 3.1: Beispiele

Erhebung	Wie groß sollte meine Stichprobe sein?
Analyse	Was ist der Mittelwert meiner Variable?
Interpretation	Unterscheiden sich zwei Gruppen systematisch?

Figure 3.1: Bildnachweis[1] [2]

3.2 Wozu Statistik?

Auch ohne Statistik sind Menschen sehr gut darin, Muster oder Zusammenhänge in ihrer Umwelt zu erkennen. Man könnte diese Fähigkeit als **intuitive Empirie** bezeichnen. Wozu also Statistik?

[1] [redacted] https://www.analyti
[2] [redacted] https://www.livescie

Probleme intuitiver Empirie (kleiner Ausschnitt)

- Kognitive Biases (z.b. Konfirmationsbias)
- Schwächen beim probabilistischen Denken (Stichwort *Lotto*)
- Übergeneralisierung / Tendenz zur vorschnellen Verallgemeinerung
- Übersensitivität: Muster erkennen, wo keine sind
- Vermischung mit eigenen Interessen (z.B. Politik)
- Begrenztes (Arbeits)Gedächtnis: Unfähigkeit, große Datenmengen im Kopf zu behalten und zu verrechnen

Figure 3.2: Bildnachweis[3]

3.3 Beispiel Prävalenzfehler

Ein neuentwickelter Coronatest wird damit beworben, dass er anhand einer Speichelprobe in 100% der Fälle eine bestehende Coronaerkrankung erkennt. Im Kleingedruckten findet sich der Hinweis, dass die Spezifität 90% beträgt, d.h. der Test weist in 10% der Fälle eine *gesunde Person fälschlicherweise als krank* aus.

[3] [redacted]https://www.arte.tv

Nehmen Sie an, die Prävalenz von Corona ist 1%, d.h. eine von hundert Personen ist mit Corona infiziert.

Schätzen Sie: Angenommen Ihr Test ist positiv, in wie viel Prozent der Fälle sind Sie dann auch tatsächlich positiv?

3.4

Placeholder image!!!

3.5 Beispiel Prävalenzfehler

Auflösung:

- von 100 Personen ist im Schnitt 1 Person krank. Diese wird durch den Test auf jeden Fall korrekt als krank erkannt.
- 99 Personen sind nicht erkrankt, und bei 10% dieser Personen gibt der Test ein falsch-positives Ergebnis (10 Personen).
- D.h. bei 100 Personen haben 11 Personen ein positives Ergebnis, aber bei 10 von 11 Personen (91%) ist dieses Ergebnis falsch!

Unsere empirische Intuition versagt in diesem Fall häufig. Im konkreten Beispiel spricht man auch vom **Prävalenzfehler** oder der **Base rate fallacy**.

3.6 Beispiel Meditonsin

> Meditonsin ist das meistverkaufte homöopathische Erkältungsmittel in Deutschland, mit einem Jahresumsatz von deutlich über zwanzig Millionen Euro.

> Jedenfalls entwickelte ein Chemieprofessor für Medice eine Apparatur, die alle Zutaten von Meditonsin bis heute in 250-Liter-Fässern vermischt und am Ende zehnmal kräftig in Richtung Erdmittelpunkt schüttelt.

[4] [redacted]https://thedecisionla

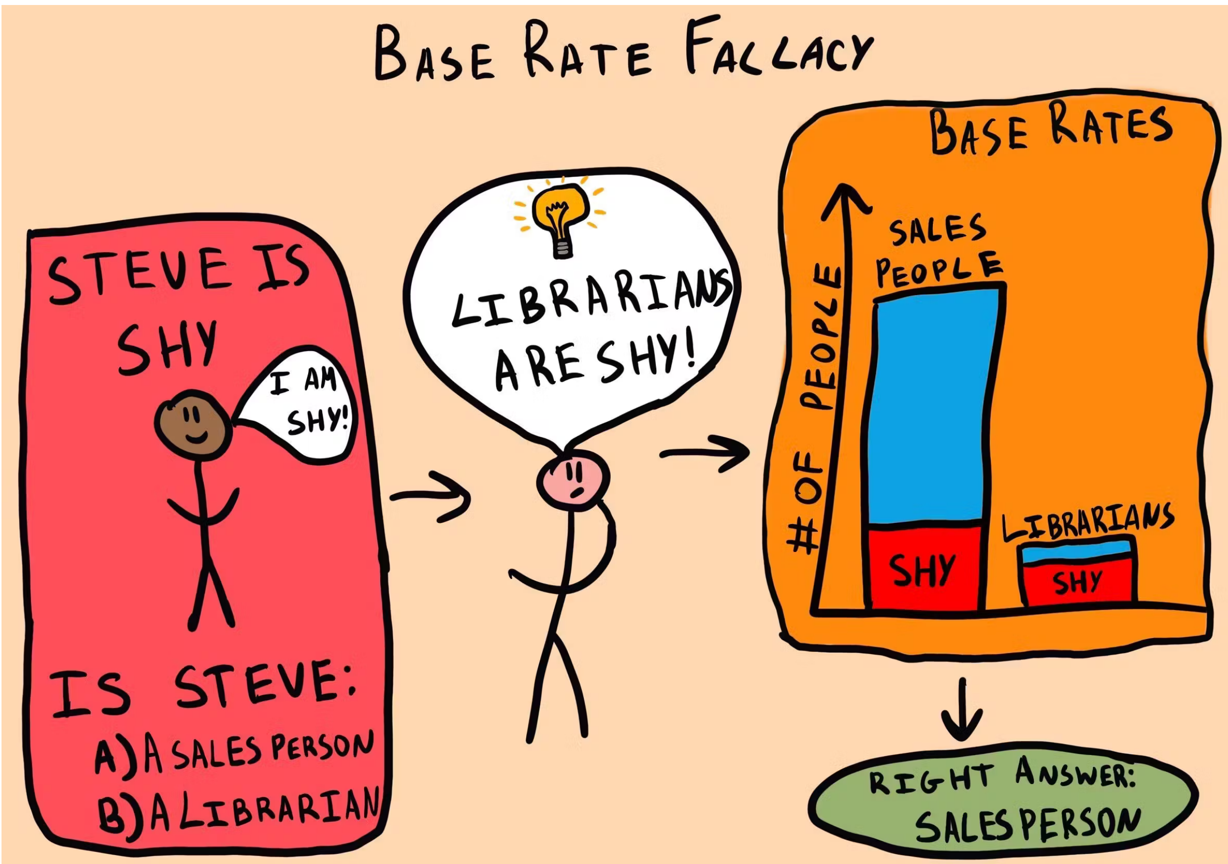


Figure 3.3: **Weiteres Beispiel:** ist die Häufigkeit/Prävelenz von "Sales people" deutlich größer als die der "Librarians", so ist die Wahrscheinlichkeit groß, dass eine schüchterne Person mit großer Wahrscheinlichkeit eine "Sales person" – selbst dann, wenn Sales people mehrheitlich eher nicht schüchtern sind.[4]

> Meditonsin enthält laut Beipackzettel Tollkirsche und Eisenhut in D5-Auflösung [0,00001%] sowie Quecksilber in D8-Auflösung [0,00000001%].

Quelle: Sueddetsche Zeitung [5]

DAZ.online

Unabhängige pharmazeutische Informationen für Wissenschaft und Praxis der Deutschen Apotheker Zeitung

RECHTSKRÄFTIGES URTEIL

Meditonsin-Werbung mit Apotheken-Studie unzulässig

BERLIN - 03.05.2023, 14:45 UHR

Meditonsin-Hersteller Medice darf für seine homöopathischen Tropfen nicht mehr damit werben, dass in einer „apothekenbasierten Beobachtungsstudie" die gute Wirksamkeit und Verträglichkeit des Mittels nachgewiesen wurde. Ein entsprechendes Urteil des Landgerichts Dortmund aus dem vergangenen Herbst ist jetzt rechtskräftig geworden. Geklagt hatte die Verbraucherzentrale Nordrhein-Westfalen.

Figure 3.4: Laut einem Gerichtsurteil aus dem Mai 2023 darf Meditonsin nicht mehr mit einer vorgeblich wissenschaftlich bewiesenen Wirksamkeit werben.[6]

3.7 Beispiel Meditonsin

Woher kommt das große Vertrauten in ein Medikament, bei dem eine Flasche im Schnitt kaum ein Molekül der vorgeblichen Wirksubstanzen enthält?

[5] [redacted] https://sz-
[6] https://www.astrono

Warum entsteht bei vielen Menschen subjektiv der Eindruck, Medikamente wie Meditonsin würden wirken?

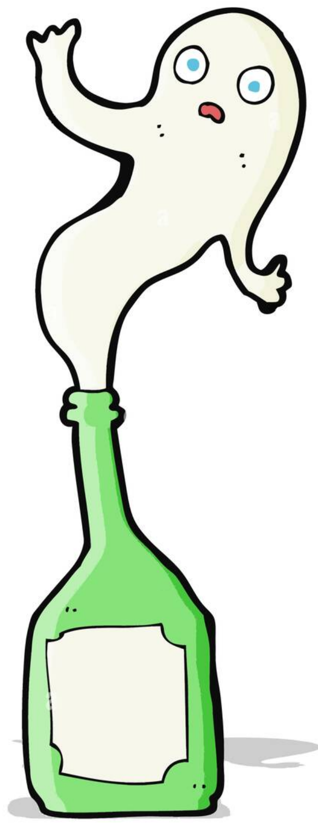


Figure 3.5: Der Geist in der Flasche.[7]

3.8 Beispiel Meditonsin

Eine Erklärung ist mangelndes Bewusstsein für einen Effekt der auch als **Regression zur Mitte** bezeichnet wird.

- Meditonsin wird vermutlich häufig eingenommen, wenn eine Erkältung auf ihrem Höhepunkt ist.
- *In aller Regel verbessert* sich der Gesundheitszustand nach Erkältungen auch ohne Medikation und pendelt sich wieder auf einen Normalzustand ein (dies ist die "Regression" zur "Mitte").
- Dadurch entsteht eine subjektiv wahrgenommener Zusammenhang Einnahme Meditonsin Heilung

Statistische Methoden ermöglichen hier in Kombination mit einem adäquaten Studiendesign (Interventions- und Placebogruppe) eine objektivere Einschätzung.

Empfehlung: Quarks Science Podcast "Meditonsin, Umckaloabo, Grippostad: Die Akte Erkältungsmittel" [8]

7 [redacted] https://www.alamy.c
8 https://www.quarks.
9 https://theconversat

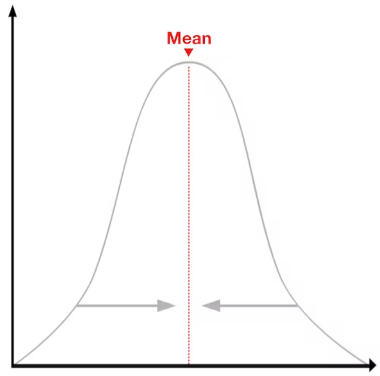


Figure 3.6: Regression zur Mitte. [9]

3.9 Deskriptive Statistik und Inferenzstatistik

In Statistik 1 und 2 werden wir fast ausschließlich Methoden aus der deskriptiven Statistik und Inferenzstatistik behandeln.

Das Ziel der **deskriptiven Statistik** ist die Beschreibung und Zusammenfassung von Daten innerhalb einer Stichprobe.

Beispielfragen: Wie hoch ist der Mittelwert einer Variable in der Stichprobe? Wie hoch ist die Korrelation zwischen Variable X und Variable Y in der Stichprobe?

Das Ziel der **Inferenzstatistik** sind Schlussfolgerungen über eine Population auf der Grundlage von Stichproben.

Beispielfragen: Wie groß ist die durchschnittliche Körpergröße der Bevölkerung, basierend auf einer Stichprobe von 500 Personen? Gibt es einen signifikanten Unterschied im durchschnittlichen Einkommen zwischen zwei verschiedenen Berufsgruppen?

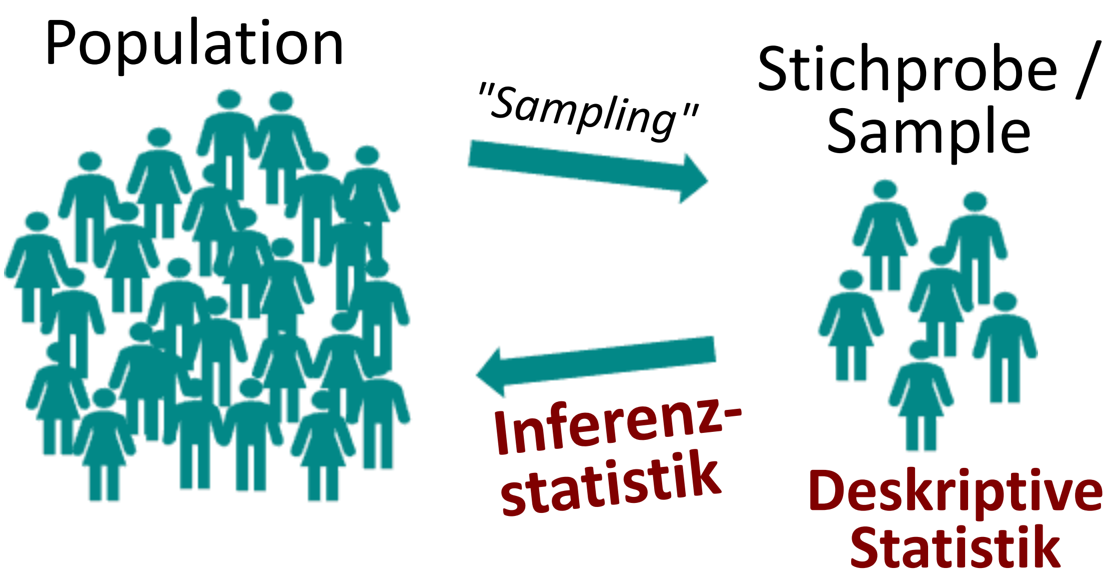


Figure 3.7: Bildnachweis[10]

Ein alternativer Ausdruck für Inferenzstastistik ist **Induktive Statistik**.

10 [redacted]https://datatab.net/

3.10

Zurück zum Thema: wie entwickelt man gute Hypothesen? Und was ist eigentlich eine Hypothese?

4 Hypothesen

4.1 Der Forschungsprozess

```
setup:
    - element: "#hypothesis"
      modifier: function() { this.node.style.fill = 'green'; }
    - element: "#hypothesisbg"
      modifier: function() { this.node.style.fill = '#d8ffe2';}
```

4.2 Was ist eine Hypothese?

Hypthesen: in Form einer **logischen Aussage** formulierte **Annahme** zur Erklärung **empirischer Phänomene**.

Beispiele

- Es gibt einen positiven Zusammenhang zwischen der Feinstaubbelastung einer Stadt und der Prävalenz von Lungenerkrankungen.
- Eine vorherige Impfung mit dem Corona-Impfstoff verringert die Symptomschwere im Fall einer Infektion.

4.3 Gute Hypothesen, schlechte Hyothesen

Eine wissenschaftliche Hypothese muss drei Grundvoraussetzungen erfüllen: **Widerspruchsfreiheit**, **Widerlegbarkeit**, **Operationalisierbarkeit**.

Beispiele

- Alle Lebewesen benötigen Sauerstoff, aber manche Lebenwesen sind anaerob [~~widerspruchsfrei~~, ~~widerlegbar~~, operationalisierbar]
- Könnten Babies sprechen, würden sie hauptsächlich ihre Gefühle zum Ausdruck bringen [widerspruchsfrei, operationalsierbar, ~~widerlegbar~~]
- Bewusstsein ist eine notwendige Voraussetzung um zu fühlen [widerspruchsfrei, ~~operationalisierbar~~, ~~widerlegbar~~].

Beachte: Hypothese, die nicht widerspruchsfrei oder nicht operationalisierbar sind, sind auch nicht widerlegbar (die umgekehrte Aussage gilt nicht in allen Fällen).

Operationalisieru Messbarmachung eines theoretischen Konstrukts (Schwerkraft, Intelligenz, psychische Resilienz, …) durch ein

4.4

Eine *gute* wissenschaftliche Hypothese erfüllt darüber hinaus weitere Kriterien:

Prädiktivität	*Hypothesen sollten präzise Vorhersagen ermöglichen*Negativbeispiel: Das Selbst ist in ein Es, Ich, Über-Ich aufgeteiltPositivbeispiel: Eine tägliche morgendliche Meditationsübung verringert das subjektive und objektive Stresslevel
Spezifität	*Hypothesen sollten spezifisch und konkret sein*Negativbeispiel: Frühe Kindheitstraumata wirken sich nachteilig auf spätere Beziehungen ausPositivbeispiel: Kindheitstraumata in einem vorsprachlichen Alter, verursacht durch wichtige Bezugspersonen, erhöhen die Wahrscheinlichkeit von Bindungsangst in engen romantischen Beziehungen
Sparsamkeit	*Hypothesen sollten keine unnötige Komplexität beinhalten*Negativbeispiel: Die Coronapandemie verstärkt das Gefühl der Vereinsamung in älteren Bevölkerungsschichten, falls diese an einem Dienstag befragt wurden.Positivbeispiel: Die Coronapandemie verstärkt das Gefühl der Vereinsamung in älteren Bevölkerungsschichten~~, falls diese an einem Dienstag befragt wurden~~.
Relevanz	*Hypothesen sollten von theoretischer oder gesellschaftlicher Relevanz sein*Negativbeispiel: Menschen mit roten Autos bevorzugen rote GemüsesortenPositivbeispiel: Nach persönlichen Erfolgserlebnissen neigen Menschen zu riskantem Fahrvehalten

4.5 Gerichtete und ungerichtete Hypothesen

Ungerichtete Hypothesen geben keine Richtung des Effektes an:

- Frauen und Männer haben unterschiedliche viele soziale Kontakte
- Die Punktzahl im Abitur hängt zusammen mit der späteren Lebenszufriedenheit

Gerichtete Hypothesen geben die Richtung des Effektes an:

- Frauen haben *mehr* soziale Kontakte als Männer
- Eine höhere Punktzahl im Abitur korreliert *negativ* mit der späteren Lebenszufriedenheit

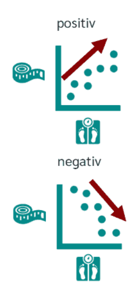


Figure 4.1: Bildnachweis[1]

[1] [redacted]https://datatab.de/t

4.6 Probabilistische Hypothesen

In der Psychologie sind Hypothesen in aller Regel **probabilistisch**, d.h. sie machen Aussagen über Wahrscheinlichkeiten und Mittelwerte.

Beispiel deterministische Hypothese: Menschen mit Alkoholabhängigkeit entwickeln eine Depression.

Beispiel probabilistische Hypothese: Menschen mit Alkoholabhängigkeit haben eine *erhöhte Wahrscheinlichkeit* für eine Depression.Alternativ: Menschen mit Alkoholabhängigkeit haben *im Mittel* eine höhere Ausprägung depressiver Symptome.

- Psychologische Phänomene sind in der Regel von einer Vielzahl von Faktoren beeinflusst, die zum Großteil nicht bekannt sind (biologisch–z.B. Blutdruck, psychologisch–z.B. Stimmung, situativ–z.B. Versuchslabor versus Realität).
- Diese Faktoren erhöhen die Variabilität von Beobachtungen und führen dazu, dass die meisten psychologischen Phänomene nur durch Mittelwertbildung über eine Reihe von Messwiederholungen sichtbar werden.

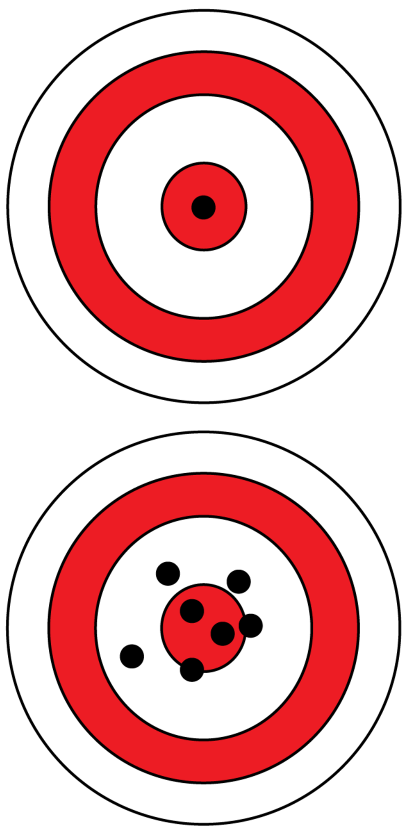


Figure 4.2: Bildnachweis[2]

4.7

Ok, Task Force: wie lauten eure (Arbeits-)Hypothesen zu den Hintergründen von Paradoxia?

[2] [redacted] https://www.toolsgr

Hypothese 1: Paradoxia wird durch eine virale Infektion verursacht. *Die Blutproben von Menschen mit Paradoxia sollten daher eine höhere Menge von Entzündungsmarkern aufweisen.*

Hypothese 2: Die Blaufärbung der Haare ist ein reines Tik-Tok-Phänomen. *Betroffene in Ländern, in denen TikTok gesperrt ist, sollten daher keine erhöhte Neigung zur Blaufärbung der Haare haben.*

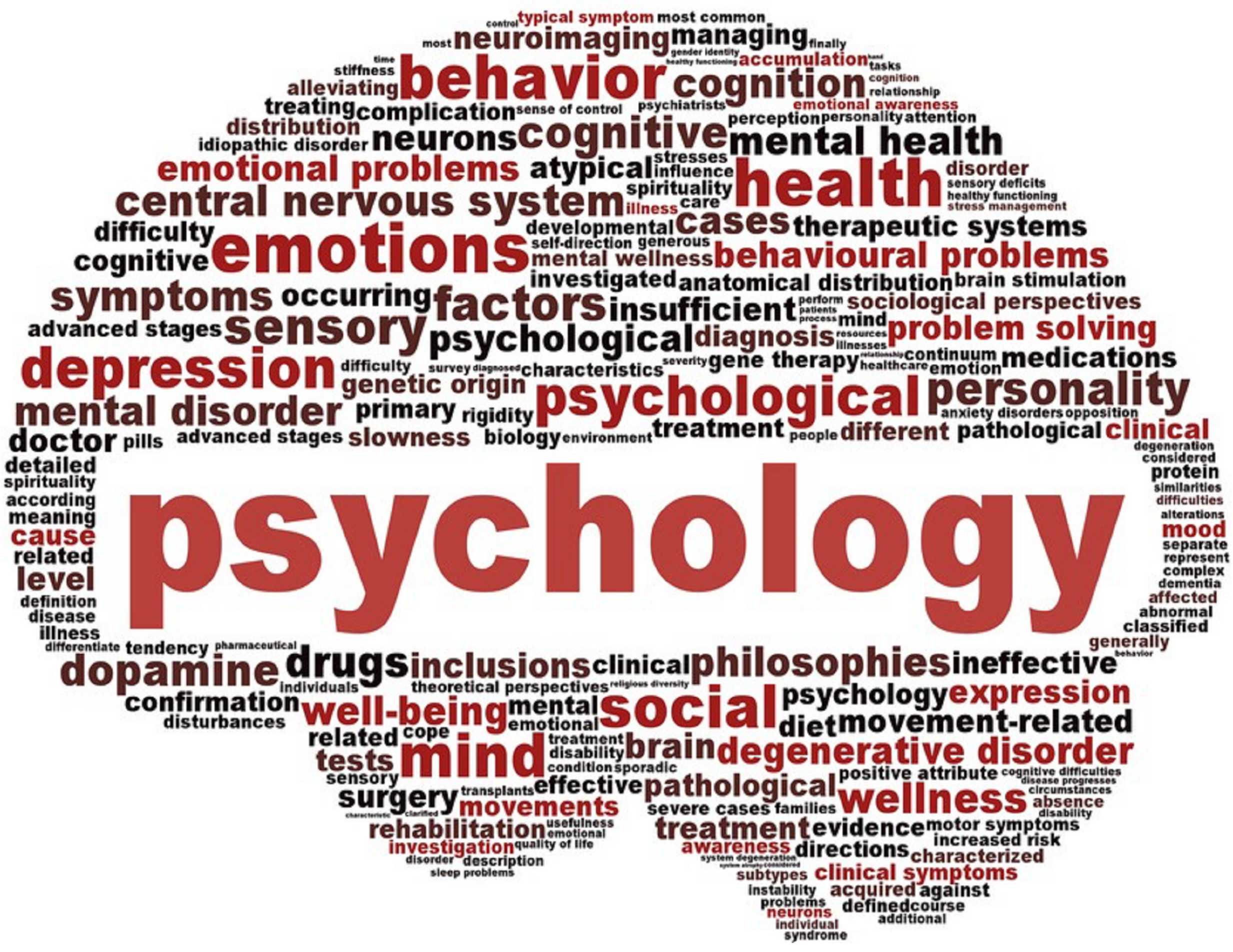
Part II

Vorlesung 02: Messen in der Psychologie

5 Vorlesung 02: Messen in der Psychologie

5.1 Ohne Maßband oder Waage – Wie misst man Psyche?

- Die meisten Variablen menschlicher Psychologie sind keine “natürlichen Größen”, die sich mit konventionellen Instrumenten (z.B. Meterstab) messen ließen – sie sind latent.
- In der Regel basieren diese Variablen auf bestimmten Vorstellungen oder Theorien über die Psyche (Optimismus, Ängstlichkeit, Intelligenz, usw.) – man spricht hier auch von **Konstrukten**.
- Um messbar zu werden, müssen Konstrukte **operationalisiert** werden, d.h. es müssen manifeste messbare **Variablen** definiert werden, die – ggf. in Kombination (z.B. Fragebogen) – Rückschluss auf das Konstrukt erlauben.

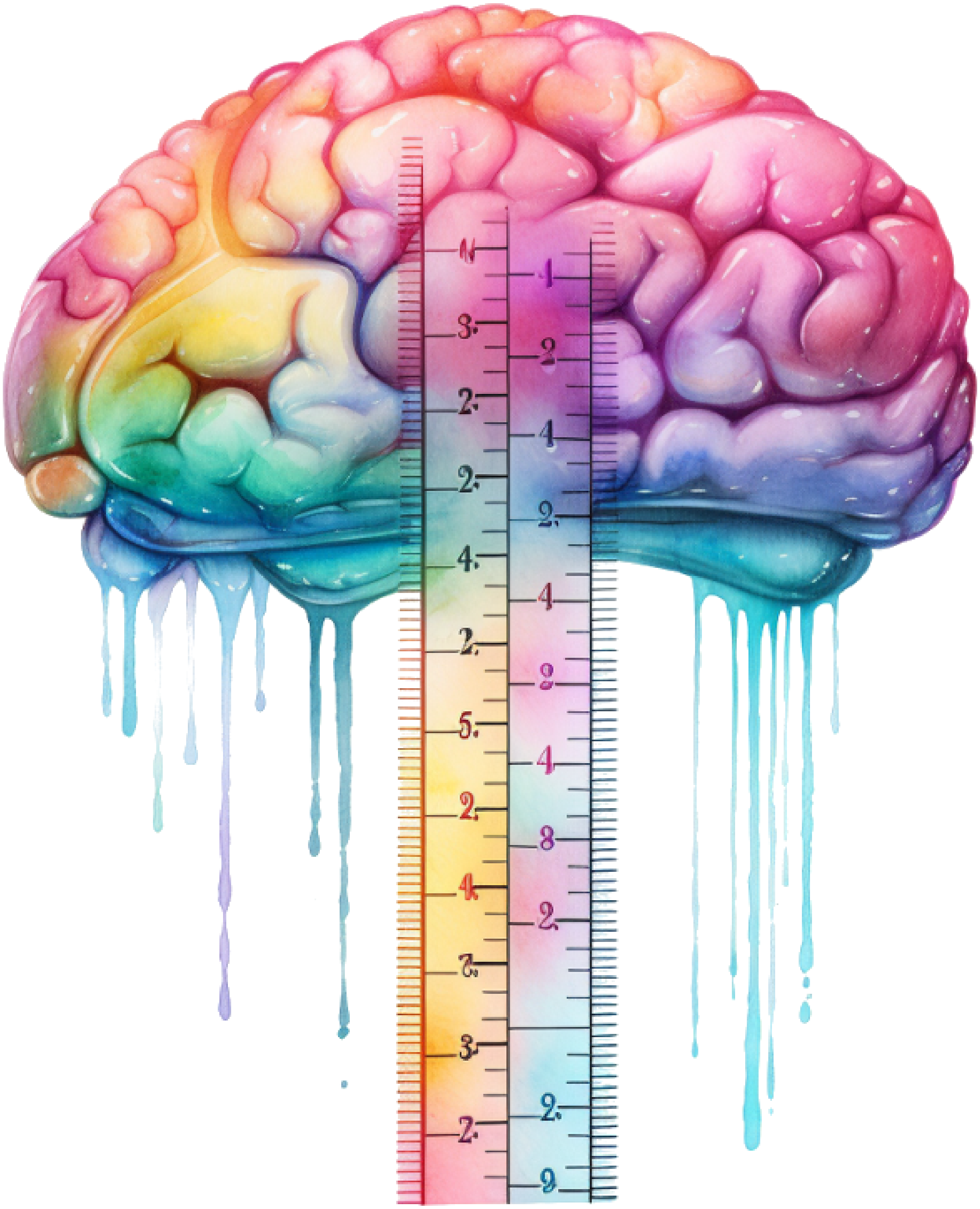


[1] [redacted]https://twitter.com/

Figure 5.1: Bildnachweis[1]

5.2 Skala und Skalenniveau

- Eine **Skala** bezeichnet in der Psychologie die systematische Zuordnung der Ausprägungen eines empirischen Merkmals (z.B. Depressivität, Konzentrationsfähigkeit, Stresslevel) zu Zahlenwerten oder Kategorien.
- Diese Zuordnung erfolgt entweder mit einem Messinstrument oder einer Messvorschrift
  - Messinstrumente: Test zur Bestimmung der Cortisolkonzentration im Speichel, Eyetracking-Gerät, EEG-Gerät, …
  - Beispiele Messvorschriften: Fragebogen, strukturiertes Interview, psychologischer Test
- Kurz gesagt: die Messung einer **Variablen** erfordert eine **Skala**
- Skalen werden in verschiedene **Skalenniveaus** unterteilt, deren Aussagekraft sich unterscheidet:
  - Nominalskala
  - Ordinalskala
  - Intervallskala
  - Verhältnisskala



5.3 Skalenniveaus

Table 5.1: **Kategoriale / qualitative Skalenniveaus**

Nominalskala	Verschiedene Ausprägungen des Merkmals entsprechen qualitativen (zumeist verbalen) Kategorien	Studiengang (Psychologie, Medizin, …), Geschlecht (männlich / weiblich)
Ordinalskala	Wie Nominalskala, aber die Kategorien lassen sich in eine natürliche Reihenfolge bringen; daher häufig mit Zahlen benannt, diese Zahlen stellen aber kein "Ausmaß" dar	Likert-Rating (z.B. Skala von 1 bis 10), Bildungsstand (Mittelschule, Abitur, Studium, Doktor, …)

Table 5.2: **Metrische / quantitative Skalenniveaus** (immer Zahlenwerte!)

Intervallskala	Zahlenwerte; Abstände zwischen zwei Werten sind interpretierbar	Schulnote (Note 1-6), IQ (96, 114, …)
Verhältnisskala	Wie Intervallskala, aber es gibt einen natürlichen Nullpunkt und daher können Verhältnisse gebildet werden (z.B. "doppelt so viel Einkommen in €")	Anzahl Freunde (2, 6, …), Körpergröße (173cm, 199cm, …)

Figure 5.2: Skalenniveaus werden manchmal mit den Etagen eines Hauses verglichen, wobei die Nominalskala die niedrigste Etage darstellt. In der Abbildung wird zusätzlich zwischen Intervall- und Absolutskala unterschieden. Im Gegensatz zur Verhältnisskala, bei der die Einheiten häufig frei gewählt werden können (Meter oder Zentimeter, Euro oder Cent, ..), basiert die Absolutskala auf der natürlichen Einheit "Anzahl" (Anzahl Personen, Anzahl richtiger Antworten, usw.). Davon abgesehen sind Verhältnis- und Absolutskala identisch und die genannte Unterscheidung ist ohne weitere praktische Relevanz.

5.4

Achtung bei Nominalskalen: zum Teil werden den Ausprägungen einer nominalskalierten Variable Zahlen zugeordnet (die sog. Kodierung – Beispiel Haarfarbe: "hellblond" = 1, "dunkelblond" = 2, "hellbraun" = 3, usw.). Diese Kodierung ändert aber nichts an den Eigenschaften der Variable, insbesondere dürfen nominalskalierte Variablen auch nach der Kodierung nicht miteinander verechnet werden (d.h. keine Addition, kein Verhältnis, kein Mittelwert, usw.).

5.5 Skalenniveaus: Zusammenfassung

Skalenniveau	Variablentypen	Messbare Eigenschaften			
Nominal	kategorial, dichotom	Häufigkeiten			
Ordinal	diskret	Häufigkeiten	Rangfolge		
Intervall	diskret, kontinuierlich	Häufigkeiten	Rangfolge	Abstand interpretierbar	
Verhältnis	diskret, kontinuierlich	Häufigkeiten	Rangfolge	Abstand interpretierbar	Natürlicher Nullpunk

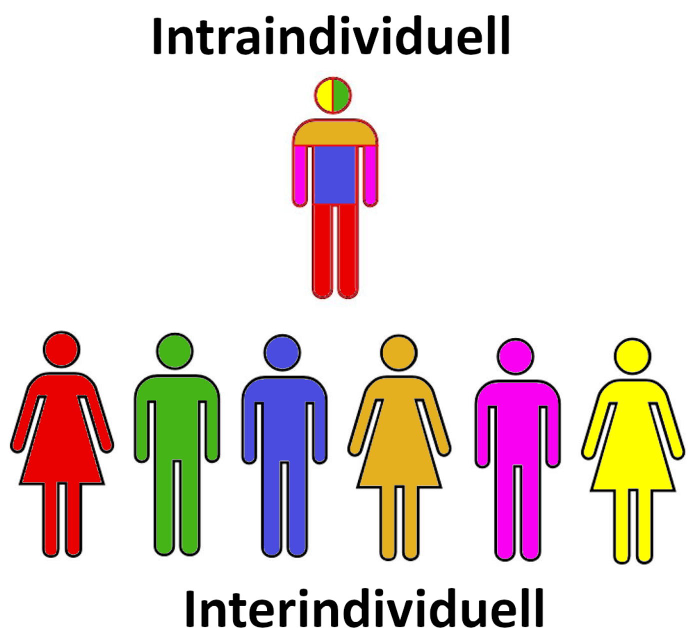
5.6 Was ist eine Variable?

- Die Grundannahme psychologischer Forschung ist, dass sich Erleben und Verhalten durch bestimmte Merkmale messen und beschreiben lässt
- Diese Merkmale sind in aller Regel *variabel*, sowohl innerhalb eines Individuums (intraindividuell), als auch zwischen Individuen (interindividuell)

Variables Merkmal = Variable

Beispiele

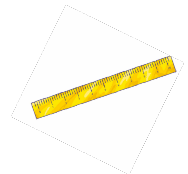
- Intraindividuelle Variable (unterschiedliche Werte bei einer Person über die Zeit): Müdigkeit einer Person über den Tagesverlauf; Hippocampusvolumen einer Person zu verschiedenen Jahreszeiten
- Interindividuelle Variable (unterschiedliche Werte bei verschiedenen Personen): Alter verschiedener Personen; Intelligenzquotient verschiedener Personen



5.7 Variablentypen

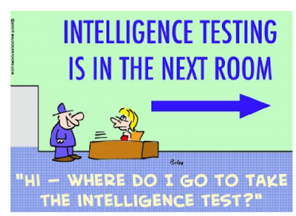
Variablen können verschiedene Arten von Ausprägungen aufweisen:

Typ	Quali/Quanti	Definition	Skala	Beispiele
Dichotom	Qualitativ (=Kategorien)	Zwei mögliche Ausprägungen	Nominal	schwanger/nicht schwanger, Raucher/Nichtraucher 
Kategorial	Qualitativ (=Kategorien)	Eine begrenzte Zahl von Ausprägungen	Nominal, Ordinal	Haarfarbe, Schulabschluss, Musikgeschmack 

Typ	Quali/Quanti	Definition	Skala	Beispiele	
Diskret	Quantitativ (=Zahlenwerte)	Gestufte Ausprägungen, die sich der Größe nach ordnen lassen	Ordinal, Intervall, Verhältnis	Anzahl von Geschwistern, Schulnoten	
Kontinuierlich	Quantitativ (=Zahlenwerte)	Stufenlose Ausprägung	Intervall, Verhältnis	Alter, Intelligenz	

NB: Die Art der Ausprägung einer Variable ist eine wichtige Information für die Auswahl einer geeigneten deskriptiven oder inferentiellen statistischen Methode

5.8 Manifeste und Latente Variablen

Manifeste Variable	Direkt messbar oder beobachtbar	Alter, Geschlecht, Herzfrequenz, Schulnoten, Einkommen	
Latente Variable	Nicht direkt messbar oder beobachtbar – die Ausprägungen der Variablen müssen erschlossen werden	Intelligenz, Einstellung gegenüber Minderheiten, Wohlbefinden	

In der Psychologie basieren viele Variablen auf theoretischen Konstrukten (wie Intelligenz) und können daher nicht direkt beobachtet, abgefragt oder gemessen werden .

Latente Variable sind daher eher die Regel als die Ausnahme.

5.9 Abhängige und unabhängige Variablen

In vielen Studien wird eine Variable in verschiedenen Studien- oder Experimentalbedingungen gemessen.

Um Studierende etwas zu verwirren ( ) werden auch die Studien- bzw. Experimentalbedingungen als Variablen bezeichnet, genauer als **unabhängige Variablen**. Sie sind unabhängig, weil sie i.d.R. von den Durchführenden einer Studie frei bestimmt werden.

Die tatsächlich gemessene Variable wird demgegenüber als **abhängige Variable** bezeichnet, da die Werte der Variable in Abhängigkeit von der jeweiligen Studien- btw. Experimentalbedingung unterschiedlich ausfallen können.

Table 5.6: Beispiele

Unabhängige Variable	Gruppenzugehörigkeit (z.B. Patienten vs. Kontrollen), Experimentalbedingungen (z.B. mit versus ohne Aufmerksamkeit, Messzeitpunkt 1 versus Messzeitpunkt 2)
Abhängige Variable	Herzrate, Fragebogenscore, Anzahl von Fehlern in einer Aufgabe


Untersuchen Sie doch einfach noch Bedingung Y und Z. Das sollte kein Problem sein, oder?
Prof.
Abhängigkeits-
verhältnis
Muss abhängige Variablen messen
Denkt sich unabhängige Variablen aus

6 Das Experiment

6.1 Was ist ein Experiment?

- Ein Experiment ist eine Überführung der Forschungsfrage in eine objektivierbare und quantifizierbare Untersuchung
- Im Zentrum des Experimentes steht die **unabhängige Variable**, durch die das Experiment in zwei oder mehr **experimentelle Bedingungen** unterteilt wird.

Definition

Ein **Experiment** ist eine **systematische Veränderung in einer unabhängigen Variable** (Treatment / Manipulation), mit dem Ziel, ursächlich eine **Veränderung in der abhängigen Variable** (d.h. der gemessenen Variable) herbeizuführen.

... was heißt das konkret?

6.2 Beispiel

Was schmeck besser: Coca-Cola oder Pepsi Cola?

Experiment: Blindverkostung durch N Versuchspersonen, die jeweils Coca-Cola und Pepsi Cola in einem neutralen Gefäß probieren und auf einer Skala von 1 bis 10 bewerten.

- **Unabhängige Variable:** Getränkemarke (zwei Werte: Cola, Pepsi; kategorial)
- **Abhängige Variable:** Bewertung durch die Versuchspersonen (10 Werte, 1-10; diskret)

Im genannten Beispiel handelt sich um ein Experiment mit **within-subject Design**, da jede Versuchsperson alle (hier zwei) Bedingungen der unabhängigen Variable durchläuft.

6.3

Wie sähe das Experiment in einem between-subject Design aus?

6.4 Within- und between-subject Design

Design	Definition	Beispiel
Within-subject	Alle Versuchspersonen durchlaufen alle Experimentalbedingungen	Jede VP probiert Cola und Pepsi
Between-subject	Verschiedene Versuchspersonen*gruppen* durchlaufen jeweils nur eine Experimentalbedingung	Gruppe 1 probiert Cola, Gruppe 2 probiert Pepsi
Mixed	Mindestens eine Experimentalbedingung wird innerhalb der Teilnehmer variiert (within-subject) und eine Experimentalbedingung zwischen den Teilnehmern variiert (between-subject)	Gruppe 1 probiert Cola mit und ohne Eis, Gruppe 2 probiert Pepsi mit und ohne Eis

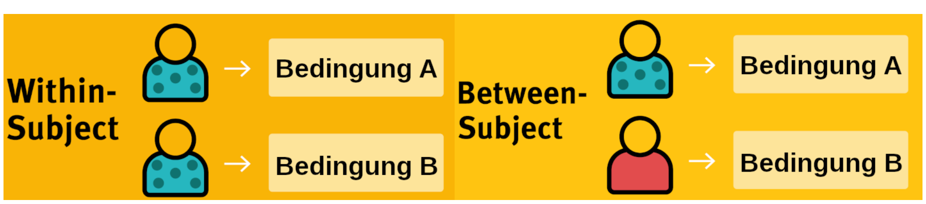


Figure 6.1: Bildnachweis[1]

[1] [redacted]https://www.nngrou

6.5 Within-subject & Messwiederholung (ein kleiner Rant)

Ein synonymer Begriff für das *within-subject Design* ist *Design mit Messwiederholung* (engl. repeated measures).

Messwiederholung bezeichnet, exakt wie das within-subject Design, die geplante Durchführung mehrerer Experimentalbedingungen an einer VP.

Messwiederholung ist ein recht unglücklich gewählter und verwirrender Begriff:

- Ursprünglich bezeichneter er vermutlich allein die Messung einer VP zu verschiedenen Zeitpunkten (z.B. Test vor und nach einer Intervention). Hier ergibt der Begriff "Wiederholung" Sinn. In der heutigen Anwendung bezieht sich der Begriff Messwiederholung allerdings nicht nur auf den within-subject Faktor "Zeit" (mit verschiedenen Messzeitpunkten als within-subject Experimentalbedingungen), sondern *auf jede Art* von within-subject Faktor (z.B. mit und ohne Eis im Cola-Beispiel). "Wiederholung" verwirrt in diesem Zusammenhang, da unterschiedliche Experimentalbedingungen getestet werden (wenn auch in derselben Person).
- Experimente beinhalten häufig eine tatsächliche Art von Messwiederholung, nämlich die wiederholte Durchführung ein und derselben Experimentalbedingung (auch "Trial" genannt), um durch Mittelwertbildung die Präzision der Messung zu erhöhen. Hier wäre der Begriff Messwiederholung besser angebracht.
- Aus dem Begriff selbst wird nicht deutlich, dass eine Messung tatsächlich innerhalb derselben VP wiederholt wird – der Begriff *within-subject* ist hier präziser.

Der Ausdruck *Messwiederholung* oder *Repeated Measures* ist allerdings so verbreitet (in wissenschaftlichen Artikeln, Statistik-Software usw.), dass er hier Erwähnung finden muss.

6.6 Beispiel between-subject Design: das Asch-Experiment

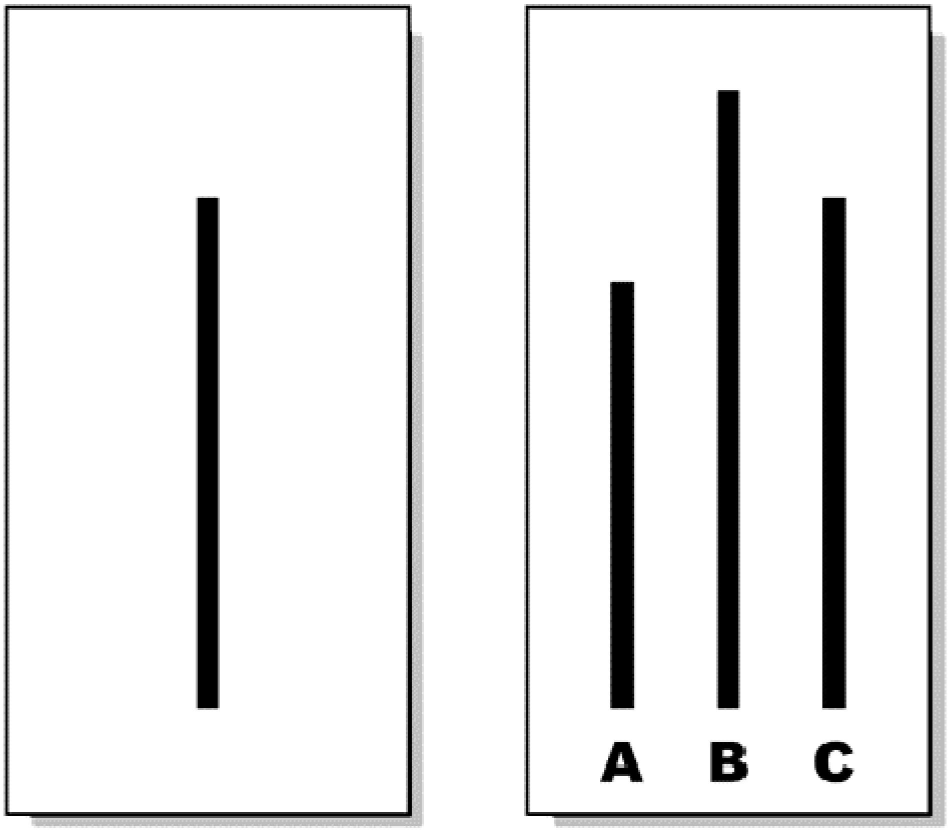
- Konstrukt: gruppenkonformes Verhalten
- Experiment: die Versuchsperson (VP) musste auswählen, welche von drei Vergleichlinien dieselbe Länge wie eine Referenzlinie hat. Außer der VP waren noch weitere vorgebliche VPen anwesend (in Wirklichkeit Vertraute der Experimentalleitung – die "Konfidenten").
  - **Unabhängige Variable:** zwei Gruppen (Gruppe A: Konfidenten geben richtige Antworten; Gruppe B: Konfidenten geben in 2/3 der Fälle eine falsche Antwort)
  - **Abhängige Variable:** Prozentzahl der Fehler der VP

Ergebnis:

- 1% Fehlentscheidungen in Kontrollgruppe
- 37% Fehlentscheidungen in der Experimentalgruppe

Die Operationalsierung des Konstruktes "gruppenkonformes Verhalten" entspricht hier also einem experimentellen Kontrast:

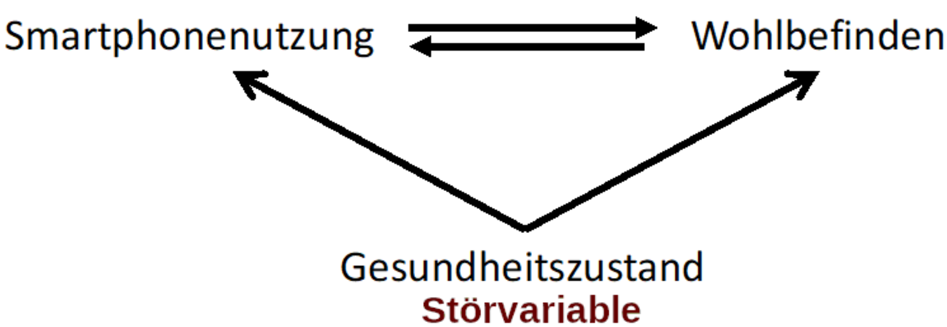
Fehlerrate Experimentalgruppe $-$ Fehlerrate Kontrollgruppe



6.7 Ist jede wissenschaftliche Studie ein Experiment?

Nein.

- In **Beobachtungsstudien** werden Phänomene oder Zusammenhänge *ohne Manipulation einer unabhängigen Variable* untersucht.
- Beispiele: Quasi-Experiment, Längsschnittstudien, Fall-Kontroll-Studie
- Im Gegensatz zum Experiment(*) können in Beobachtungsstudien **keine kausalen Schlüsse** gezogen werden
- Warum? Jeder Unterschied zwischen Gruppen oder Zusammenhang zweier Variablen zwischen Personen kann immer auch durch (bekannte oder unbekannte) Störvariablen verursacht sein.



→ Mehr dazu im Modul Forschungsmethoden!

(*)Angenommen wird hier ein “echtes Experiment”, d.h. mit randomisierter Zuordnung einer Stichprobe zu Experimentalbedingungen (das sog. **randomisiert-kontrollierte Experiment**)

6.8

Ok, Task Force, zurück zur ersten Hypothese:

Hypothese 1: Paradoxia wird durch eine virale Infektion verursacht. *Die Blutproben von Menschen mit Paradoxia sollten daher eine höhere Menge von Entzündungsmarkern aufweisen.*

Warum wäre zur Überprüfung dieser Hypothese ein echtes Experiment ethisch schwierig? Welche Probleme weist eine reine Beobachtungsstudie (Vergleich der Entzündungsmarker von Paradoxikern und Kontrollen) auf?

7 Darstellung von Daten

7.1 Darstellung kategorialer Daten: die Häufigkeitstabelle

- **Reminder:** Kategoriale Variable = begrenzte Zahl von Ausprägungen; keine natürliche Reihenfolge

Table 7.1: Häufigkeitsverteilung von Haarfarben unter Paradoxikern (N=800)

Haarfarbe	Absolute Häufigkeit (n)	Relative Häufigkeit (%)
Blau (gefärbt)	440	55 %
Grün (gefärbt)	152	19 %
Braun (natürlich)	104	13 %
Blond (natürlich)	64	8 %
Schwarz (natürlich)	40	5 %

Figure 7.1: Bildnachweis[1]

- **Absolute Häufigkeiten:** geben an, wie oft die jeweiligen Ausprägungen einer Variable vorkommen (Anzahl)
- **Relative Häufigkeiten:** geben an, wie häufig die jeweiligen Ausprägungen in Bezug zu allen Fällen vorkommen (Prozent oder Anteil)

[1] [redacted]https://www.holleew

7.2 Darstellung kategorialer Daten: die Häufigkeitstabelle

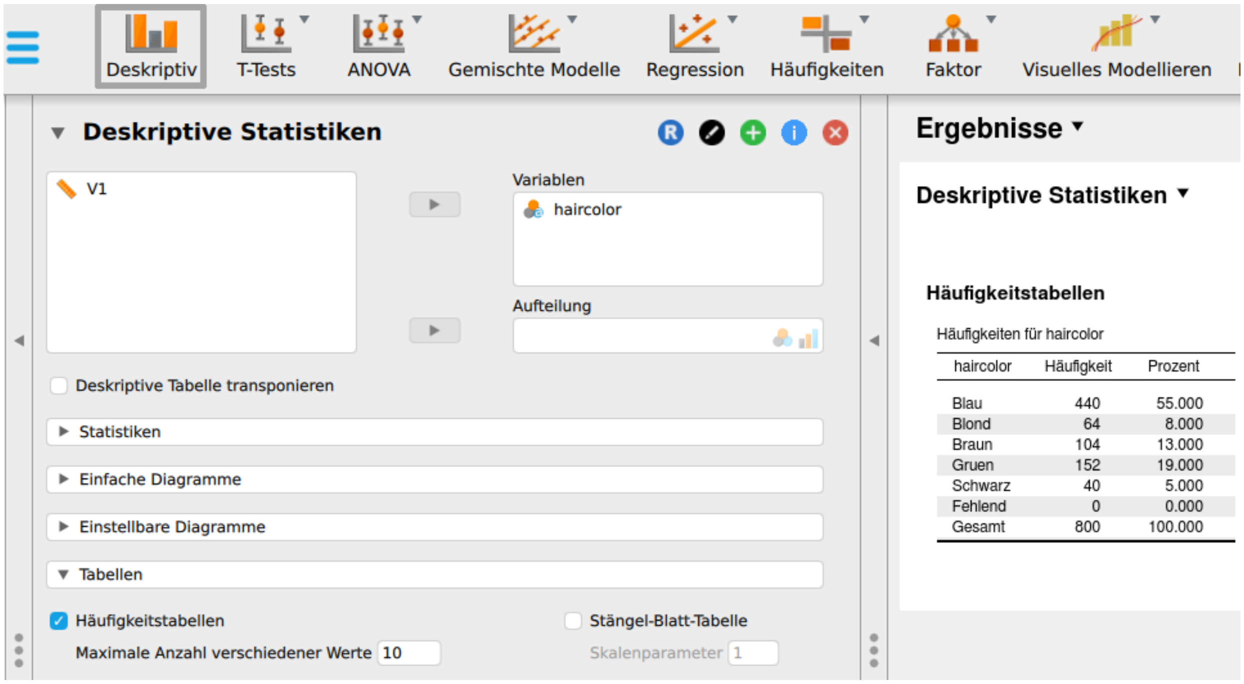
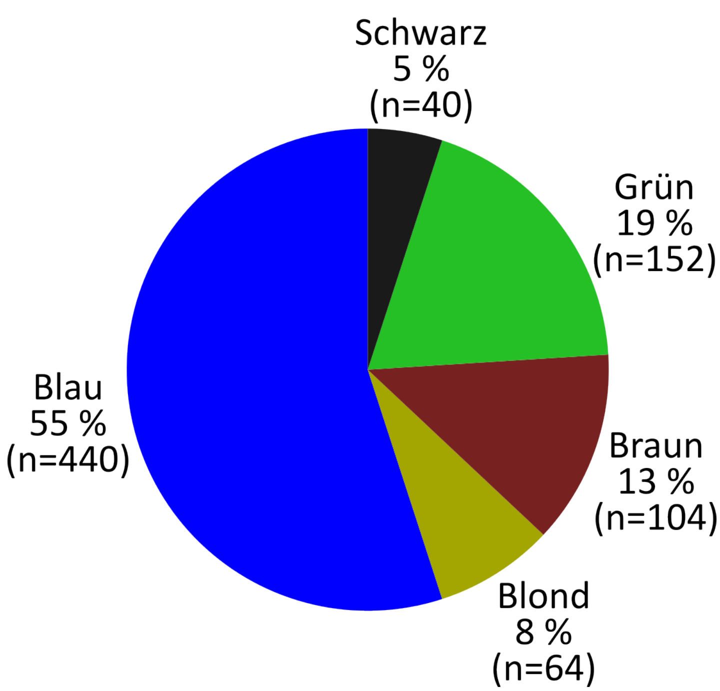


Figure 7.2: Darstellung von Häufigkeitstabellen in JASP

7.3 Darstellung kategorialer Daten: Kreisdiagramme

- Vorteil: übersichtliche Darstellung der Verteilung kategorialer Variablen
- Nachteil: verschiedene Bedingungen können nicht direkt verglichen werden (zwei Kreisdiagramme notwendig)



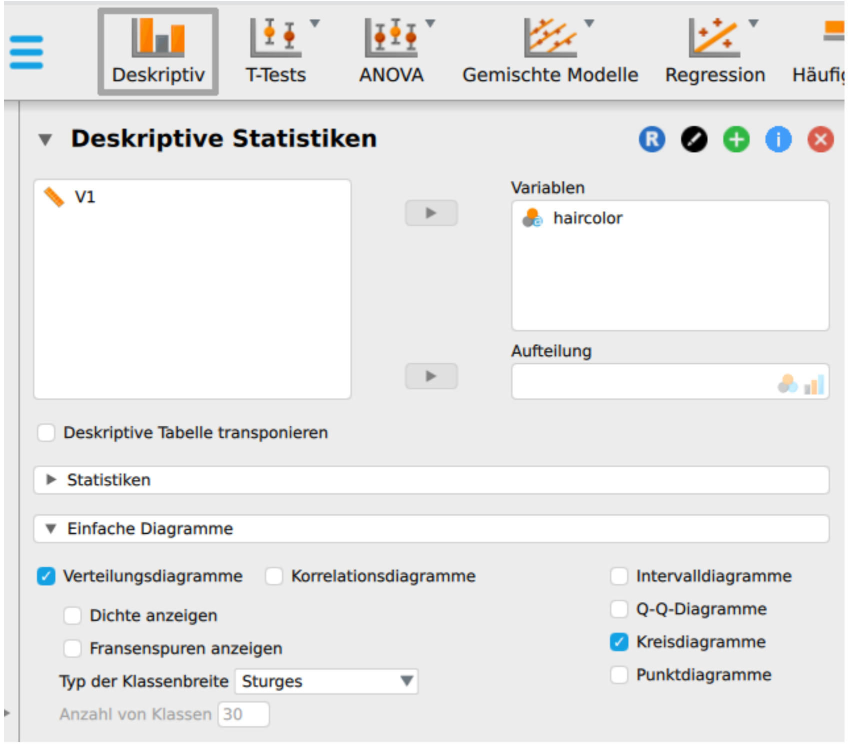


Figure 7.3: Kreisdiagramm in JASP

7.4 Darstellung kategorialer Daten: Häufigkeitsdiagramm

- Das Häufigkeitsdiagramm erlaubt einen besseren Vergleich von zwei Gruppen oder Bedingungen

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

fontsize = 12
order = ['Schwarz', 'Blond', 'Braun', 'Gruen', 'Blau']

df = pd.read_csv('data/haircolor_raw.csv')
count0 = df[df.group == 0].groupby('haircolor').id.count().reindex(order)
count1 = df[df.group == 1].groupby('haircolor').id.count().reindex(order)
plt.barh(np.arange(0, 12.5, 2.5), count0.values, label='Paradoxia')
plt.barh(np.arange(1, 13.5, 2.5), count1.values, label='Kontroll')
plt.xticks(fontsize=fontsize)
plt.yticks(np.arange(0.5, 12.5, 2.5), count0.index, fontsize=fontsize)
plt.xlabel('Anzahl', fontsize=fontsize)
plt.legend(fontsize=fontsize)
plt.show()
```

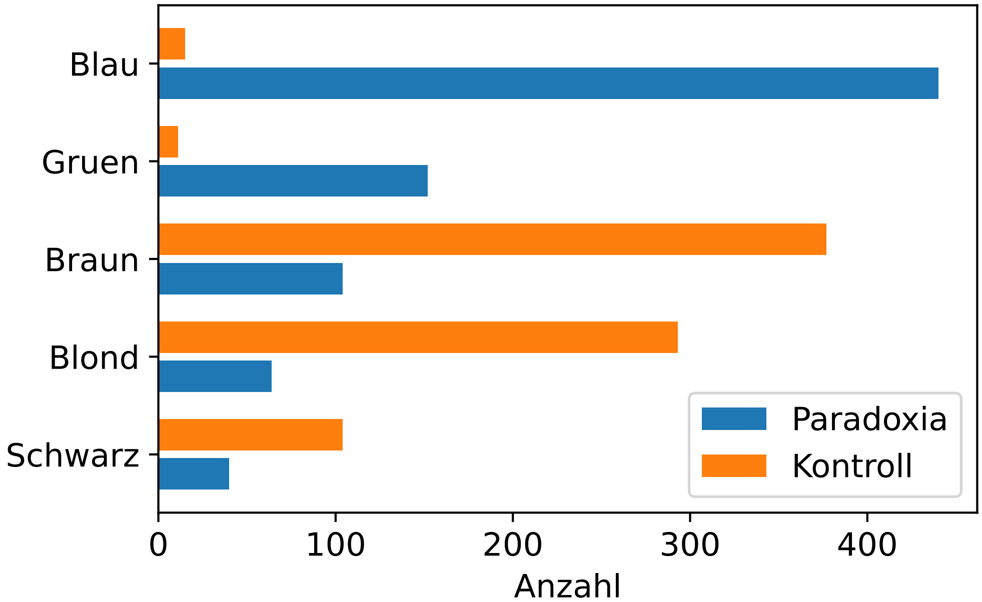


Figure 7.4: Verteilung der Haarfarbe bei Paradoxikern und Kontrollen in Form eines Häufigkeitsdiagramms. Die Form der Darstellung wird auch als Balken- oder Säulendiagramm bezeichnet.

7.5 Darstellung quantitativer Daten: das Histogramm

- Erinnerung: quantitative Variablen sind Zahlenwerte (diskret oder kontinuierlich)
- Beispiele: Körpergröße (z.B. in *cm*), durchschnittliches Zuspätkommen einer Kohorte in Statistik 1 (z.B. in *sec*), emotionaler Intelligenzquotient (einheitslos)

Die wichtigste Darstellungsform quantitativer Daten ist das Histogramm:

```
setup:
  - element: "#image1"
    modifier: attr
    parameters:
      - class: fragment
        data-fragment-index: "3"
  - element: "#binsize_text"
    modifier: attr
    parameters:
      - class: fragment
        data-fragment-index: "4"
  - element: "#binsize_bracket"
    modifier: attr
    parameters:
      - class: fragment
        data-fragment-index: "4"
  - element: "#totalbg"
    modifier: attr
```

```yaml
      parameters:
        - class: fragment
          data-fragment-index: "5"
    - element: "#total"
      modifier: attr
      parameters:
        - class: fragment
          data-fragment-index: "5"
```

7.6

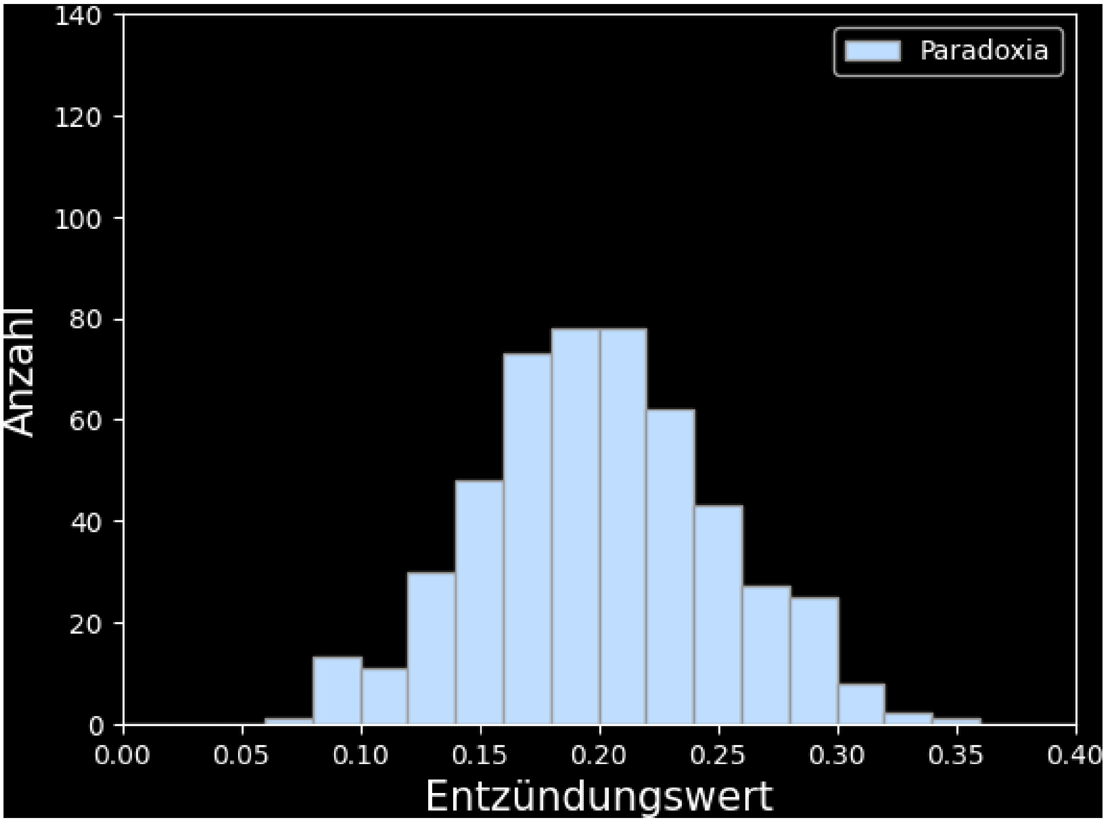
Die Daten der ersten Beobachtungsstudie zu Paradoxia sind frisch eingetroffen!

Table 7.2: Table 1. Results.

id	group	inflammatory value (0-1)
1	control	0.12
2	control	0.03
...	...	...
500	control	0.19
501	paradoxia	0.28
502	paradoxia	0.15
...	...	...
1000	paradoxia	0.19

7.7

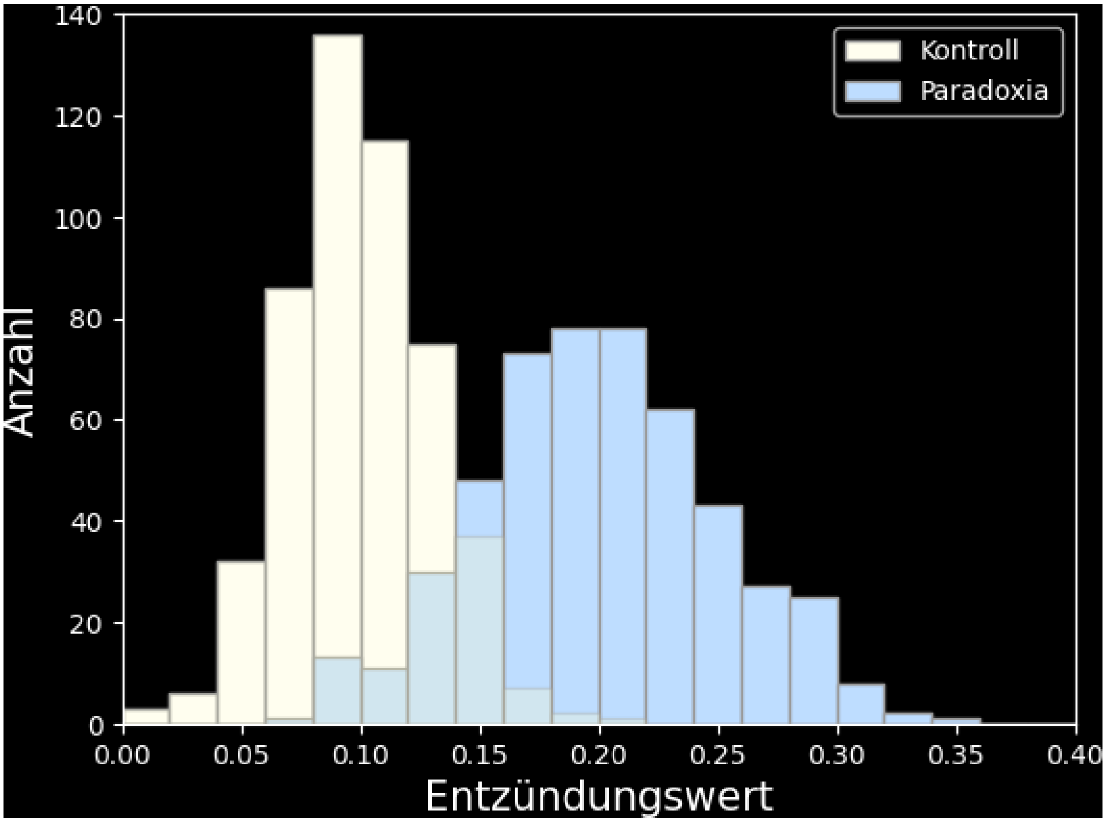
Hier ist das Histogramm der Entzündungswerte der Paradoxiker:



Was ist die Intervallgröße? Überschlagen Sie: passt das Histogramm zur angegeben Stichprobe von N=500 Paradoxikern? Was schätzen Sie ist der Mittelwert der Entzündungsmarker?

7.8

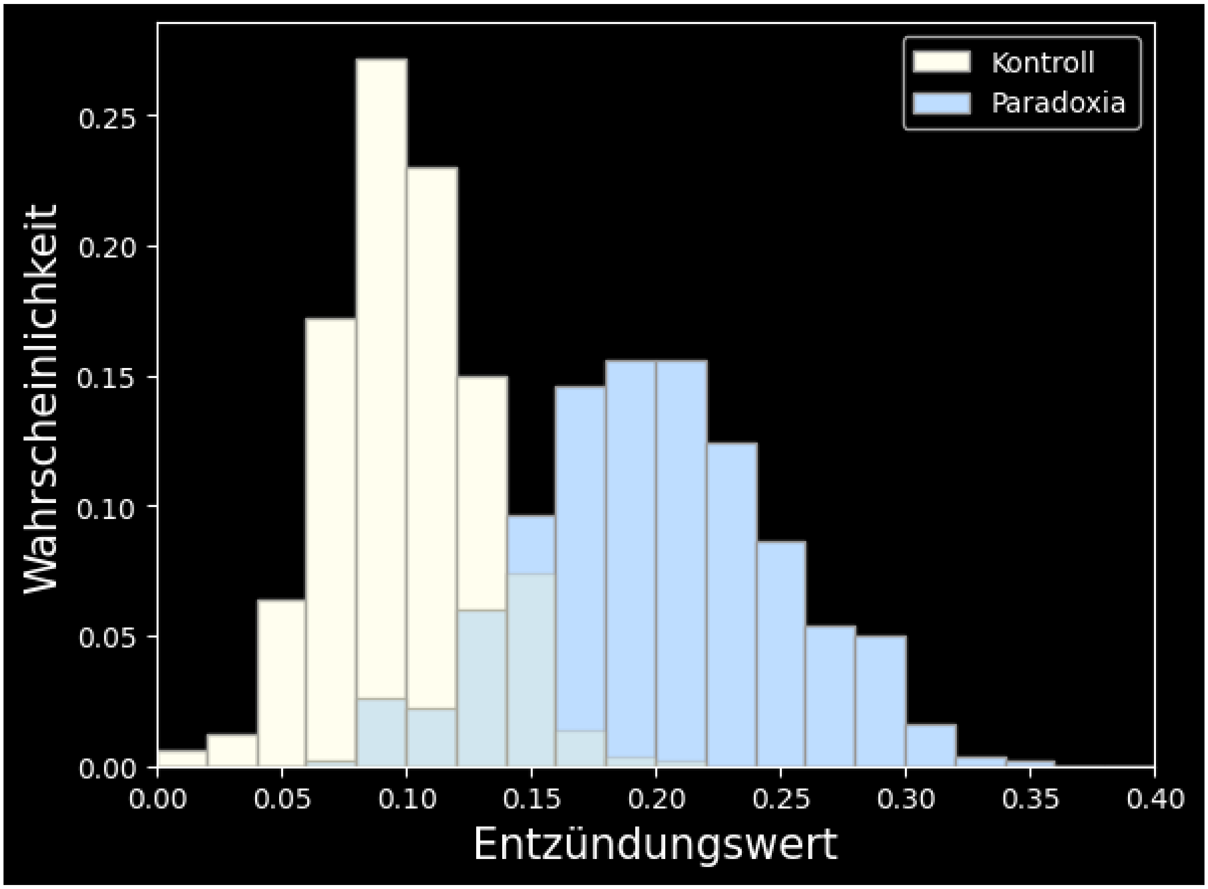
Und hier der Vergleich mit der Kontrollgruppe:



Diese erste Studie erbringt also tatsächlich Evidenz für einen erhöhten Entzüdungswert bei Paradoxikern, und favorisiert damit die Hypothese 1 der Task Force!

7.9

Statt der Anzahl (absolute Häufigkeit) kann auch die Wahrscheinlichkeit (relative Häufigkeit) dargestellt werden:



Jeder Wert in dieser Abbildung gibt also die Wahrscheinlichkeit an, dass ein Entzündungswert im Intervall des jeweiligen Balkens liegt.

Während sich die Balken eines Histogramms mit absoluter Häufigkeit (Anzahl) zur Stichprobengröße aufaddieren, addieren sie sich beim Histogramm mit relativer Häufigkeit (Wahrscheinlichkeit) zu 1!

7.10 Grundgesamtheit und Stichprobe

- **Grundgesamtheit** (auch **Population**) bezeichnet die Gesamtheit der Personen, für die wir uns in einer wissenschaftlichen Studie interessieren (z.B. Personen mit einer bestimmten Erkrankung, Kinder zwischen 3 und 6, die gesamte Menschheit)
- In den meisten Fällen ist es nicht möglich alle Mitglieder der Grundgesamtheit zu untersuchen (Vollerhebung) – wir müssen daher eine **Stichprobe** ziehen.
- Die Annahme ist, dass die Stichprobe ein möglichst **repräsentatives Abbild der Grundgesamtheit** ist und sich Ergebnisse in der Stichprobe auf die Grundgesamtheit übertragen lassen.

7.11 Zufallsvariable

- Eine **Zufallsvariable** ist eine Variable, die zu einem gewissen Grad vom Zufall (oder uns zufällig erscheinenden Faktoren) abhängt.

[2] [redacted] https://datatab.de/t

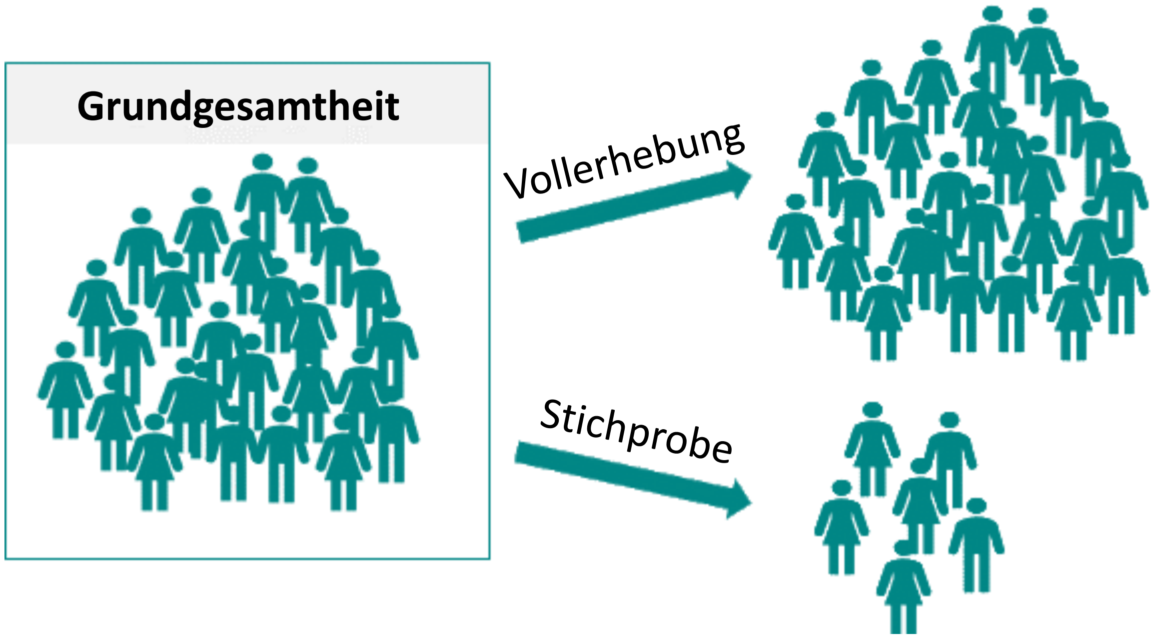


Figure 7.5: Bildnachweis[2]

- In der Psychologie kommt "Zufall" an zwei Stellen ins Spiel:
  - **Messfehler** (unbekannte, uns zufällig erscheinende, Einflüsse auf die Genauigkeit der Messung)
  - Beim **zufälligen Ziehen einer Stichprobe** aus der Grundgesamtheit
- Ein wichtiger Aspekt von Zufallsvariablen ist, dass die Zufälligkeit mathematisch beschrieben werden kann – durch **Wahrscheinlichkeitsverteilungen** (wir kommen darauf später noch einmal zurück)
  - Beispiel: jeder Wurf eines Würfels ist zufällig, aber wir können sagen, dass die Wahrscheinlichkeit jeder Zahl exakt 1/6 ist.

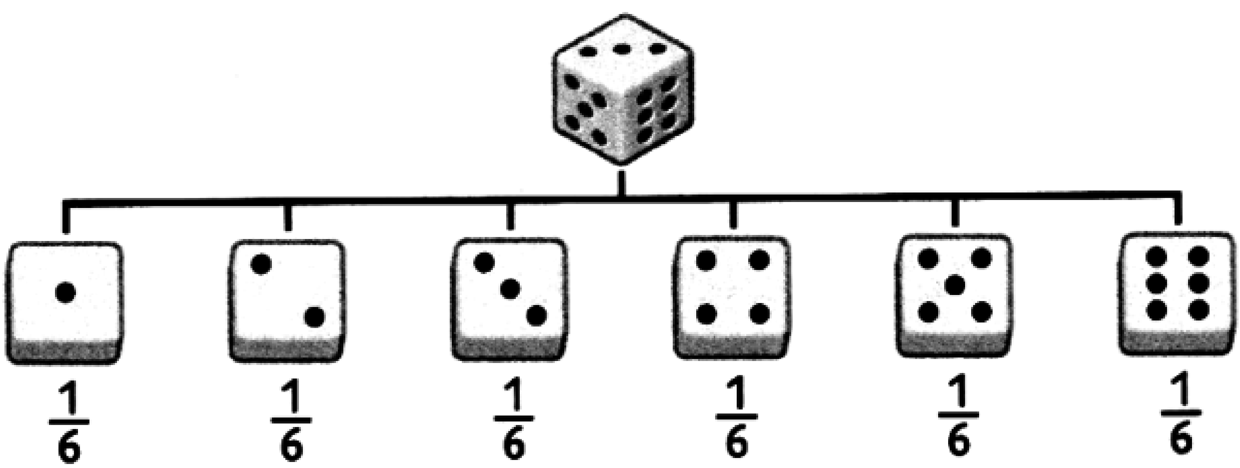


Figure 7.6: Bildnachweis[3]

[3] [redacted]https://www.studien

7.12 Zufallsvariable

- Praktisch gesehen, sind nahezu alle Variablen in der Psychologie Zufallsvariablen, daher meint "Variable" in aller Regel "Zufallsvariable"
- Die Ausnahme der Regel wäre der seltene Fall, dass eine Variable *für alle Personen der Grundgesamtkeit bekannt* ist und ihre Messung *keinem Messfehler* unterliegt (z.B. Geburtsjahr aller amerikanischen Präsidenten bis zum Jahr 2023).
- Zufallsvariablen werden mit großen Lettern bezeichnet, i.d.R. X (oder Y/Z falls zwei/drei Zufallsvariablen betrachtet werden).
- Einzelne Beobachtungen einer Zufallsvariable (z.B. der gemessene IQ von *bestimmten Personen*) werden mit kleinen Lettern bezeichnet (x_i, y_i, z_i), wobei i der Durchnummerierung dient.
- Der Vektor $\mathbf{x}$ (beachte die Fettschrift) bezeichnet die Gesamtheit der Beobachtungen $x_1..x_N$

Table 7.3: Beispiel: die Zufallsvariable X bezeichnet in der Tabelle den in einer Stichprobe gemessenen Intelligenzquotienten

	X
x_1	99
x_2	112
x_3	104
...	...

Part III

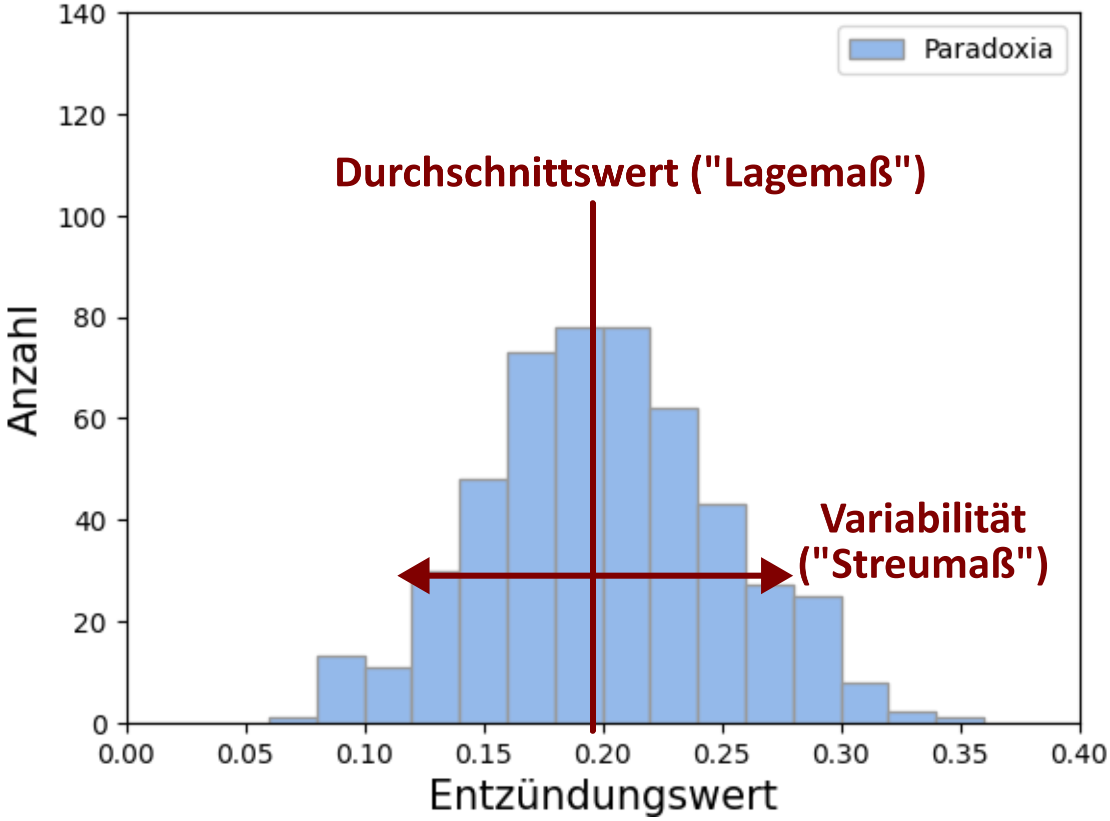
Vorlesung 03: Lage- und Streuungsmaße

8 Vorlesung 03: Lage- und Streuungsmaße

9 Lagemaße

9.1

Wie würden Sie die Daten in folgender Abbildung mit möglichst wenigen Worten beschreiben?



9.2 Der Forschungsprozess

```
setup:
    - element: "#results"
      modifier: function() { this.node.style.fill = 'green'; }
    - element: "#resultsbg"
      modifier: function() { this.node.style.fill = '#d8ffe2';}
```

9.3 Lagemaße

- Der weithin bekannte **Durchschnittswert** oder **Mittelwert** ist ein Beispiel für ein **Lagemaß** von Verteilungen.
- Man spricht dabei auch von der **zentralen Tendenz** (engl. *central tendency*) einer Verteilung
- Das Lagemaß ist neben dem **Streumaß** der wesentliche Parameter, um die Verteilung von Daten effizient (d.h. mit wenigen Parametern) zu beschreiben
- Die drei wichtigsten Lagemaße sind:
  - Mittelwert (gemeint ist das *arithmetische Mittel*)
  - Median
  - Modus (auch Modalwert)

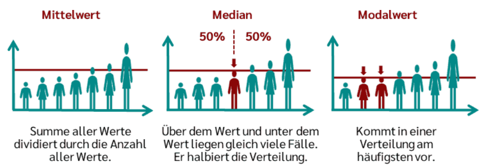


Figure 9.1: Bildnachweis[1]

9.4 Mittelwert

- Berechnung (verbal):
  1. Addiere alle Werte
  2. Teile durch die Anzahl der Werte
- Berechnung (Formel):

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

- Notation: Strich über dem kleinen Letter der Zufallsvariable = Mittelwert der Zufallsvariable

Beispiel

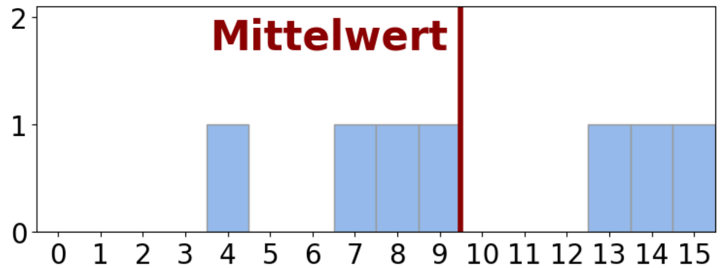
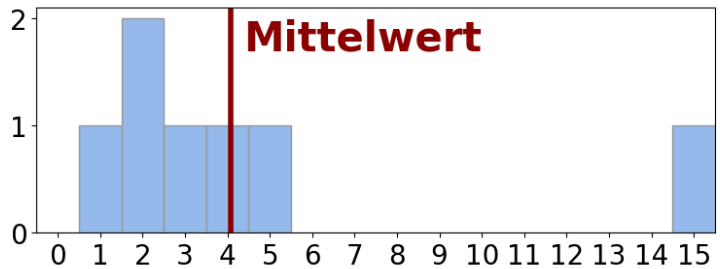
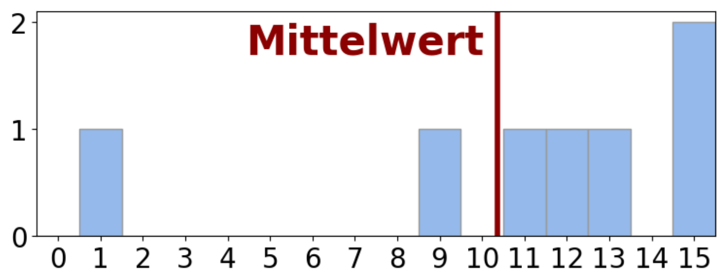
Folgende Beobachtungen der Zufallsvariable X "Punktzahl in der Abi-Matheprüfung" werden in einer Stichprobe von 7 Psychologiestudierenden gemacht: $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14\}$

$$\bar{x} = \frac{1}{7}(13 + 7 + 15 + 8 + 4 + 9 + 14) == \frac{1}{7} \cdot 70 = 10$$

[1] [redacted]https://datatab.de/t

9.5 Wann ist der Mittelwert sinnvoll?

- Der Mittelwert ist ein sinnvolles Lagemaß, wenn er nicht durch einzelne **Ausreißer** (extreme Werte) dominiert bzw. verzerrt wird.

$\mathbf{x}$ (z.B. Punktzahlen im Abi)	$\bar{X}$	Histogramm	Mittelwert sinnvoll?
$\{13, 7, 15, 8, 4, 9, 14\}$	10		
$\{3, 1, 4, 2, 2, 6, 15\}$	4.7		
$\{13, 15, 11, 12, 9, 14, 1\}$	10.7		

9.6 Median

- Gibt es dominante Ausreißer in den Daten, so ist häufig der **Median** das sinnvollere Lagemaß
- Berechnung:

  1. Alle N Werte der Größe nach aufreihen
  2. Der Wert, der genau in der Mitte liegt, ist der Median
  3. Bei einer geraden Anzahl von Werten bilden zwei Werte die Mitte – in diesem Fall ist der Median der Mittelwert dieser beiden Werte
  4. $\left(\text{Alternativ mit Formel: } Tiefe_{\text{Median}} = \frac{N+1}{2}\right)$

Beispiel

Folgende Beobachtungen der Zufallsvariable X "Punktzahl in der Abi-Matheprüfung" werden in einer Stichprobe von 7 Psychologiestudierenden gemacht: $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14\}$

$$\text{Sortierte Reihenfolge: } \mathbf{x} = \{4, 7, 8, \mathbf{9}, 13, 14, 15\} \rightarrow \text{ Median} = \tilde{x} = 9$$

Wir fügen den Wert eines weiteren Studierenden hinzu: $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14, 10\}$

$$\text{Sortierte Reihenfolge: } \mathbf{x} = \{4, 7, 8, \mathbf{9}, \mathbf{10}, 13, 14, 15\} \rightarrow \text{ Median} = \tilde{x} = \frac{9 + 10}{2} = 9.5$$

9.7 Median

- Der **Median ist außerdem sinnvoll** bei:
  - ordinalen Daten wie etwa diskreten Ratings (Skala von 1 bis 10), bei denen die Abstände zwischen Zahlen nicht interpretierbar sind.
  - schiefen Verteilungen der Daten (dazu kommen wir noch)

- Der **Median ist nicht sinnvoll**, wenn:
  - auch der Mittelwert ein sinnvolles Lagemaß darstellt (der Mittelwert hat einige hilfreiche mathematische Eigenschaften[2]).
  - in der zugrundeliegenden Verteilung zwei oder mehr Wertebereiche deutlich häufiger vorkommen als andere Wertebereiche, und diese Bereiche nicht überlappen (z.B. hat der Median von $\{2, 2, 2, 2, 9, 9, 9\}$ den wenig aussagekräftigen Wert 2).

9.8 Modus

- In manchen Fällen ist es interessant zu wissen, was der **häufigste Wert** in einem Datensatz ist – dies ist der **Modus**.
- Berechnung:
  1. Zähle die Häufigkeit aller vorkommenden Werte
  2. Der häufigste Wert ist der Modus
- Im Fall von **kategorialen Variablen** ist der Modus das einzig mögliche Lagemaß (Beispiel: aus welchem Bundesland kommen die meisten von Ihnen?)
- Ein weiterer sinnvoller Anwendungsfall können schiefe Verteilungen sein:

[2] [redacted]Beispielsweise Datenpunkte. Beim Median ist dies nicht gegeben. Auch beziehen sich viele statistische Standardtests auf den Mittelwert und nicht den Median.

[3] [redacted]https://youtu.be/in

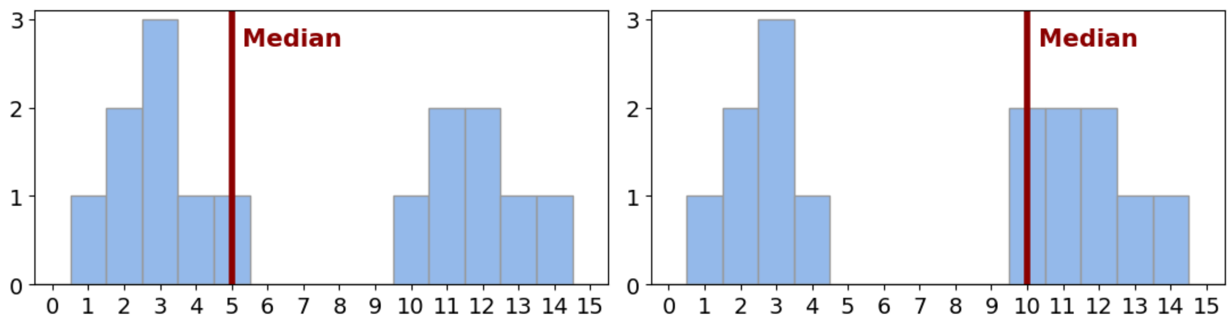


Figure 9.2: Für diese "bimodale" Verteilung der Daten ist der Median eine schlechte Wahl. Der Wert springt mehr oder weniger zufällig zwischen den beiden dominanten Wertebereichen der Daten (1-5 und 10-14), je nachdem, welcher Bereich mehr Werte aufweist.

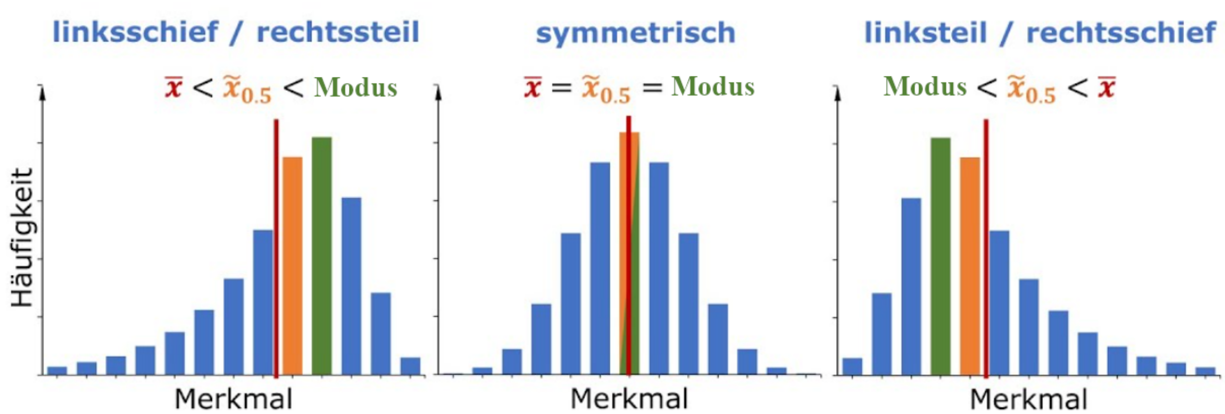
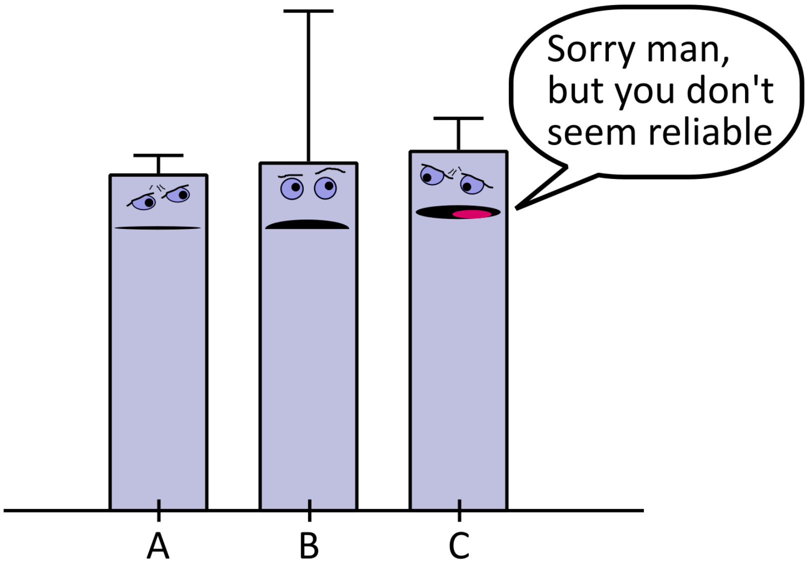


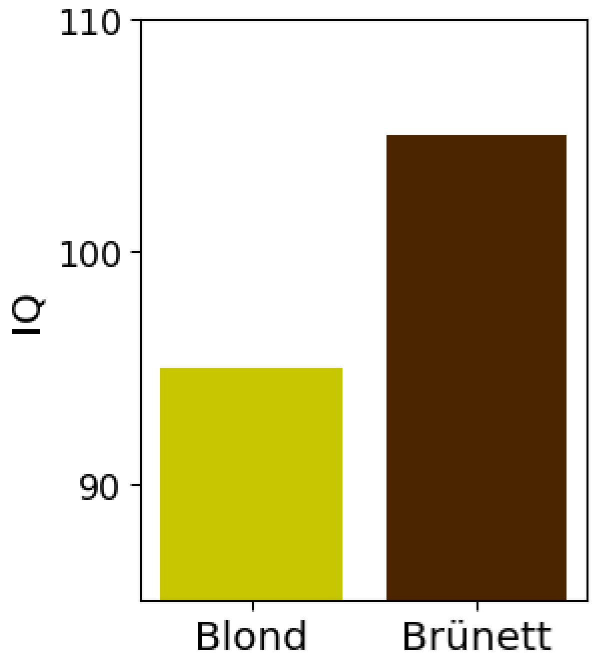
Figure 9.3: Bildnachweis[3]

10 Streuungsmaße

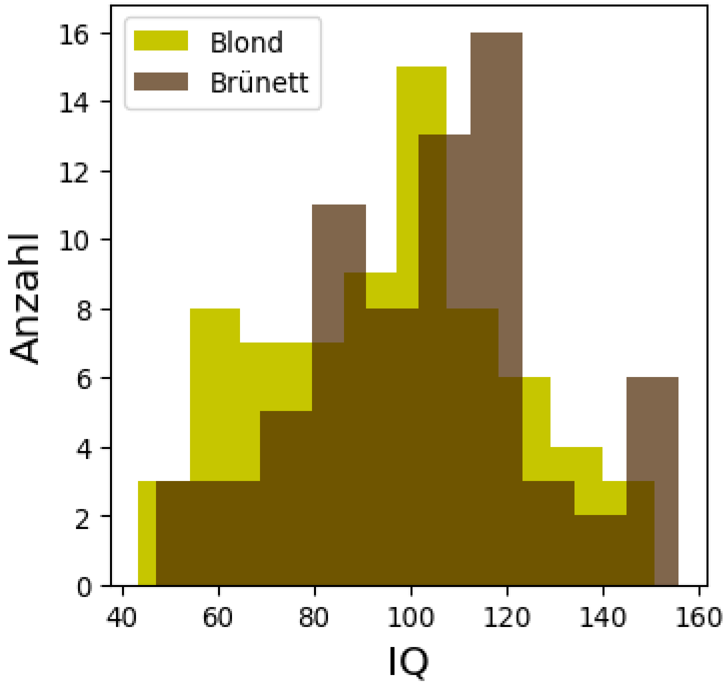


10.1 Warum sind Streuungsmaße wichtig?

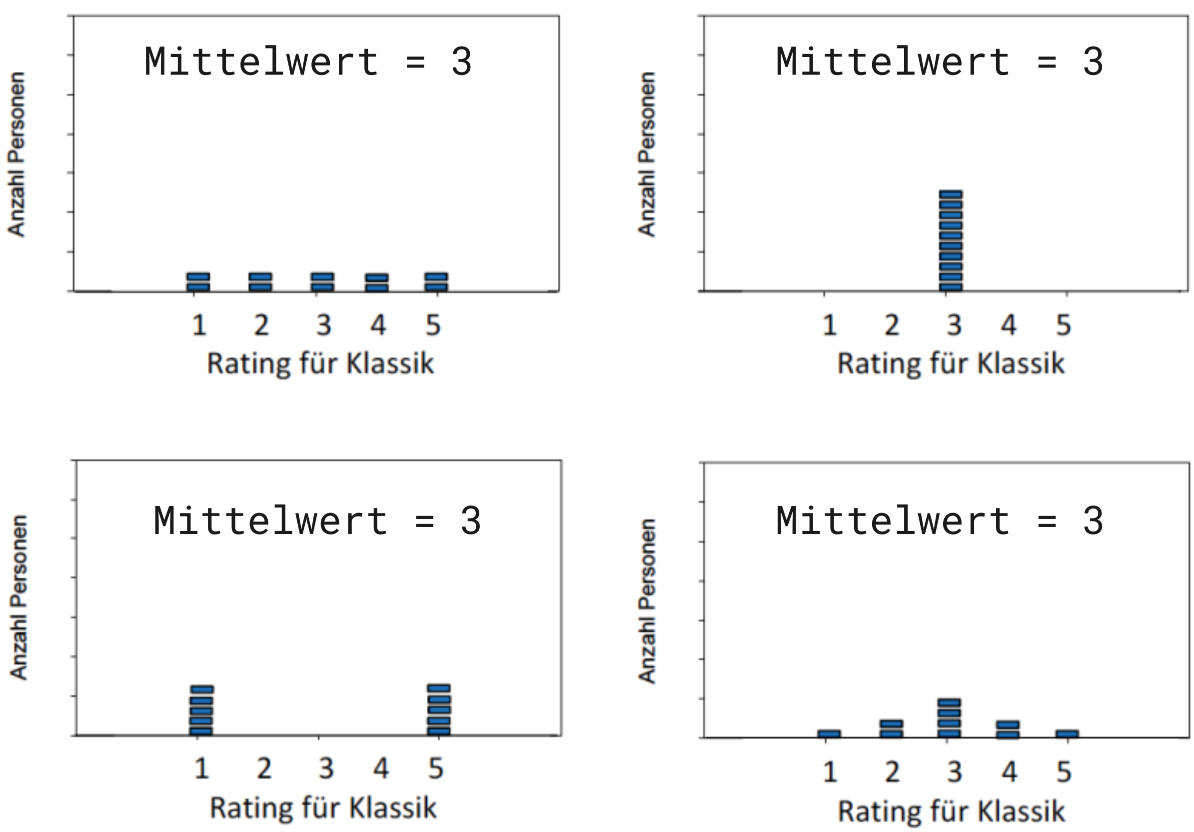
"In unserer Studie waren brünette Menschen im Schnitt 10 IQ-Punkte schlauer als blonde Menschen"



Behind the scenes:

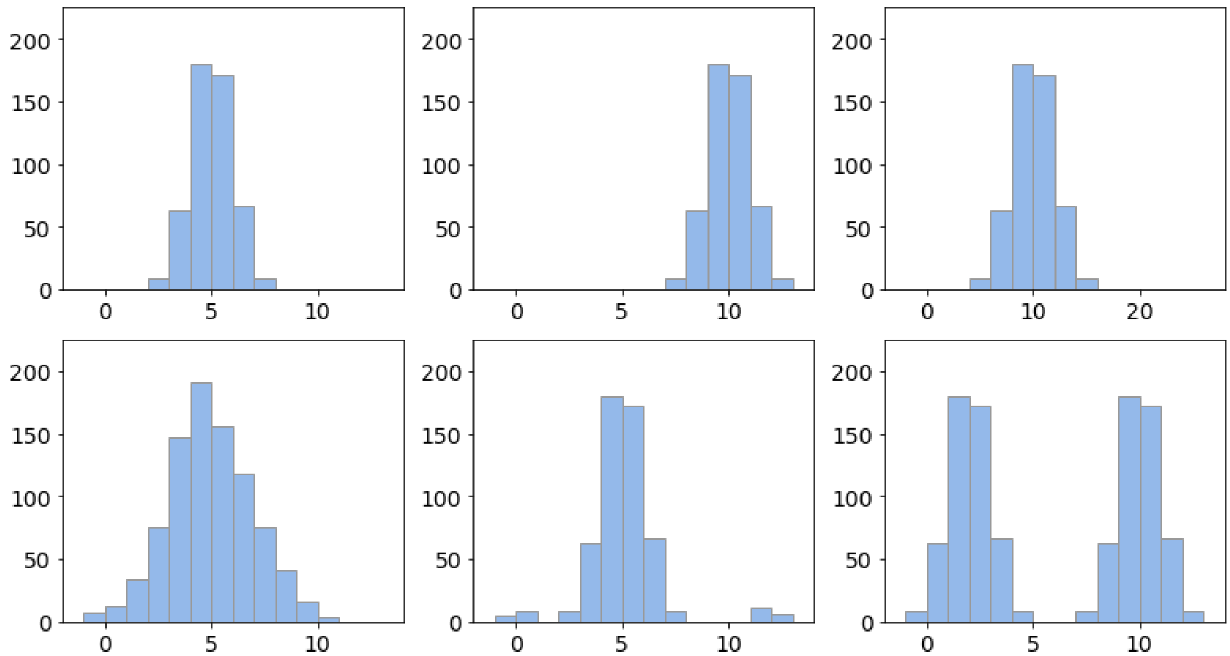


10.2 Warum sind Streuungsmaße wichtig?



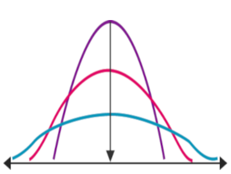
10.3

Wie schätzen Sie die Streuung / Variabilität folgender Verteilungen ein?



10.4 Streuungsmaße

- **Streuungsmaße** geben die Variabilität von Daten anann entweder
- Lagemaße sollten durch die Angabe eines Streuungsmaßes ergänzt werden
- Streuung von Daten kann ein Ausdruck von *echter Variabilität* in der Stichprobe sein oder eine Folge der *Messungenauigkeit* (häufig beides)
- Je nach Skalenniveau und Zweck können verschiedene Streuungsmaße bestimmt werden:
  - Spannweite (Range)
  - Interquartilsabstand
  - Varianz
  - Standardabweichung



10.5 Spannweite / Range

- Die **Spannweite** oder **Range** ist die Differenz zwischen dem kleinsten und dem größten Wert:

$$Range = x_{max} - x_{min}$$

- Einfachstes **Streuungsmaß** — sinnvoll, um dem Leser einen Eindruck der gesamten Spannbreite von Daten zu geben.
- Allerdings kaum Aussagekraft über die tatsächliche Variabilität der Daten
  - Beispiel: die Spannbreite von $\mathbf{x} = \{1, 10, 10, 10, 10, 10, 10, 10, 10, 10, 101\}$ ist $101 - 1 = 100$, obwohl die Daten bis auf die zwei Ausreißer 1 und 101 keine Variabiltät aufweisen.

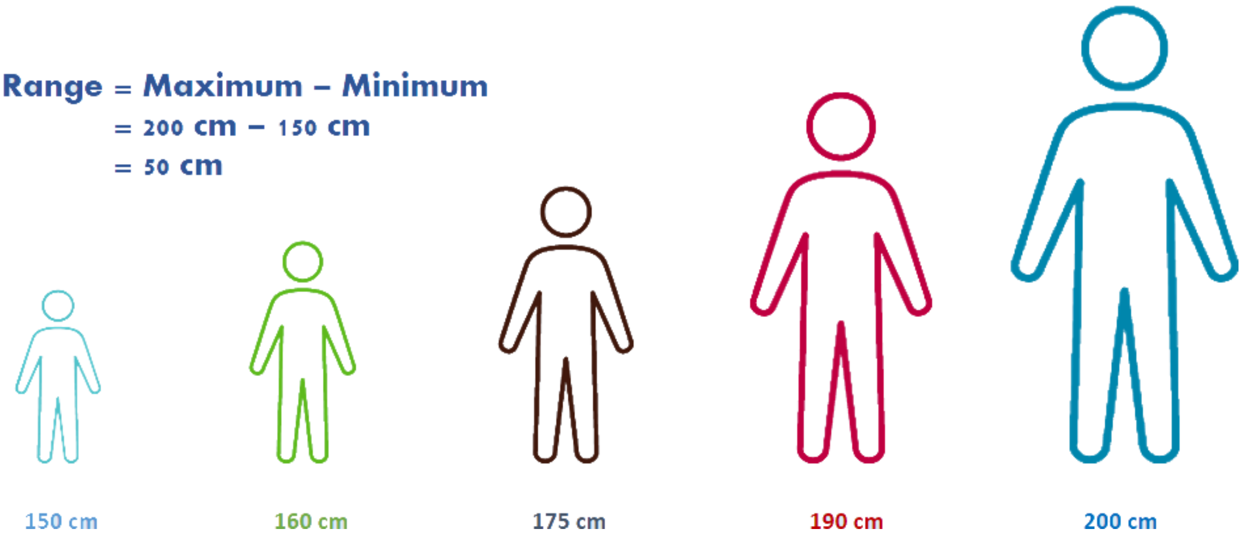
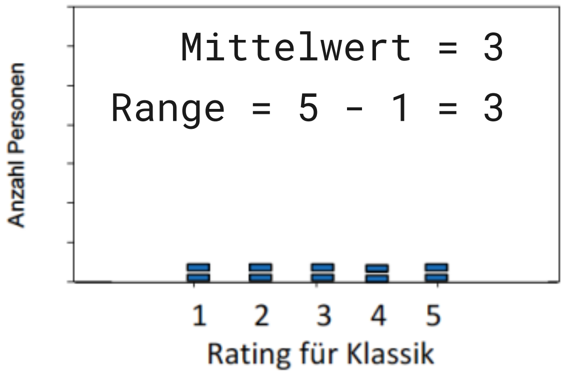
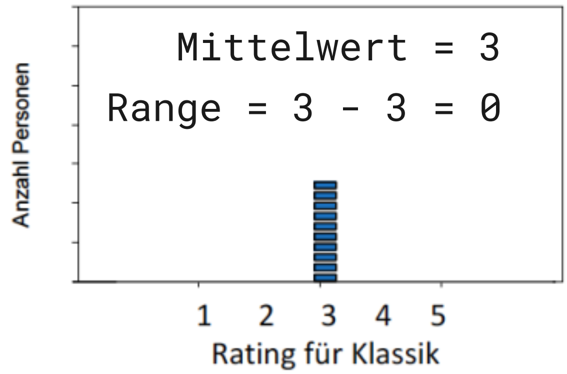


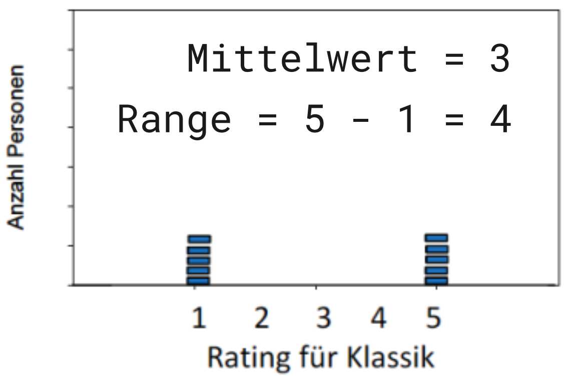
Figure 10.1: Bildnachweis[1]

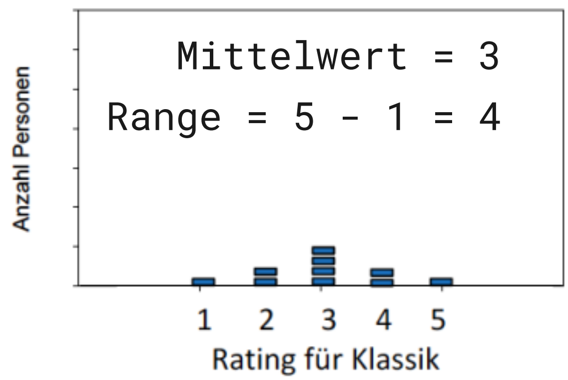
[1] [redacted]https://www.shiksha

10.6 Range: Beispiele









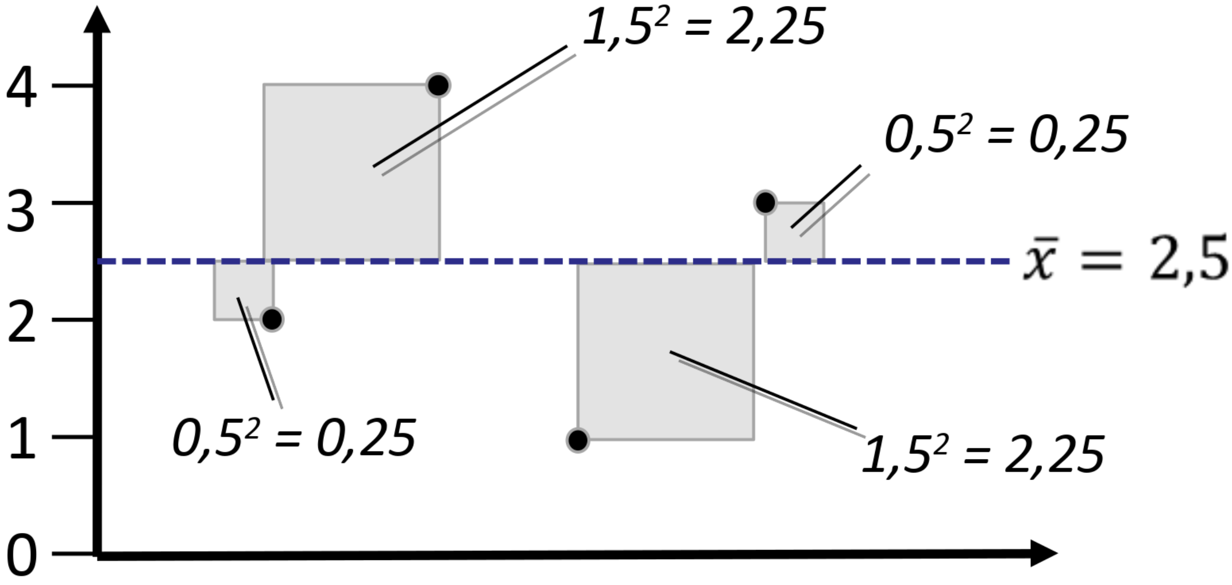
10.7 Varianz

- Die Varianz ist definiert als die **Summe der quadrierten Abweichungen aller Werte vom Mittelwert**:

$$Var(X) = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

(Die Varianz ist gleich der quadrierten Standardabweichung s – letztere lernen wir noch kennen)

- Durch die Quadrierung wird verhindert, dass sich positive und negative Abweichungen gegenseitig aufheben.



$$
\begin{aligned}
s^2 &= 0{,}25 + 2{,}25 + 0{,}25 + 2{,}25 \\
&= 5
\end{aligned}
$$

10.8 Warum werden werden nicht einfach die Absolutwerte der Differenzen genommen?

Prinzipiell wäre auch eine Formel für die Varianz mit Absolutabständen denkbar:

$$Var_{\text{absolut}}(X) = \frac{1}{N}\sum_{i=1}^{N} |x_i - \bar{X}|$$

Über die Gründe, warum sich Var_{absolut} nicht durchgesetzt hat, streitet sich die Fachwelt. Neben historischen Gründen, gibt es aber einige Eigenschaften, die die Präferenz für Abstandsquadrate zumindest nachvollziehbar machen:

- Quadrierte Abstände gewichten Punkte, die weiter vom Mittelwert entfernt sind, höher. Dies entspricht einer "Bestrafung" von Ausreißern und kann ein wünschenswertes Verhalten sein.
- Analogie zur euklidischen Distanz (z.B. Satz des Pythagoras: $a = \sqrt{b^2 + c^2}$)
- Die Varianz mit Abstandsquadraten ist ein natürlicher Parameter der Normalverteilung
- Fortgeschritten: die Varianz mit Abstandsquadraten ist für alle x differenzierbar (hingegen ist Var_{absolut} bei $x = 0$ nicht differenzierbar)

Weitergehende Literatur [2]

10.9 Standardabweichung

- Ein Nachteil der Varianz ist, dass sie aufgrund der Abstandsquadrate in quadrierten Einheiten angegeben ist:

Beispiel

$x = \{167\,cm, 181\,cm, 154\,cm, 192\,cm, 173\,cm\} \rightarrow Var(X) = 180.4\,cm^2$

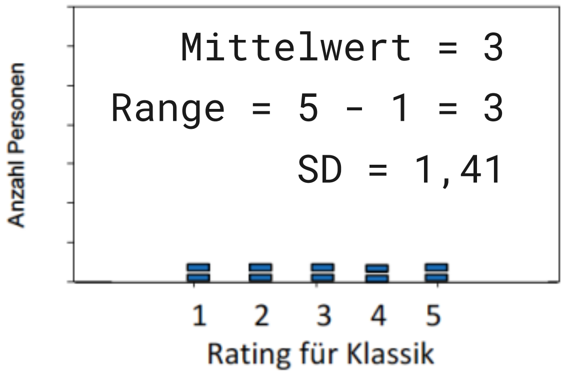
- Quadrierte Einheiten sind jedoch wenig intuitiv und schwer zu interpretieren.
- Aus diesem Grund wird häufig die Standardabweichung angegeben, welche die Wurzel der Varianz darstellt:

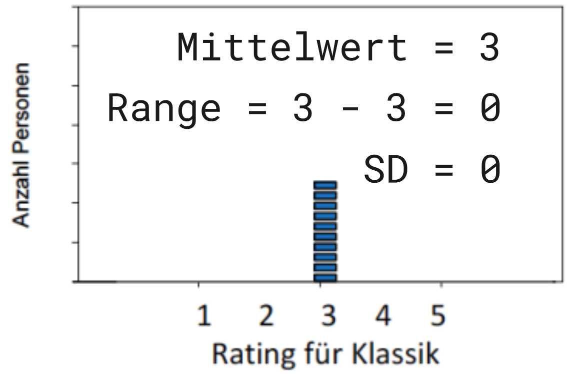
$$s = \sqrt{Var(X)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2}$$

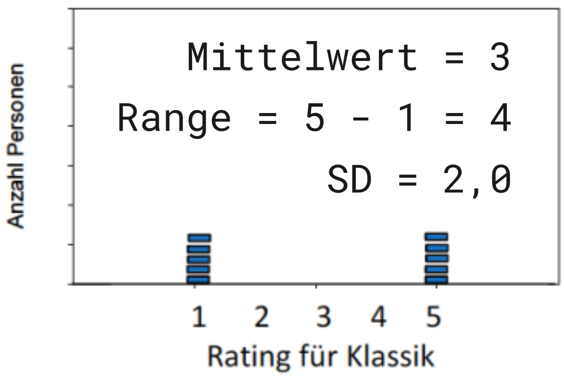
- Die Standardabweichung drückt die Streuung in den **Rohwerten der Skala** aus.
- Sie wird häufig auch mit SD (für *standard deviation*) abgekürzt.

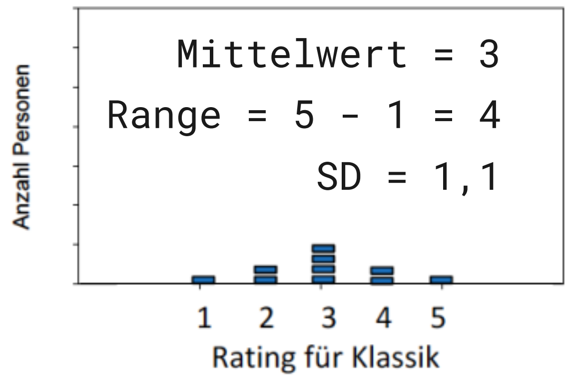
[2] [redacted]https://web.archive.

10.10 Standardabweichung: Klassik-Beispiele









10.11 Interquartilsabstand (interquartile range = IQR)

- **Quartile** sind eine Spezialform von **Quantilen**
- **Quantile** teilen eine Verteilung von Daten in gleich große Abschnitte ein
  - "gleich groß" = jeder Abschnitt hat gleich viele Datenpunkte
- Beispiel *Dezile*: Einteilung der Verteilung in *10* gleich große Abschnitte
  - Hier: 20 Werte in 10 Abschnitte (Decile) á 2 Werte

$$\underbrace{1,1,}_{\text{1.Dezil}}\ \underbrace{1,2,}_{\text{2.Dezil}}\ 2,3,\ 3,5,\ 5,5,\ \underbrace{5,6,}_{\dots}$$

- Beispiel *Quintile*: Einteilung der Verteilung in *5* gleich große Abschnitte
  - Hier: 30 Werte in 5 Abschnitte (Quintile) á 6 Werte

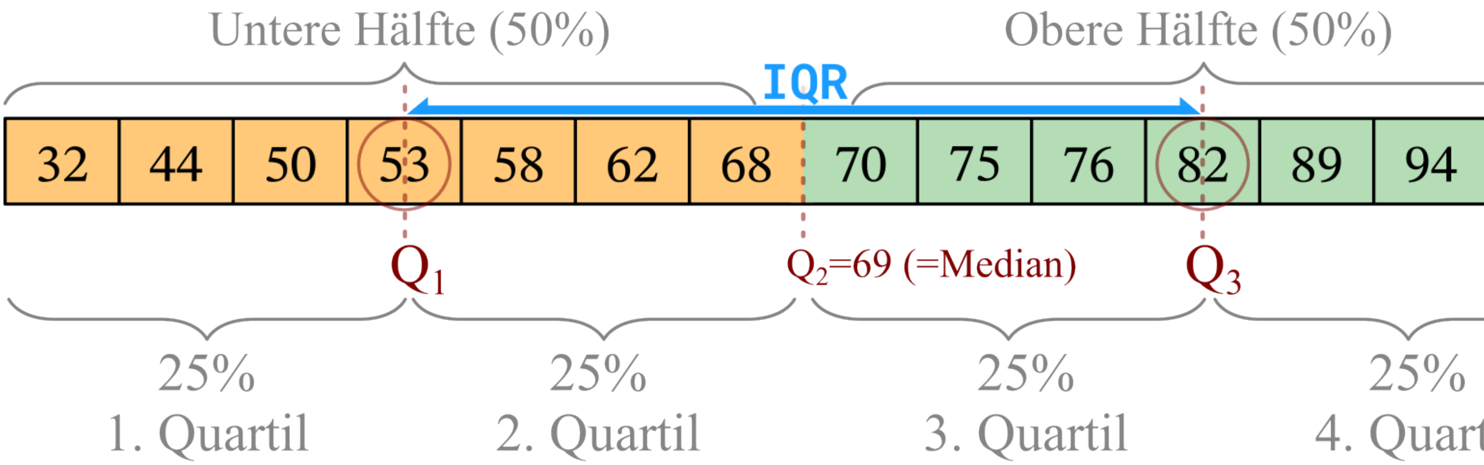
$$\underbrace{1,1,1,1,2,2,}_{\text{1.Quintil}}\ \underbrace{2,2,2,3,3,3,}_{\text{2.Quintil}}\ \underbrace{4,4,5,5,5,5,}_{\text{3.Quintil}}\ \underbrace{5,6,6,7,7,7,}_{\text{4.Quintil}}\ \underbrace{8,8,9,9,9,9}_{\text{5.Quintil}}$$

- Beispiel *Quartile*: Einteilung der Verteilung in *4* gleich große Abschnitte

– Hier: 12 Werte in 4 Abschnitte (Quartile) á 3 Werte

$$\underbrace{1,1,2,}_{\text{1.Quartil}}\ \underbrace{2,2,2,}_{\text{2.Quartil}}\ \underbrace{3,4,5,}_{\text{3.Quartil}}\ \underbrace{5,6,6}_{\text{4.Quartil}}$$

10.12 Interquartilsabstand (interquartile range = IQR)



- Um eine Reihe von Daten in 4 gleich große Quartile zu teilen, sind genau drei Quartilsgrenzen notwendig
- Diese Quartilsgrenzen werden mit Q_1, Q_2, Q_3 (bezogen auf *Quartile*) bzw. mit $Q_{25\%}$, $Q_{50\%}$, $Q_{75\%}$ (bezogen auf *Quantile*) bezeichnet
- Der Interquartilsabstand (IQR) ist die Differenz aus der 75%-*Quantils*grenze und der 25%-Quantilsgrenze bzw. die Differenz aus der 3. und der 1. *Quartils*grenze:

$$IQR = Q_{75\%} - Q_{25\%} = Q_3 - Q_1$$

10.13 Interquartilsabstand (interquartile range = IQR)

- Berechnung des IQR:
  1. Sortiere alle Werte von klein nach groß
  2. Bestimme die Tiefe des Medians (runde ab bei unganzzahligem Wert): $Tiefe_{\text{Median(abgerundet)}}$
  3. Bestimme die Tiefe des Quartils: $Tiefe_{\text{Quartil}} = \frac{Tiefe_{\text{Median(abgerundet)}}+1}{2} = \frac{\text{abgerundet}\left(\frac{N+1}{2}\right)+1}{2}$
  4. Für das 25%-Quantil (Q_1) geht man **von vorne** in die Datenreihe
  5. Für das 75%-Quantil (Q_3) geht man **von hinten** in die Datenreihe

Beispiel

Folgende 11 Werte werden beobachtet: $x = \{1, 1, 2, 3, 3, 3, 4, 6, 6, 7, 9\}$
In diesem Fall ist der 6. Wert der Median, also $Tiefe_{\text{Median(abgerundet)}} = 6$
Die Tiefe des Quartils ist damit $Tiefe_{\text{Quartil}} = \frac{Tiefe_{\text{Median(abgerundet)}}+1}{2} = \frac{6+1}{3} = 3{,}5$
Der "3,5"-te Wert von vorne ist der Mittelwert aus 2 und 3 ($Q_1 = 2{,}5$), der "3,5"-te Wert von hinten ist der Mittelwert aus 6 und 6 ($Q_3 = 6$)
$IQR = Q_3 - Q_1 = 6 - 2.5 = 3{,}5$

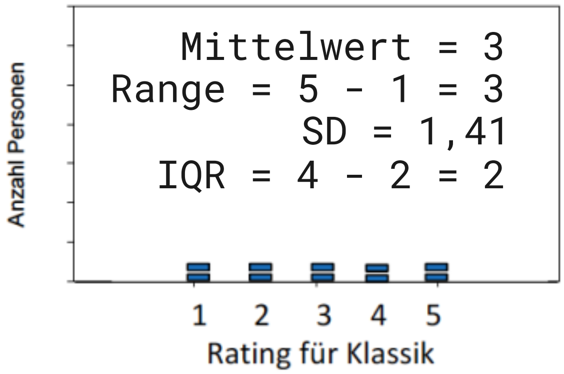
10.14 Wann ist der Interquartilsabstand ein sinnvolles Streuungsmaß?

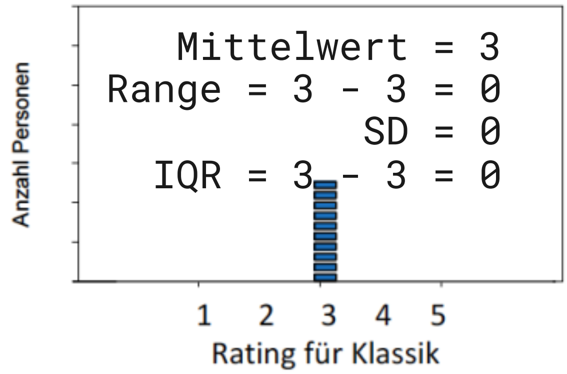
Outliers: When your dataset contains outliers, extreme values can heavily influence other dispersion measures like the range or standard deviation. The IQR is robust to outliers since it only depends on the middle 50% of the data.

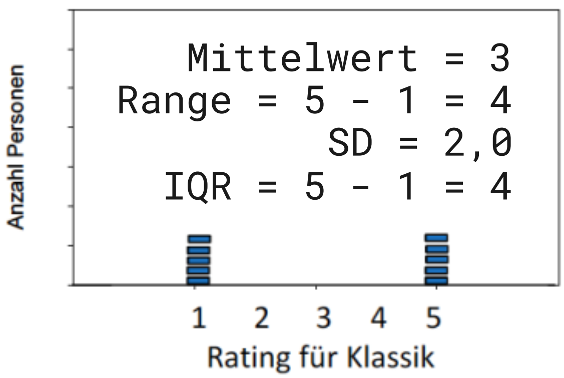
Skewed Data: If the data is not normally distributed and has a skewed shape, using the IQR can be advantageous because it is less affected by the tail of the distribution compared to other measures like the standard deviation.

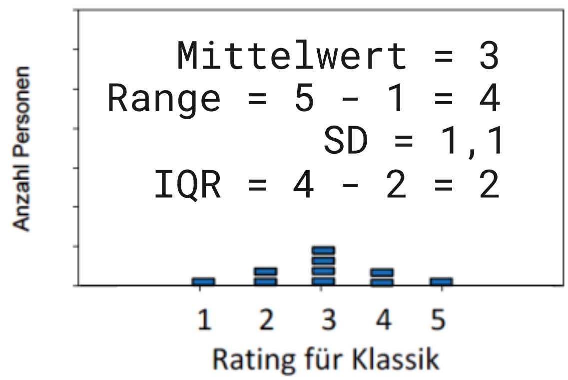
Der Interquartilsabstand ist zur Varianz, wie der Median zum Mittelwert

10.15 Interquartilsabstand: Klassik-Beispiele









10.16 Die Streuungsmaße im Vergleich

Spannbreite (Range)	Gibt die Ausdehnung des gesamten Wertebereiches an Auf kein bestimmtes Lagemaß bezogen Geringer statistischer Nutzen, manchmal interessante Zusatzinfo Maximal abhängig von Ausreißern
Varianz	Auf den Mittelwert bezogen ("wie stark streuen die Daten um den Mittelwert?") Relativ anfällig gegenüber Ausreißern Unnatürliche quadrierte Einheiten
Standardabweichung	Wie Varianz, aber natürliche unquadrierte Einheiten
Interquartilsabstand	Auf kein bestimmtes Lagemaß bezogen Jedoch häufig im Zusammenhang mit Median angegeben Sehr robust gegenüber Ausreißern

11 Darstellung von Lage- und Streuungsmaßen in Text und Bild

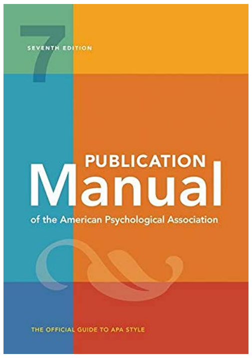
11.1 Angabe von Lage- und Streuungsmaßen in wissenschaftlichen Arbeiten

- Handelt es sich um einzelne Werte, können diese übersichtlich im Fließtext berichtet werden:

> Respondents' mean rating for self-estimated musicality was 6.1 (SD = 2.4), and the mean rating for importance of music in their life was 8.2 (SD = 1.8). Thus, partici-

- Handelt es sich um eine größere Anzahl von Werten (z.B. bei mehreren Bedingungen) bietet sich eine Darstellung in Tabellenform an.
- Auch bei dieser Darstellung sollten Lage- *und* Streuungsmaße angegeben werden.
- In wissenschaftlichen Manuskripten werden Tabellen häufig nach den APA-Richtlinien (7. Edition, 2020) formatiert:

Musikstil	*M*	*SD*
Klassik	2,9	1,1
Rock	4,1	0,8
Rap	3,3	1,2



Siehe Link[1] für Informationen zur Formatierung von Tabellen im APA-Stil

11.2 Beispieltabelle

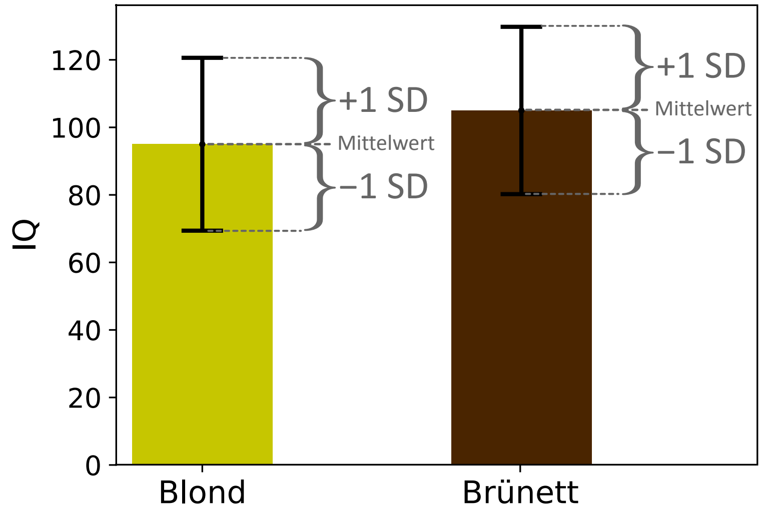
TABLE 4 *Mean ratings (and SDs) for the functions of music within the six dimensions of musical styles*

	Preference dimensions (genres)						
Statement	All styles (N = 503)	Sophisticated (n = 38)	Electronic (n = 30)	Rock (n = 218)	Rap (n = 37)	Pop (n = 88)	Beat, folk, & country music (n = 10)
Puts me in a good mood	8.2 (1.3)	8.3 (1.6)	8.4 (1.0)	8.3 (1.2)	8.0 (1.9)	7.8 (1.4)	8.7 (0.5)
Helps me chill and tune out	6.9 (2.3)	7.6 (2.1)	6.9 (2.6)	6.8 (2.4)	6.4 (2.9)	7.2 (1.7)	7.3 (2.2)
Energizes me	6.8 (2.0)	6.2 (2.3)	7.3 (1.8)	6.8 (1.9)	7.2 (2.0)	6.4 (2.1)	7.6 (1.8)
Lets me appreciate as art	6.5 (2.6)	8.1 (1.5)	6.4 (2.6)	6.7 (2.5)	5.7 (3.1)	5.7 (2.9)	6.3 (2.9)
Enables me to reminisce	6.4 (2.5)	5.2 (2.9)	6.1 (2.7)	6.8 (2.2)	6.0 (2.3)	6.6 (2.4)	7.8 (1.1)
Enables me to better understand my thoughts and feelings	6.0 (2.5)	5.8 (3.0)	5.5 (2.8)	6.2 (2.4)	5.0 (3.1)	5.9 (2.3)	6.5 (2.0)
Is what I listen to as background music	5.7 (2.8)	5.4 (2.8)	6.2 (2.3)	5.5 (2.8)	5.5 (2.9)	6.2 (2.7)	5.0 (3.1)
Is what I like to dance to	5.6 (3.2)	2.8 (2.9)	7.5 (2.8)	5.7 (3.0)	7.4 (2.3)	5.4 (3.0)	5.5 (3.2)
Expresses my identity	5.5 (2.7)	5.5 (2.3)	5.2 (2.4)	6.0 (2.6)	4.7 (2.9)	4.9 (2.7)	6.3 (2.8)
Expresses my values	5.5 (2.6)	5.7 (2.6)	4.1 (2.9)	6.0 (2.5)	5.2 (2.4)	5.1 (2.5)	6.3 (2.5)
Lets me forget my problems	5.5 (2.6)	5.8 (2.3)	6.0 (2.4)	5.3 (2.7)	4.5 (2.9)	5.1 (2.4)	6.6 (1.8)
Helps me feel close to others	5.4 (2.7)	5.0 (3.1)	5.2 (2.8)	5.6 (2.6)	5.1 (2.9)	5.3 (2.5)	5.3 (2.8)
Makes me feel ecstatic	4.9 (3.2)	3.8 (3.1)	6.5 (3.1)	5.3 (3.0)	4.3 (3.2)	3.6 (3.0)	6.8 (1.8)
Lets me experiment with different sides of my personality	4.8 (3.0)	3.6 (2.8)	4.7 (3.4)	5.1 (2.9)	4.8 (3.1)	4.4 (2.9)	4.6 (3.7)
Helps me meet people	4.4 (2.8)	3.6 (3.0)	5.1 (3.0)	4.7 (2.8)	4.3 (2.8)	3.8 (2.8)	4.0 (2.7)
Makes me identify with the artists	3.4 (2.9)	3.5 (3.3)	3.3 (3.2)	3.6 (2.8)	3.2 (2.9)	2.7 (2.5)	4.3 (3.4)
Gives me information	3.3 (2.8)	3.5 (3.1)	1.9 (2.3)	3.6 (2.7)	2.8 (2.8)	2.8 (2.4)	4.1 (3.0)

11.3 Balkendiagramm mit Fehlerbalken

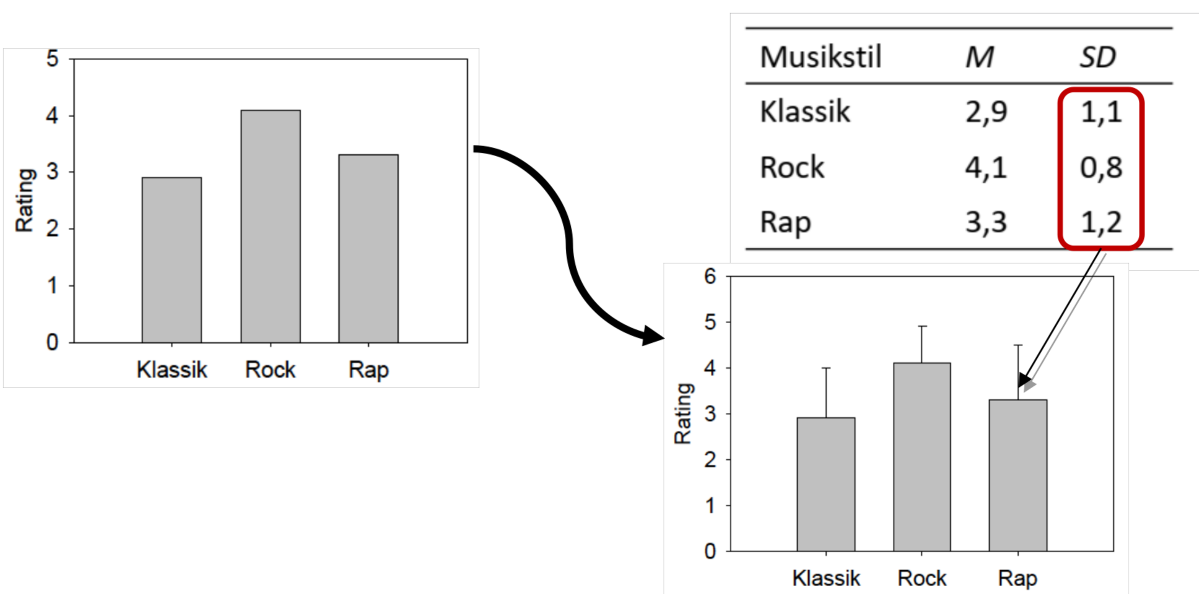
- Der **Fehlerbalken** ist die klassische graphische Repräsentation der Streuung:

[1] [redacted]https://apastyle.apa



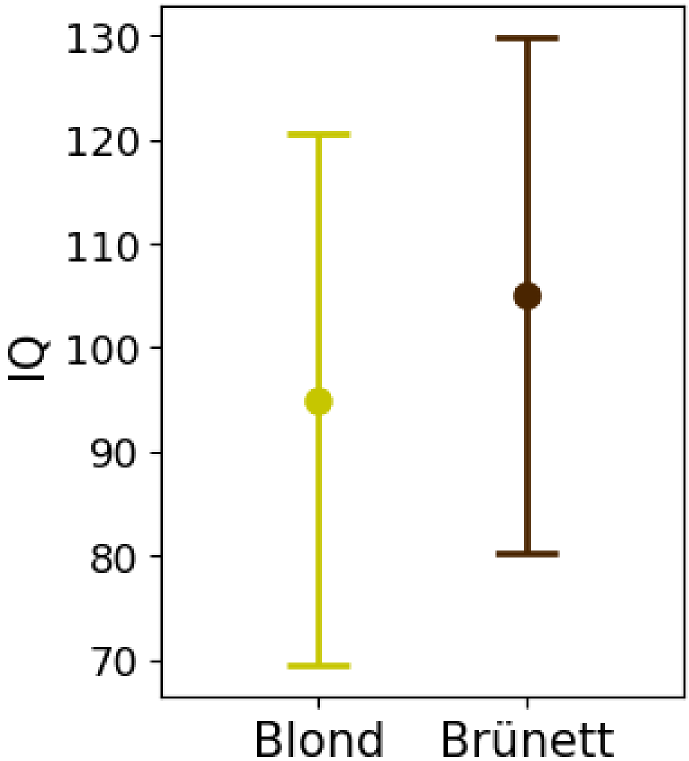
- Die Darstellung entspricht dabei dem Lagemaß (im Bild der Mittelwert) plus/minus dem Streuungsmaß (im Bild die Standardabweichung)
- Der gesamte Fehlerbalken hat also die Ausdehnung $2 \cdot$ Streuungsmaß
- Klassische Kombinationen von Lagemaß und Streuungsmaß sind:
  - Mittelwert $\pm$ Standardabweichung
  - Mittelwert $\pm$ Standardfehler (dazu kommen wir noch)
  - Mittelwert $\pm$ Konfidenzintervall (dazu kommen wir noch)
  - Median $\pm$ IQR
  - Median $\pm$ Median-Abweichung (das behandeln wir nicht)

11.4 Beispiel



11.5 Einzelner Punkt statt Balken

- Statt eines Balkens kann der Mittelwert auch durch einen einzelnen Punkt repräsentiert werden – manche empfinden das als eleganter:



11.6 Messwiederholungen: Liniendiagramme

Gerade als sich die Hinweise auf die Hypothese des viralen Ursprungs von Paradoxia verdichten, wird ein Blogpost des Chaos Computer Club (CCC) in der Öffentlichkeit bekannt. Anonymen Hackern gelang es, auf die letzten 12 Monate Tik-Tok-Historie von 1800 Personen zuzugreifen – darunter viele Paradoxiker!

11.7

Die Daten des CCC geben der alternativen Hypothese neue Nahrung, nach der es sich bei Paradoxia um ein soziales Tik-Tok-Phänomen handelt. Leider ist nicht bekannt, zu welchem Zeitpunkt die Paradoxiker in dem Datensatz erkrankt sind. Die erste Frage an Sie als Task Force lautet: welche Möglichkeiten von Ursache und Wirkung könnten plausibel sein?

11.8 The case against bar plots

- Trotz ihrer hohen Verbreitung haben Balkendiagramme (Barplots) eine Reihe von Nachteilen:

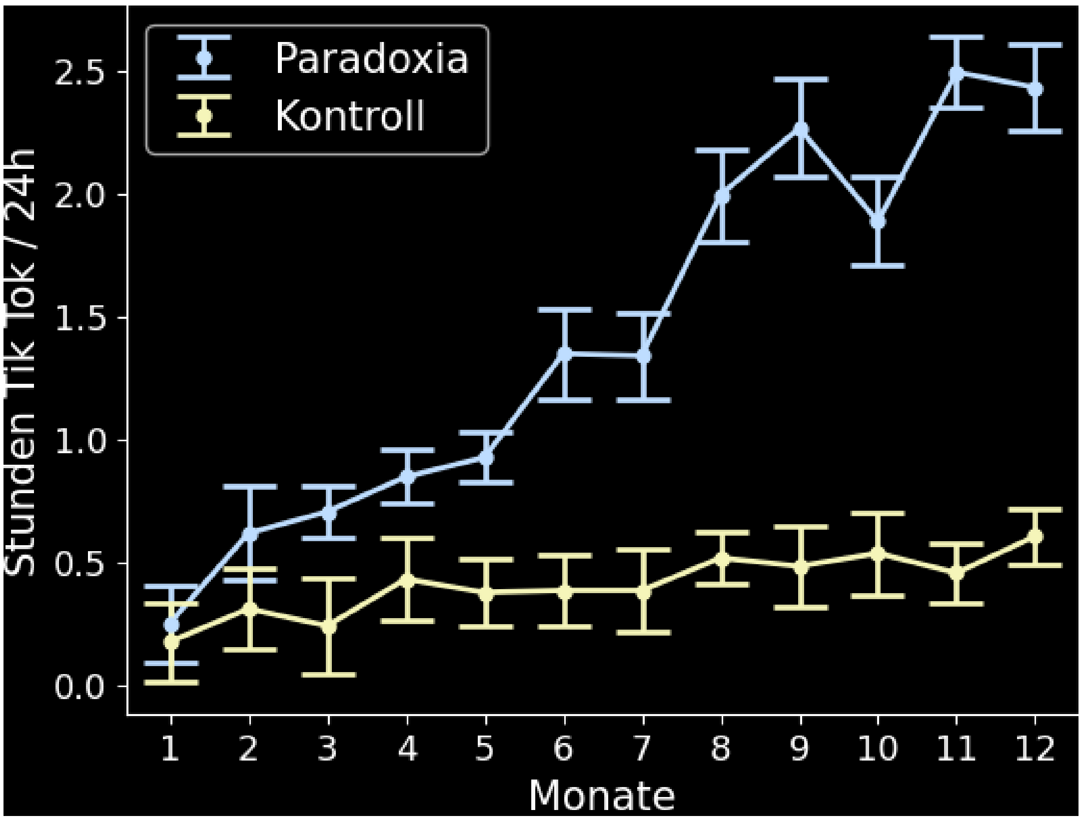
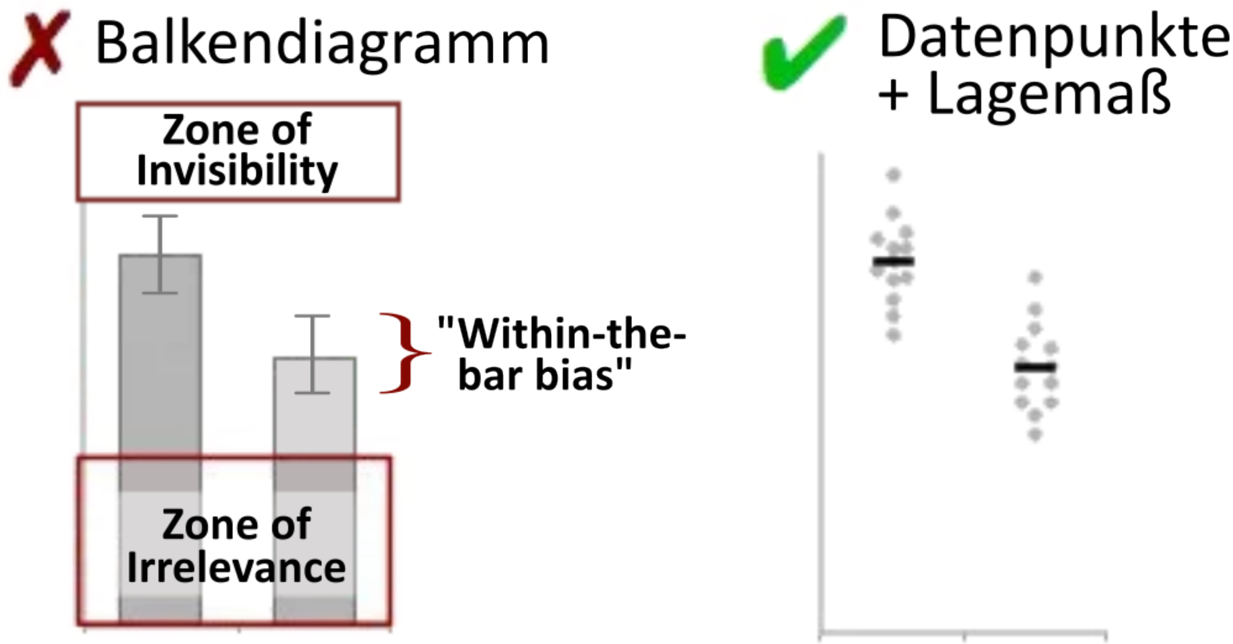


Figure 11.1: Die Abbildung zeigt ein Liniendiagramm mit Fehlerbalken. Bei wiederholten Messungen Personen bietet es sich an, die einzelnen Messzeitpunkte mit Linien zu verbinden, um den zeitlichen Zusammenhang zu unterstreichen. Überlegen Sie sich: welche Aussage macht ein einzelner Fehlerbalken in dieser Abbildung?

- Sie geben kaum Information über die spezifische Verteilung der Daten und mögliche Ausreißer
- Die intransparente Darstellungsweise verdeckt häufig, dass 1) Daten unrealistisch sind oder 2) Ausreißer das Lagemaß verzerren oder 3) die Verteilung der Daten unpassend für das verwendete Lagemaß sind.
- Sie legen den Fokus auf irrelevante Bereiche der Skala (siehe Abbildung unten)



Durch die tatsächliche Verteilung der Datenpunkte im rechten Plot wird klar, dass im Balkendiagramm ein vergleichsweise starker Fokus auf Bereiche gelegt wird, in den gar keine Daten enthalten sind ("Zone of Irrelevance"), und andererseits Extremwerte, insbesondere oberhalb des Fehlerbalkens, visuell völlig unrepräsentiert sind ("Zone of Invisibility"). Der Fehlerbalken ist außerdem leicht mit der Illusion verbunden, dass sich alle Datenpunkte innerhalb des angezeigten Bereiches befinden ("Within-the-bar bias").[2].

11.9 The case against bar plots

- Alle Datenverteilungen haben den gleichen Mittelwert und Standardabweichung

11.10 The case against bar plots

11.11 Der Box-Plot

- Eine bekannte Darstellungsform für den Median ist der **Boxplot**

[2] [redacted] Weissgerber NM (2019) Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency. Circulation 140:1506–1518.

[3] [redacted] https://www.biztory

[4] [redacted] https://blog.revoluti

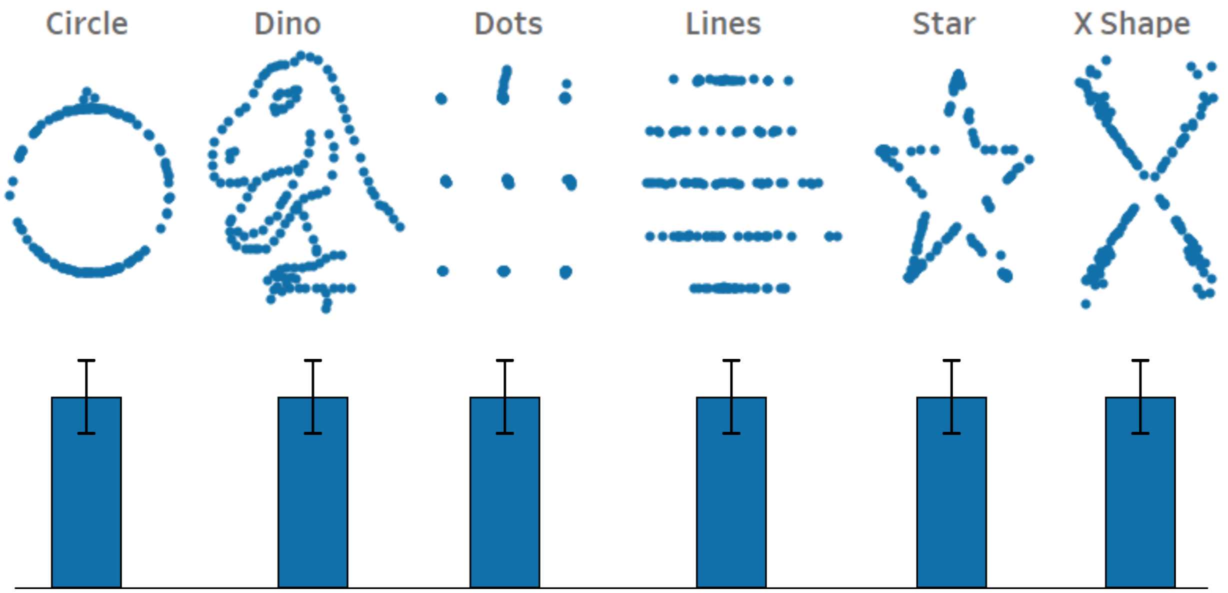


Figure 11.2: Illustration der Tatsache, dass unterschiedlichste Verteilungen von Daten im selben Balkendiagramm münden, hier am Beispiel von Mittelwert und Standardabweichung. Alle Verteilungen haben exakt den gleichen Mittelwert und Standardabweichung, sowohl entlang der y-Achse, als auch entlang der x-Achse (hier nicht dargestellt)[3].

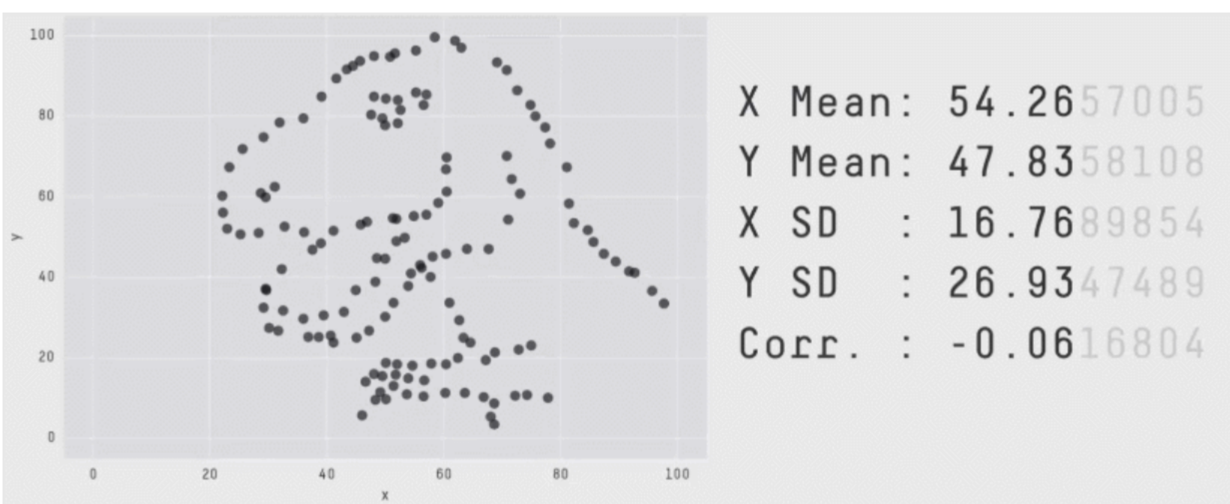
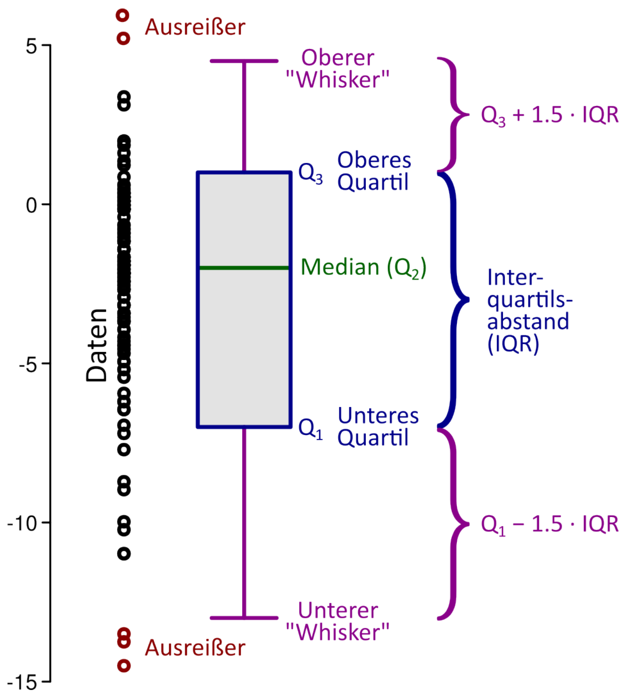


Figure 11.3: Alberto Cairo's "DataSaurus"[4].

- In einer der bekanntesten Variationen zeigt der Boxplot drei Informationen an:
    1. **Median (als einfache Linie)**
    2. **Box:** die "mittleren 50% der Daten" (25% über dem Median, 25% unter dem Median)
    3. **Antennen ("Whisker"):** der Bereich der Daten, der keine Ausreißer enthält. Häufig ist hier das Kriterium, dass die Daten im Bereich $[Q_1 - 1.5 \cdot IQR; Q_3 + 1.5 \cdot IQR]$ liegen müssen.



- Der Boxplot gibt eine schnelle Übersicht über wesentliche Kennwerte eines Datensatzen
- Zu beachten ist, dass zahlreiche Variationen des Boxplots existieren
    - Bei einer weiteren bekannten Variation zeigen die Antennen das absolute Maximum und Minimum der Daten an (also *inklusive* möglicher Ausreißer)

11.12 Moderne Darstellungsformen

Raincloud-Plot

Klassische Darstellung mit Lage- und Streuungsmaß

[5] [redacted] https://jasp-
[6] [redacted] https://jasp-

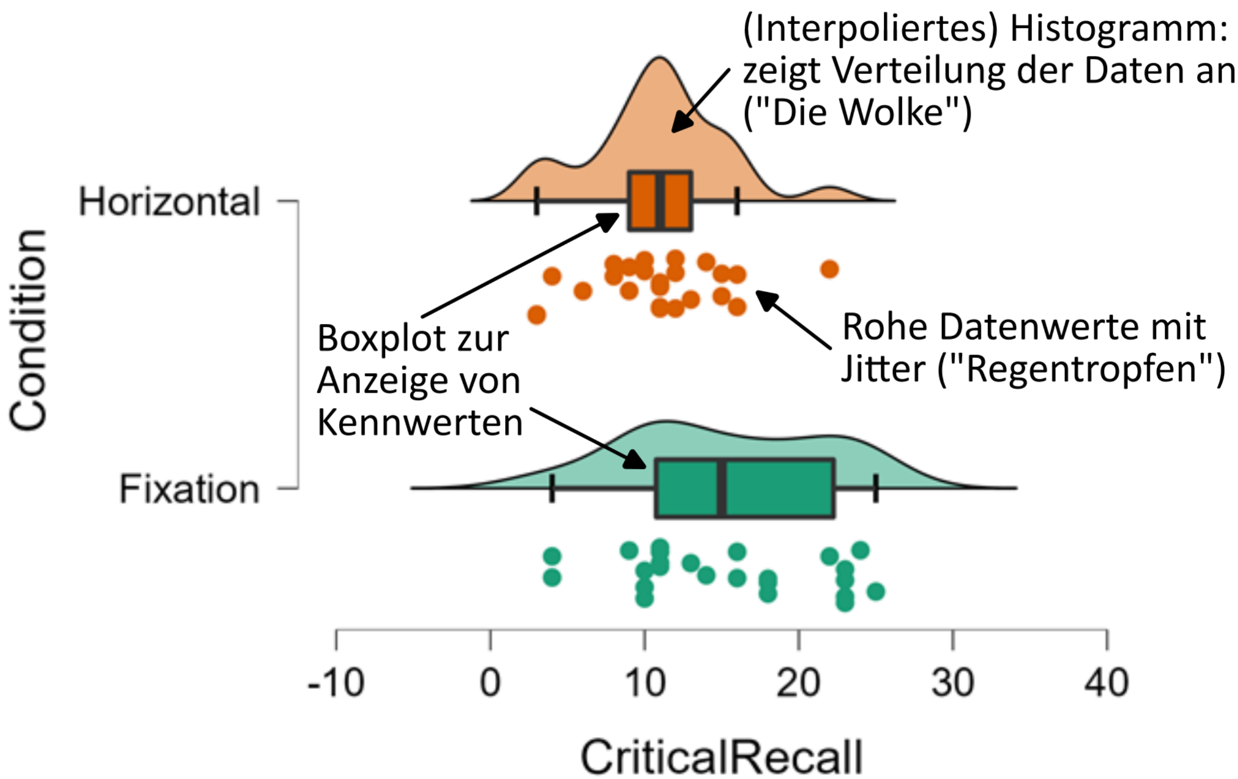


Figure 11.4: Bildnachweis[5]

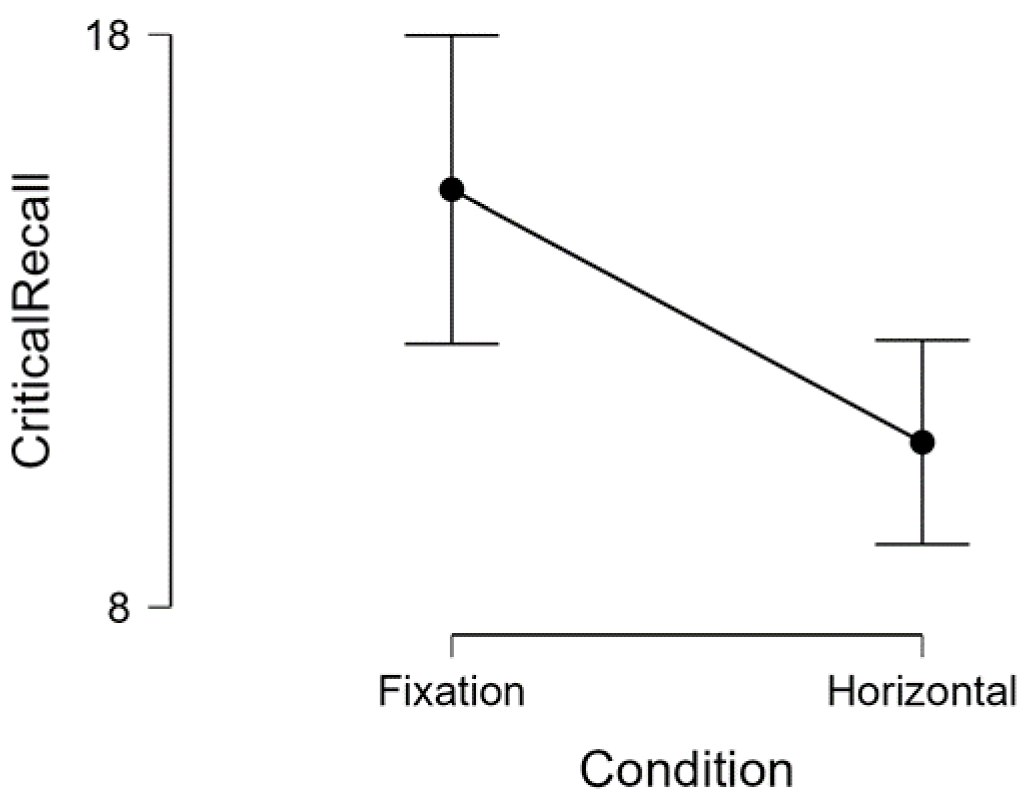


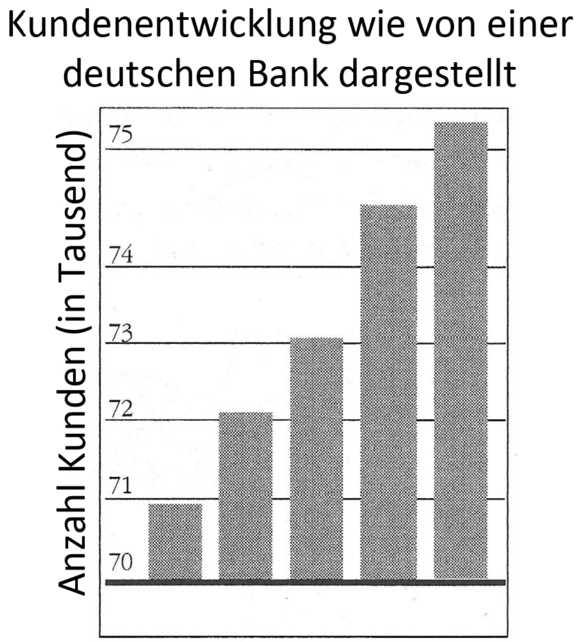
Figure 11.5: Bildnachweis[6]

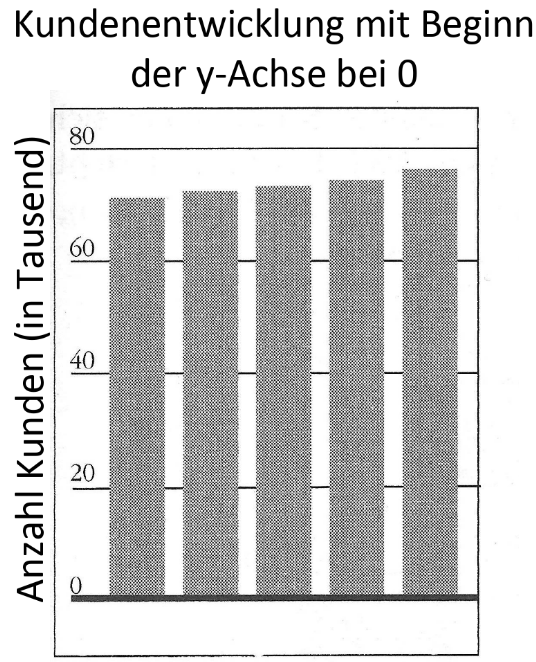
- Moderne Statistik- und Plotsoftware ermöglicht heutzutage eine verbesserte und transparentere Darstellung von Daten:
  - Anzeige einzelner Datenpunkte, meist getrennt durch “Jitter” (d.h. leichte horizontale oder vertikale Versetzung mit zufälligen Abständen, um Überschneidung der Datenpunkte zu reduzieren)
  - Anzeige der Verteilung mittels (interpolierter) Histogramme (wichtige Information für die Auswahl geeigneter Lage- und Streumaße, aber auch statistischer Tests)
  - *Zusätzliche* Anzeige von Kennwerten

12 Problematische Abbildungen

12.1 Abschneiden der y-Achse

- Das Abschneiden der y-Achse verzerrt häufig die Stärke von Effekten





- Die grundsätzliche Empfehlung ist daher, die y-Achse bei 0 beginnen zu lassen
- Es gibt aber Ausnahmen:
  - Vorliegen eines anderen natürlichen Referenzwertes (z.B. IQ-Wert 100, wenn alle Werte über 100 liegen).
  - Wären *tatsächlich vorhandene Unterschiede* zwischen Balken verschiedener Bedingungen überhaupt nicht mehr wahrnehmbar, kann ein Abschneiden der y-Achse sinnvoll sein (oder eine Logarithmus-Skala!).
  - In manchen Fällen können Messwerte niemals unter einen Mindestwert fallen. Beispielsweise sind motorische Reaktionszeiten physiologisch bedingt fast immer über 100ms – in diesem Fall ist der Bereich 0-100ms “Totraum” und kann sinnvollerweise weggelassen werden.
  - Im Idealfall wird das Abschneiden der y-Achse durch einen “Bruch” angezeigt (siehe nächste Folie).

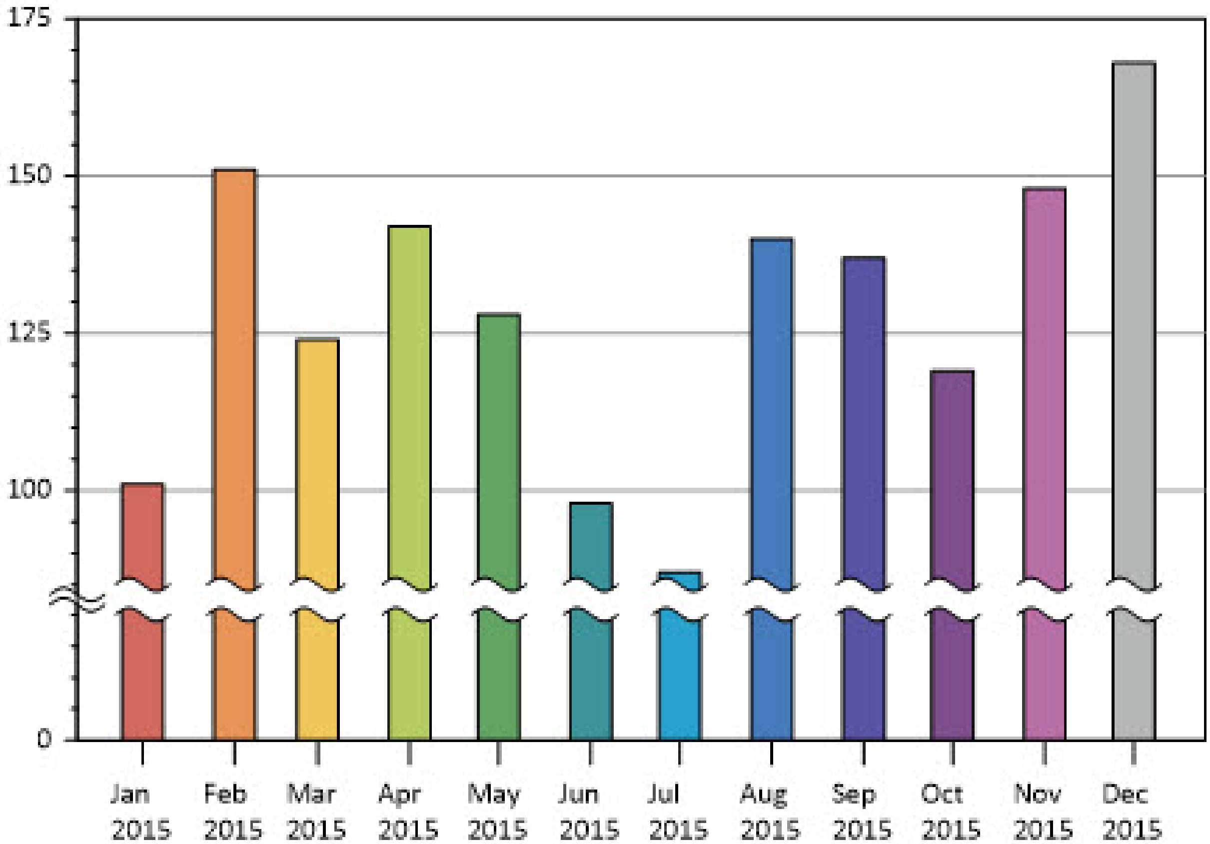


Figure 12.1: Durch den Bruch der y-Achse wird betont, dass die Balken zur besseren Übersichtlichkeit “abgeschnitten” wurden. So werden fälschliche Wahrnehmungen und Interpretationen durch die abgeschnittene Achse eher vermieden. Bildnachweis[1]

12.2 Beispiel für gebrochene y-Achse

12.3 Gleiche Intervalle auf x-Achse

- Ungleichmäßige Intervalle auf der x-Achse verzerren die Daten von Liniendiagrammen

[1] [redacted]https://support.gold

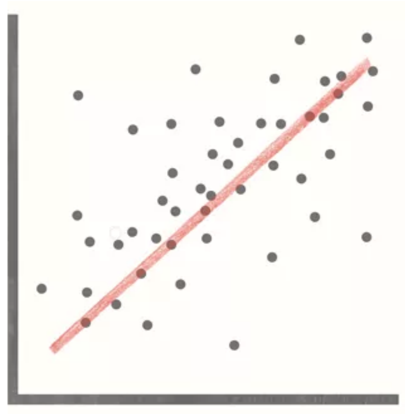
Part IV

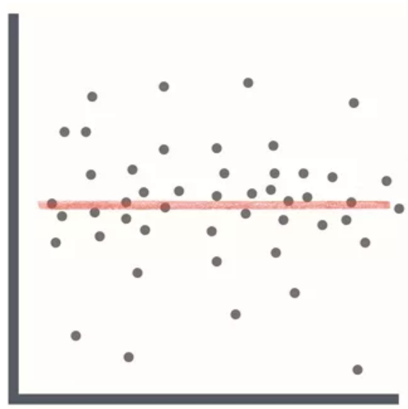
Vorlesung 04: Korrelation

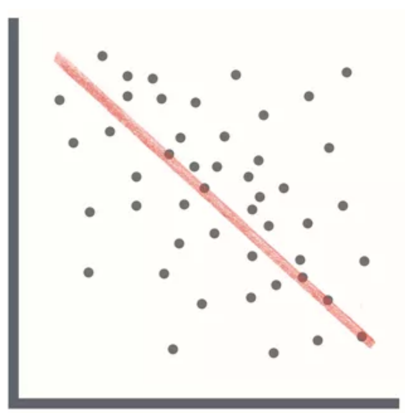
13 Vorlesung 04: Zusammenhänge

13.1 Was sind Zusammenhänge?

- Ein **Zusammenhang** beschreibt zu welchem Grad zwei Variablen (Merkmale) systematisch miteinander in Verbindung stehen
- Zusammenhänge bilden die Essenz der psychologischen Forschung—durch sie versuchen wir die Mechanik der menschlichen Psyche zu verstehen:
  - Fördert **Ausdauersport** das **psychische Wohlbefinden**?
  - Helfen **Psychotherapiestunden** bei der Überwindung einer **Depression**?
  - Wirkt sich **Bildschirmzeit** nachteilig auf die **Schlafqualität** aus?
  - Steigt durch **kindliche Frühförderung** die Wahrscheinlichkeit für einen **akadamischen Bildungsabschluss**?
- Fast jede wissenschaftliche Hypothese lässt sich als Zusammenhang formuliern – selbst Unterschiede!
  - <u>Formulierung als Unterschied</u>: unterscheiden sich Männer und Frauen in ihren verbalen Fähigkeiten?
  - <u>Formulierung als Zusammenhang</u>: steht die kategorische Variable **Geschlecht** in Zusammenhang mit der metrischen Variable **verbale Fähigkeiten**?







13.2 Zusammenhänge im engeren Sinn

- Um Unterschiede und Zusammenhänge voneinander abzugrenzen, verstehen wir nachfolgend Zusammenhänge in einem engeren Sinn:

Definition

Ein **Zusammenhang** (im engeren Sinn) beschreibt zu welchem Grad die **Variation zweier metrischer Variablen** miteinander in Verbindung steht.

- Durch die Eingrenzung auf **metrische Variablen** (diskret oder kontinuerlich) stellt etwa die Verbindung von Geschlecht und verbalen Fähigkeiten keinen Zusammenhang im engeren Sinn dar, da Geschlecht eine kategorische Variable ist.
- In den meisten Fällen sind Zusammenhänge das "schärfere statistische Schwert" als Unterschiede und es lohnt sich oft, Forschungsfragen entsprechend anzupassen:
  - Unterscheidet sich die akademische Leistung von Rauchern und Nichtrauchern? Zusammenhang **Zahl der Zigaretten pro Tag** und **akademische Leistung**
  - Unterscheidet sich der Medienkonsum von Depressiven und Kontrollen? Zusammenhang **Depressivität** und **Medienkonsum**
  - Unterscheidet sich das Depressionsrisiko zwischen Nord- und Süddeutschland? Zusammenhang **geographischer Breitengrad** und **Depressionsrisiko**

14 Kovarianz

14.1 Kovarianz

- Ziel: mathematische Größe, die zum Ausdruck bringt, wie stark die Variation zweier Variablen miteinander in Zusammenhang steht
- Wir haben bereits eine Größe für die Variation *einer* Variable — die Varianz:

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(x_i - \bar{X})$$

- Die Varianz gibt an, wie stark eine Variable X um ihren Mittelwert $\bar{X}$ schwankt
- Analog berechnet die **Kovarianz**, wie stark die *gemeinsame Schwankung zweier Variablen um ihren jeweiligen Mittelwert* ist:

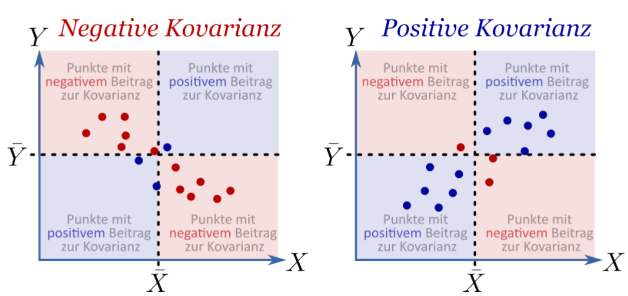
$$\text{Kovarianz:} \quad Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$$

- Im Zentrum der Kovarianz steht die Erkenntnis, dass das **mathematische Produkt** zweier Abweichungsvariablen — hier die Abweichungen $(x_i - \bar{X})$ und $(y_i - \bar{Y})$ vom Mittelwert — angibt, wie stark die beiden Abweichungen gleichsinnig variieren (ko-variieren).

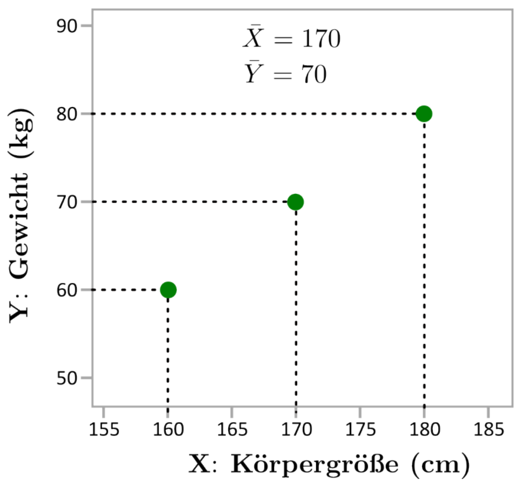
14.2 Kovarianz

Intuition:

- Sind zwei zusammengehörige Datenpunkte x_i und y_i **größer als der Mittelwert**, sind sowohl $(x_i - \bar{X})$ als auch $(y_i - \bar{Y})$ **positiv**, und damit auch das Produkt $(x_i - \bar{X})(y_i - \bar{Y})$ **positiv**.
- Sind zwei zusammengehörige Datenpunkte x_i und y_i **geringer als der Mittelwert**, sind sowohl $(x_i - \bar{X})$ als auch $(y_i - \bar{Y})$ **negativ**, und damit das Produkt $(x_i - \bar{X})(y_i - \bar{Y})$ wieder **positiv**.



14.3 Kovarianz: Größe-Gewicht-Beispiel



$$
\begin{aligned}
&Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y}) = \\
&= \frac{1}{3}[(160-170)(60-70) + (170-170)(70-70) + (180-170)(80-70)] = \\
&= \frac{1}{3}[(-10)\cdot(-10) + 0\cdot 0 + 10\cdot 10] = \\
&= \frac{1}{3}\cdot 200 = 66{,}67
\end{aligned}
$$

- Problem: die Kovarianz hängt von den Einheiten ab!
- Wird im Beispiel die Körpergröße X in der Einheit *Meter* angegeben, so lautet die Kovarianz:

$$
\begin{aligned}
Cov(X,Y) &= \frac{1}{3}[(1{,}60-1{,}70)(60-70) + (1{,}70-1{,}70)(70-70) + (1{,}80-1{,}70)(80-70)] = \\
&= \frac{1}{3}[(-0{,}10)\cdot(-10) + 0\cdot 0 + 0{,}10\cdot 10] = \frac{1}{3}\cdot 2 = 0{,}667
\end{aligned}
$$

- Kovarianzen sind also **nicht vergleichbar wenn sich Einheiten unterscheiden**, und erst recht nicht, wenn sich die Variablen unterscheiden.

15 Pearson-Korrelation

15.1 Pearson-Korrelation

- Die **Pearson-Korrelation** schafft Abhilfe für das Problem der mangelnden Vergleichbarkeit
- Der Schlüssel: die Kovarianz wird mit der Standardabweichung beider Variablen normalisiert:

$$\text{Korrelation: } r_{XY} = \frac{Cov(X,Y)}{s_X s_Y} = \frac{1}{N s_X s_Y} \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y})$$

- Durch das Teilen durch die Standardabweichungen s_X und s_Y werden die Einheiten herausgekürzt — die Korrelation ist also eine **einheitslose Größe**.
- Die Korrelation kann Werte zwischen -1 (perfekter negativer Zusammenhang) und $+1$ (perfekter positiver Zusammenhang) annehmen.

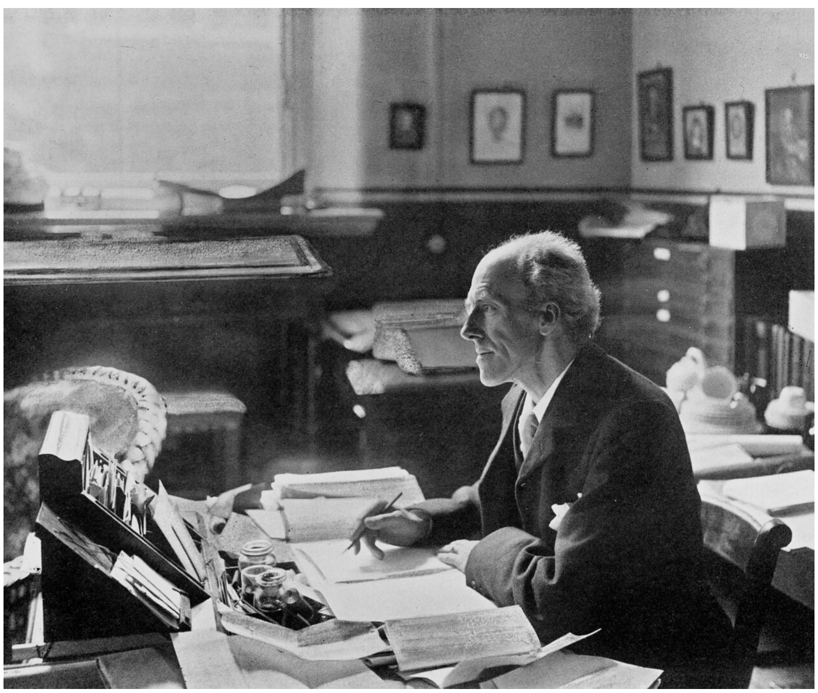


Figure 15.1: Der britische Mathematiker Karl Pearson in seinem Büro im Jahr 1910. Neben dem Korrelationskoeffizienten verdanken wir Pearson viele andere statistische Konzepte wie die Hauptkomponentenanalyse oder den p-Wert. Später wurden seine Ansichten zu Eugenik kritisch hinterfragt. Bildnachweis[1]

[1] [redacted] http://www.learn-

15.2 Warum ist die Korrelation auf −1 bis 1 beschränkt?

Darstellung der Korrelation nur mit (Ko)Varianzen:

$$r_{XY} = \frac{Cov(X,Y)}{s_X s_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Die Korrelation sollte maximal ($r = 1$) sein, wenn X mit sich selbst korreliert wird ($Y = X$). Zu berücksichtigen ist, dass die Kovarianz *von X mit sich selbst* gleich der Varianz ist:

$$r_{XX} = \frac{Cov(X,X)}{\sqrt{Var(X)}\sqrt{Var(X)}} = \frac{Var(X)}{\sqrt{Var(X)}\sqrt{Var(X)}} = \frac{Var(X)}{Var(X)} = 1$$

Umgekehrt sollte die Korrelation maximal negativ sein ($r = -1$), wenn Y genau das Inverse von X ist, also $Y = -X$. Unter Berücksichtigung von $Var(X) = Var(-X)$ gilt:

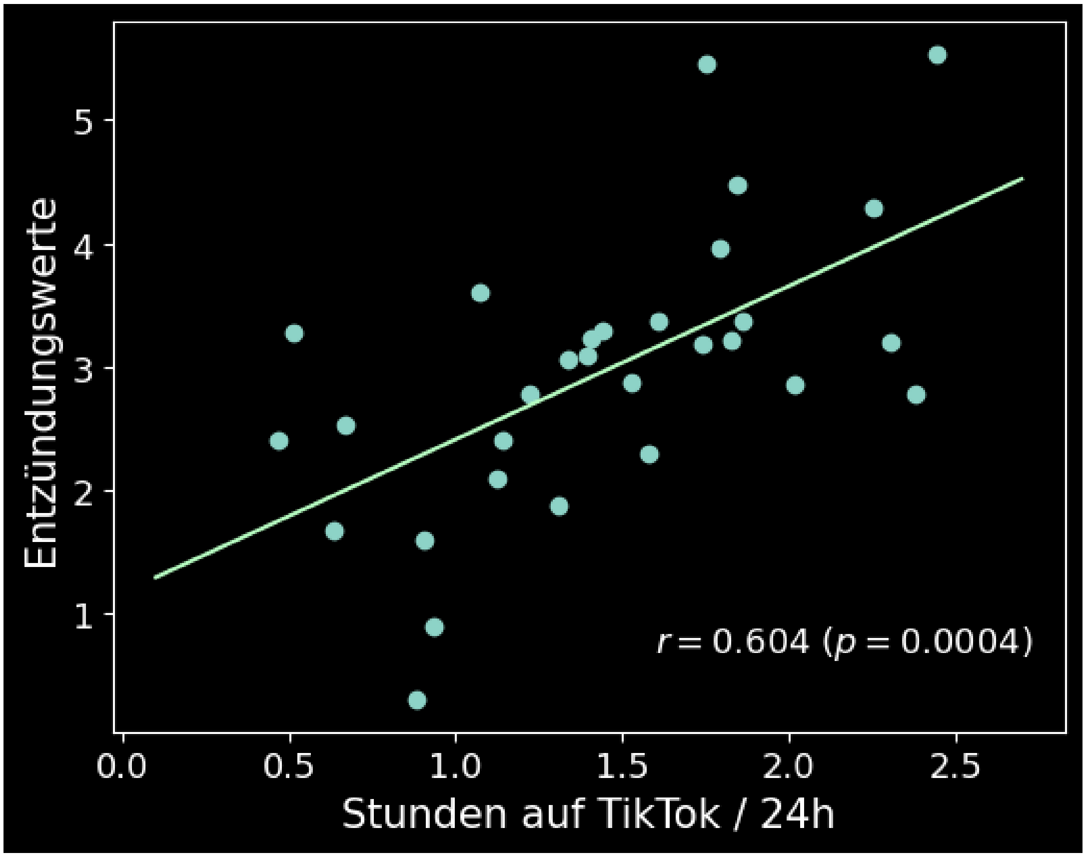
$$r_{X(-X)} = \frac{Cov(X,-X)}{\sqrt{Var(X)}\sqrt{Var(-X)}} = \frac{-Cov(X,X)}{\sqrt{Var(X)}\sqrt{Var(X)}} = \frac{-Var(X)}{Var(X)} = -1$$

15.3 Das Streudiagramm

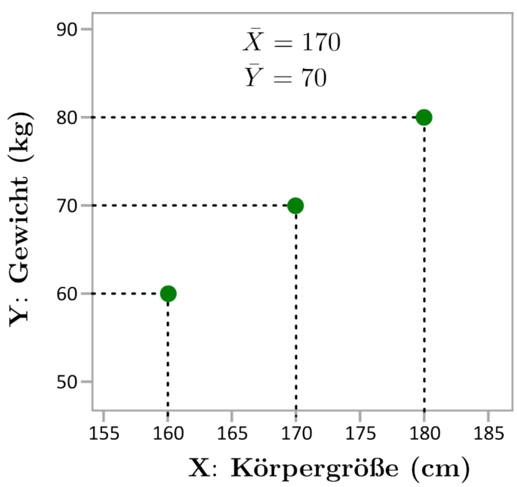
- Von einer kleinen Stichprobe im TikTok-Datensatz des CCC sind auch Entzündungswerte bekannt
- In der Task Force fragen Sie sich, ob es unter den Paradoxikern einen Zusammenhang zwischen der Zahl der Stunden auf TikTok und den Entzündungswerten gibt

Sie fassen Ihr Ergebnis in einem **Streudiagramm** zusammen.

- Das Streudiagramm (engl. *scatter plot*) ist die häufigste Visualisierung eins biavariaten Zusammenhangs
- Bei der Korrelation ist es dabei willkürlich, welche Variable auf der X- und Y-Achse liegt
- Häufig wird zusätzlich zur "Punktwolke" auch eine **Regressionsgerade** angegeben, sowie die Stärke des Zusammenhangs (r=.., p=..)



15.4 Pearson-Korrelation: Größe-Gewicht-Beispiel



Wir hatten:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y}) = \frac{200}{3}$$

Und berechnen nun die Korrelation $r_{XY} = \frac{Cov(X,Y)}{s_X s_Y}$

$s_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2} = \sqrt{\frac{1}{3}[(-10)^2 + 0^2 + 10^2]} = \sqrt{\frac{200}{3}}$

$s_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2} = \sqrt{\frac{1}{3}[(-10)^2 + 0^2 + 10^2]} = \sqrt{\frac{200}{3}}$

$$r_{XY} = \frac{\frac{200}{3}}{\sqrt{\frac{200}{3}} \cdot \sqrt{\frac{200}{3}}} = 1$$

- Wäre die Körpergröße in der Einheit *Meter* angeben, so kürzt sich der Faktor 100 (1m = 100cm) nicht nur im Zähler, sondern auch im Nenner:

$$r_{XY} = \frac{\frac{2{,}00}{3}}{\sqrt{\frac{2{,}00}{3}} \cdot \sqrt{\frac{2{,}00}{3}}} = 1$$

Die Normalisierung mit der Standardabweichung sorgt somit dafür, dass die willkürliche Einheit keine Rolle spielt.

15.5 Interpretation der Pearson-Korrelation

- Die Pearson-Korrelation zeigt an, **wie linear** der Zusammenhang zweier Variablen ausgeprägt ist
- Die Pearson-Korrelation ist dabei nicht von der Steigung einer gedachten Gerade abhängig.

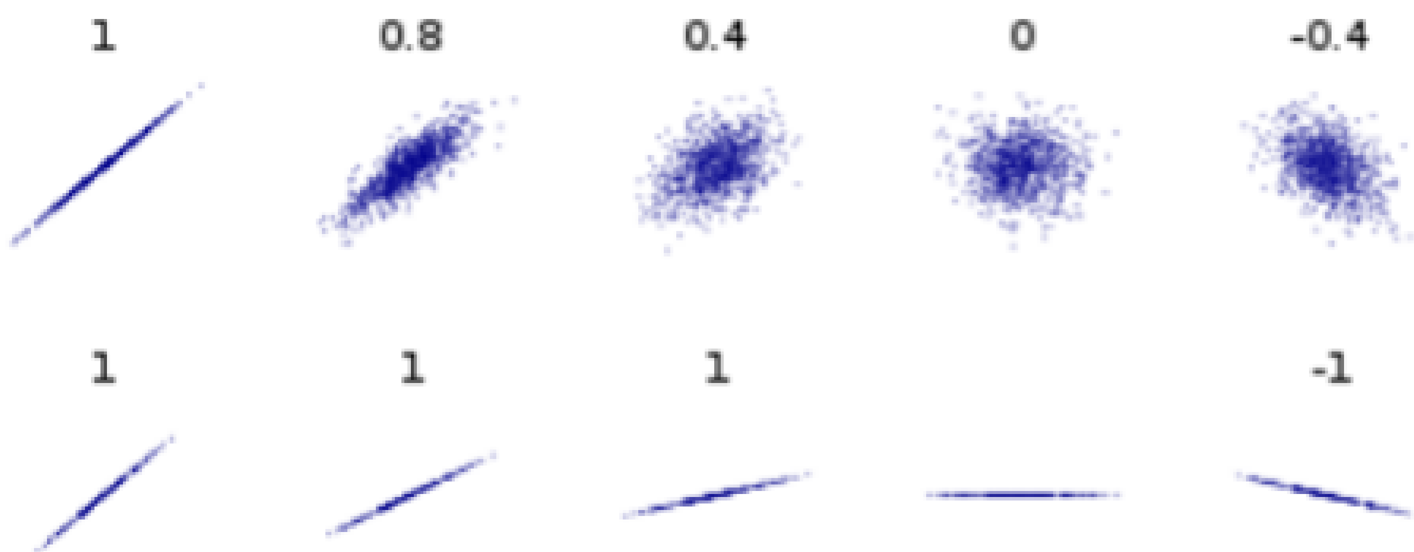


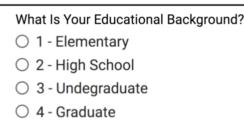
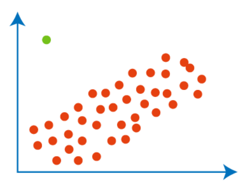
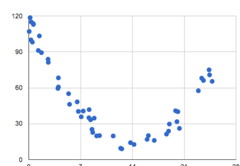
Figure 15.2: Korrelationskoeffizienten für verschiedene hypothetische Streudiagramme

- Ein Zusammenhang zweier Variablen kann extrem schwach sein (z.B. so, dass eine Verdopplung von X nur einer 0.1%-Steigerung von Y entspricht) und dennoch kann die Korrelation stark sein (= nah an ± 1), wenn die Punkte exakt auf einer Geraden liegen.
- Auf einen Satz gemünzt kann man sagen:

Die Pearson-Korrelation misst, wie gut bivariate Daten durch eine Gerade abgebildet werden können.

15.6 Voraussetzungen für das Berechnen der Pearson-Korrelation

- Die Pearson-Korrelation *kann immer* berechnet werden, solange beide Variablen aus Zahlenwerten bestehen.
- Es gibt jedoch weitere Kriterien, die für die Sinnhaftigkeit und Interpretierbarkeit der Pearson-Korrelation wichtig sind:

Kriterium	Falls Kriterium nicht erfüllt?	Beispiel	Mögliche Abhilfe?
Daten haben mindestens Intervallskalenniveau	Keine Aussage über Linearität des untersuchten Zusammenhangs möglich Korrelation nicht interpretierbar		Rangkorrelation
Keine Ausreißer	Korrelationskoeffizient kann massiv verzerrt sein		Ausreißer entfernen oder Rangkorrelation
Zusammenhang der Daten wird nicht durch *nicht-linearen* Anteil dominiert	Pearson-Korrelation falsches Modell Linearität der Daten wird verzerrt wiedergegeben, da die Korrelation vom nicht-linearen Teil beeinflusst wird		Komplexeres Modell, das den nicht-linearen Anteil berücksichtigt

15.7 Mythen zur Pearson-Korrelation

- In vielen Quellen finden sich darüber hinaus **unzutreffende Behauptungen** zur Verwendung der Pearson-Korrelation:

	Behauptung	Fact
Mythos 1	Die Variablen müssen kontinuierlich sein	**Pearson-Korrelation ist valide für diskrete Daten**, solange diese mindestens Intervallskalenniveau aufweisen. Tatsächlich gibt es sogar eine Variante der Pearson-Korrelation, bei der beide Variablen binär sind (Phi-Koeffizient).
Mythos 2	Die beiden Variablen müssen normalverteilt sein.	Die beiden Variablen müssen nicht normalverteilt sein. Korrekt ist, dass die Daten für die Berechnung eines p-Wertes **bivariat normalverteilt** sein sollten.
Mythos 3	Die Variablen müssen einen linearen Zusammenhang aufweisen	Gegenbeispiel: wenn Daten nur aus zufälligem Rauschen bestehen, sind sie hochgradig nicht-linear, aber dennoch gibt der Pearson-Koeffizient idR korrekterweise an, dass die Korrelation ungefähr 0 ist.Zusammenhänge in der Psychologie sind *sehr selten* eindeutig linear, dennoch kann es sinnvoll sein, die Pearson-Korrelation anzuwenden. Besser ist es daher zu sagen (wie in der Folie zuvor), dass der **Zusammenhang nicht zu stark durch einen nicht-linearen Anteil dominiert** werden sollten.
Mythos 4	Die Variablen müssen varianzhomogen sein	Varianzhomogenität (auch Homoskedastizität) meint, dass Y-Werte ähnliche Varianz in verschiedenen Abschnitten der X-Achse haben und umgekehrt. Jedoch ist Varianzhomogenität keine zwingende Voraussetzung für die Anwendung der Pearson-Korrelation. Korrekt ist, dass der erhaltene Korrelationskoeffizient **ungenauer** und der **p-Wert nicht mehr korrekt** ist.

15.8 Korrelationsmatrix

- Eine **Korrelation** bestimmt immer den Zusammenhang zwischen **zwei Variablen**
- Gibt es mehr als zwei Variablen (z.B. die "Big Five"), bietet sich eine Darstellung aller paarweisen Korrelationen an – die **Korrelationsmatrix**

Trait	O	C	E	A	N
Openness	1.00	.13	.35*	.23*	-.02
Conscientiousness		1.00	.37*	.10	-.40*
Extraversion			1.00	.01	-.26*
Agreeableness				1.00	-.10
Neuroticism					1.00

Nebendiagonalelemente

Diagonalelemente

*$p < .001$

Figure 15.3: Tabellarische Korrelationsmatrix. Sterne kennzeichnen häufig das Signifikanzniveau.

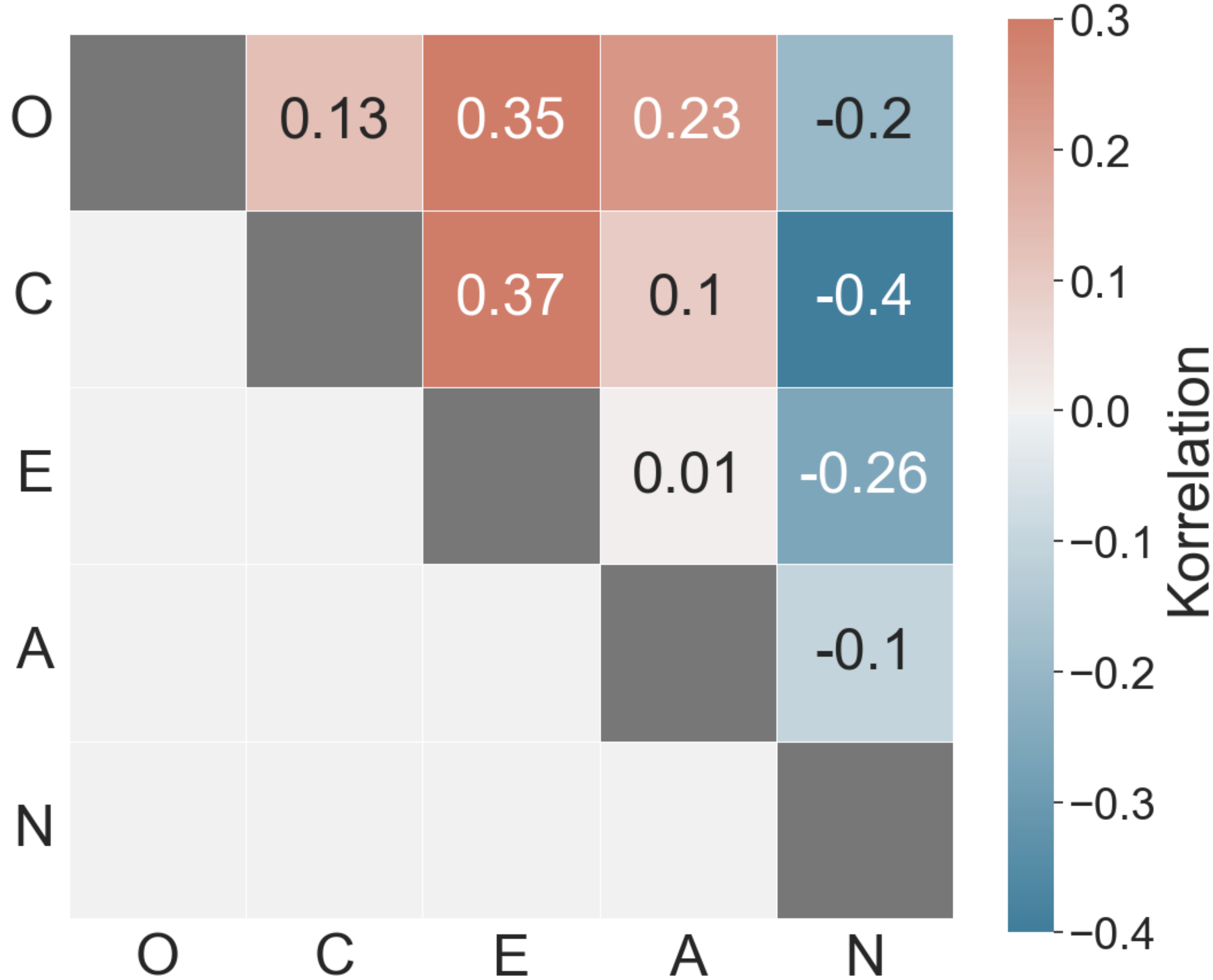


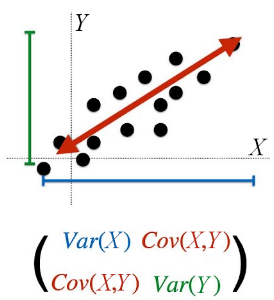
Figure 15.4: Korrelationsmatrix als “Heatmap” – die Einfärbung ist ein visuelles Hilfsmittel zur intuitiven und schnellen Erfassung der Korrelationsstruktur eines Variablen-Sets.

- Die Nebendiagonalelemente sind der interessante Teil der Korrelationsmatrix, sie geben die Korrelationen verschiedener Variablen an
- Die Diagonalelemente, also die Korrelationen von Variablen mit sich selbst, sind immer 1

15.9 Kovarianzmatrix

- Eine analoge Matrix-Darstellung gibt es auch für die **Kovarianz**
- Im Unterschied zur Korrelationsmatrix sind die Diagonalelemente der Kovarianzmatrix nicht 1, sondern geben die **Varianz** der Variable an.
- Sei $\mathbf{X}$ (beachte Fettschrift) ein Vektor von n Variablen $X_1..X_n$, so ist die zugehörige Kovarianzematrix $\text{Cov}(\mathbf{X})$:

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Var}(X_n) \end{pmatrix}$$



- Im Gegensatz zur Korrelationsmatrix ist die Kovarianzmatrix selten das "Endprodukt" einer Analyse, sondern meist ein Zwischenschritt in fortgeschritteneren statistischen Analysen wie der **Hauptkomponentenanalyse**

16 Rangkorrelationen

16.1 Was ist eine Rangkorrelation?

- **Rangkorrelationsmaße** ermöglichen es, Zusammenhänge von ordinalskalierten und nicht-linearen Variablem zu untersuchen.
  - Dies ist mit der Pearson-Korrelation nicht möglich
- "Rang" bezieht sich auf das Ordinalskalenniveau, d.h. dass die Daten nur als "Rang" oder "Reihenfolge" interpretiert werden können.
- Im Gegensatz zur Pearson-Korrelation bewerten Rangkorrelationen nicht die Linearität eines Zusammenhangs, sondern die **Monotonie des Zusammmenhangs**.
- Ein weiterer Anwendungsfall sind intervallskalierte Daten, die Ausreißer aufweisen und daher die Pearson-Korrelation verzerren.
  - Rangkorrelationen sind **robust gegenüber Ausreißern**
  - In diesem Fall werden die Variablen künstlich in Ränge umgewandelt.
- Hier behandeln wir zwei Maße für die Rangkorrelation:
  - **Spearman-Korrelation**
  - **Kendall'sches Tau**

Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

16.2 Berechnung von Rängen

- Liegen die Variablen nicht als Ränge vor, müssen sie zunächst in Ränge umgewandelt werden:
  1. Werte der Variable sortieren
  2. (Unnormierte) Ränge zuordnen
  3. Normierung: Gleiche Werte erhalten den Mittelwert ihrer Ränge
  4. (Optional) Variablen in ihre ursprüngliche Reihenfolge bringen

Ausgangswerte		Werte sortieren und Ränge bilden				(Optional) Rücksortieren		
Index	Wert	Index	Wert	Rang	Normiert	Index	Wert	Rang Normiert
1	1,5	5	1,0	1	1	1	1,5	2,5
2	1,5	1	1,5	2	$(2+3)/2$	2	1,5	2,5
3	4,0	2	1,5	3	$= 2{,}5$	3	4,0	5,0
4	3,0	4	3,0	4	4	4	3,0	4,0
5	1,0	3	4,0	5	5	5	1,0	1,0
6	5,0	6	5,0	6	$(6+7)/2$	6	5,0	6,5
7	5,0	7	5,0	7	$= 6{,}5$	7	5,0	6,5
8	9,5	8	9,5	8	8	8	9,5	8,0

- Die Tabelle gibt die Rangberechnung *einer* Variablen an (z.B. X) – für die andere Variable (Y) muss das analoge Prozedere durchgeführt werden.

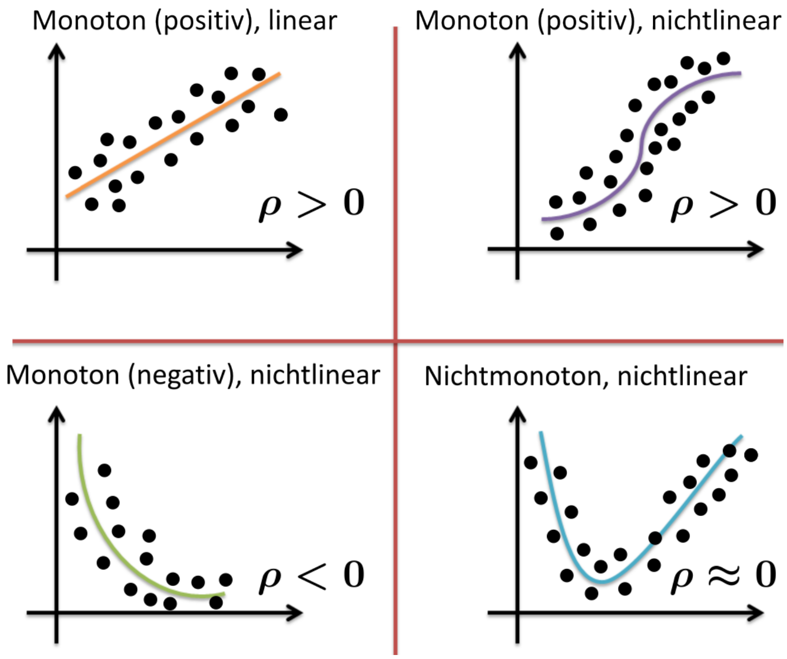
16.3 Spearman-Korrelation

- Die Spearman-Korrelation ist identisch zur Pearson-Korrelation, wenn die Variablen X und Y als Ränge $R(X)$ und $R(Y)$ vorliegen (Spearman wird häufig mit dem Buchstaben ρ bezeichnet):

$$\rho_{XY} = \frac{cov(R(X), R(Y))}{s_{R(X)} s_{R(Y)}}$$

wobei $s_{R(X)}$ und $s_{R(Y)}$ die Standardabweichungen der Ränge von X und Y sind.

- Wie die Pearson-Korrelation nimmt die Spearman-Korrelation Werte zwischen -1 und $+1$ an:
  - Ein positiver Wert impliziert eine positiven monotonen Zusammenhang
  - Ein negativer Wert impliziert eine negativen monotonen Zusammenhang
  - Ein Wert nahe bei 0 impliziert einen schwachen (oder keinen) monotonen Zusammenhang



16.4 Kendalls Tau

- Eine Alternative Rangkorrelation zu Spearman ist **Kendalls Tau**
- Kendalls Tau vergleicht inwieweit die Rangfolge *aller* Paare x_i, x_j mit der Rangfolge *aller* Paare y_i, y_j übereinstimmt
- Dazu wird die Zahl der konkordanten (übereinstimmenden) und diskordanten (nicht übereinstimmenden) Paare gezählt

$$\tau = \frac{K - D}{K + D}$$

K ist die Anzahl der konkordanten Paare und **D** die Anzahl der diskordanten Paare.

Beispiel

$$X = \begin{pmatrix} 9 \\ 3 \\ 7 \\ 5 \end{pmatrix} Y = \begin{pmatrix} 18 \\ 7 \\ 8 \\ 21 \end{pmatrix}$$

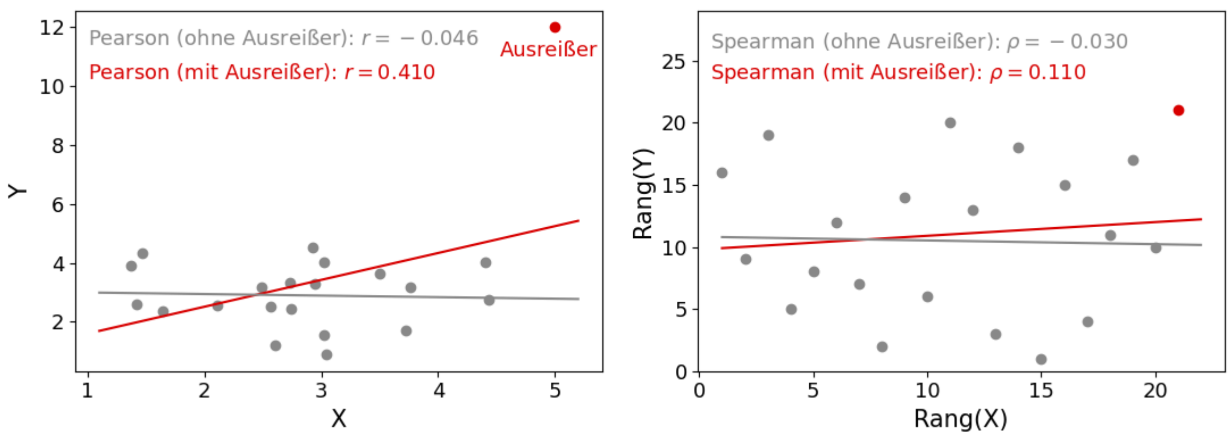
- Die Paare von X sind: $(9,3),(9,7),(9,5),(3,7),(3,5),(7,5)$
- Die Paare von Y sind: $(18,7),(18,8),(18,21),(7,8),(7,21),(8,21)$
- Die Paare $(x_1 = 9, x_3 = 7)$ und $(y_1 = 18, y_3 = 2)$ wären konkordant, da die Rangfolge des X-Paares $(x_1 > x_3)$ gleich der Rangfolge des entsprechenden Y-Paares $(y_1 > y_3)$ ist
- Die Paare$(x_1 = 9, x_4 = 5)$ und $(y_1 = 18, y_4 = 21)$ wären diskonkordant, da die Rangfolge des X-Paares $(x_1 > x_4)$ ungleich der Rangfolge des entsprechenden Y-Paares $(y_1 < y_4)$ ist

Insgesamt gibt es im Beispiel 4 konkordante Paare und 2 diskordante Paare (prüfe nach!), daher gilt:

$$\tau = \frac{K - D}{K + D} = \frac{4 - 2}{3 + 3} = \frac{2}{6} = 0.333..$$

16.5 Rangkorrelationen: Robust gegen Ausreißer

- Beispiel Pearson vs. Spearman:



- Ein einziger Ausreißer verändert die Pearson-Korrelation im Beispiel von $r = -0.046$ nach $r = 0.410$
- Demgegnüber ist die Spearman-Korrelation "robuster" gegenüber dem Ausreißer – sie verändert sich "lediglich" von $\rho = -0.030$ nach $\rho = 0.110$.
- Intution: während der *Wert* es Ausreißers deutlich über dem zweithöchsten Y-Wert liegt, ist der *Rang* nur um 1 höher.

16.6 Wann Spearman und wann Kendall?

- Beide Rangkorrelationskoeffizienten bestimmen die Monotonie eines Zusammenhangs
- **Kendall ist robuster bei kleinen Stichproben** und ist in diesen Fällen bevorzugt
- **Spearman ist etwas weniger sensitiv gegenüber Rangbindungen** (also wenn zwei Werte den gleichen Rang haben) und ist daher bevorzugt, wenn es viele Rangbindungen gibt[1]

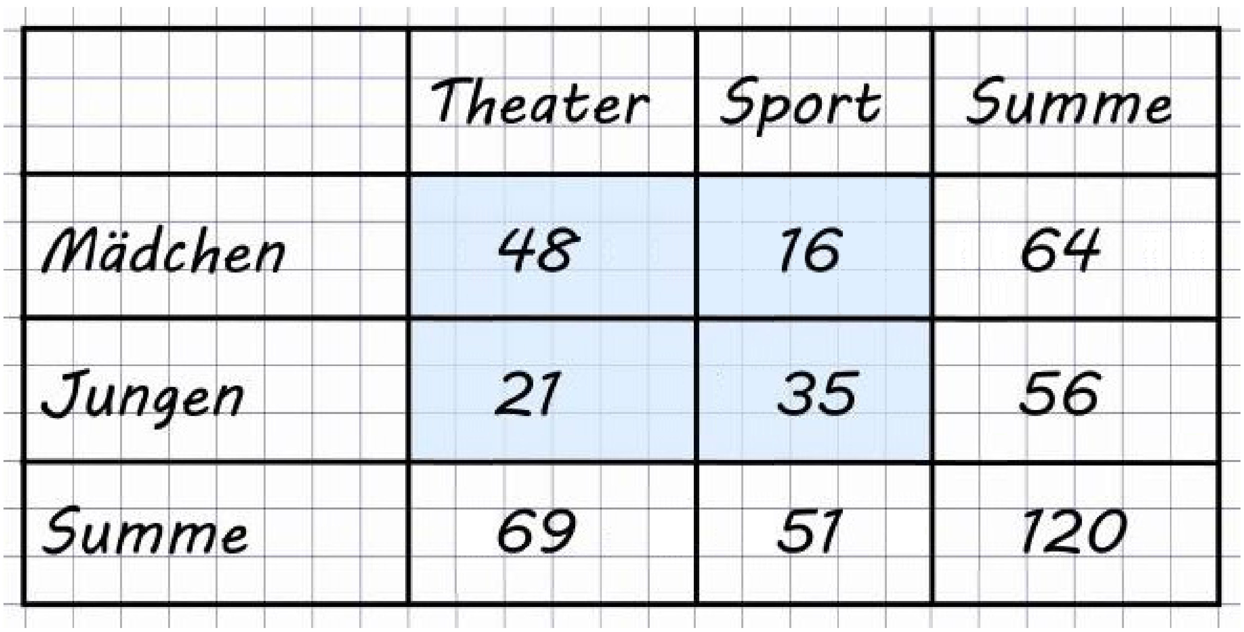
[1] [redacted]Puth association between two measured traits. Animal Behaviour 102:77–84.

- Beachte: auch bei den Kendall'schen Paaren gibt es Rangbindungen, und zwar dann, wenn die verglichenen *Paare* zwischen X und Y genau identisch sind
- Diese Paare sind weder konkordant noch diskordant und es gibt verschiedene Algorithmen diese Fälle zu berücksichtigen (hier nicht behandelt)

- **In Abwesenheit von Rangbindungen liefert Kendall präzisere Schätzungen** und ist Kendall zu bevorzugen.[2]
- In der Statistik wird Kendall häufig als "Default"-Rangkorrelation empfohlen[3]:
  - idR präzisere Schätzung des Populationsparameters
  - Standardfehler ist bekannt (für Spearman gibt es lediglich Approximationen[4])
- Faktisch ist aber Spearman der weitaus verbreitetere Korrelationskoeffizient – womöglich weil er idR größer als der Kendall-Koeffizient ist

16.7 Der Phi-Koeffizient

- Spezialfall: **beide Variablen haben nur zwei Ausprägungen**
- Darstellbar in der **Vierfeldertafel**:



	Theater	Sport	Summe
Mädchen	48	16	64
Jungen	21	35	56
Summe	69	51	120

Figure 16.1: Bildnachweis[5]

- In die vier Felder werden die Häufigkeiten der jeweiligen Variablen-Kombination eingetragen (im Beispiel blaugefärbt)
- Optional: In der letzten Zeile/Spalte die Summe

- Wie alle Korrelationskoeffizienten beantwortet der **Phi-Koeffizient** die Frage, ob zwei Variablen miteinander zusammenhängen

[2] [redacted]
association between two measured traits. Animal Behaviour 102:77–84.
[3] [redacted]

- Im Beispiel: hängt das gewählte Fach vom Geschlecht ab?

$$\text{Formel:} \quad Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- Auch der Phi-Koeffizient hat einen Wertebereich von $[-1; 1]$

	X_1	X_2	Σ
Y_1	a	b	a+b
Y_2	c	d	c+d
Σ	a+c	b+d	N

- NB: der Phi-Koeffizient ist identisch mit der Pearson-Korrelation und der Spearman-Korrelation, wenn beide dichotomen Variablen mit 0 und 1 kodiert werden.

16.8 Phi-Koeffizient: Beispiel

		Haben Sie ein Haustier ?		
		ja	nein	Gesamt
Geschlecht	männlich	2	8	10
	weiblich	12	11	23
Gesamt		14	19	33

$$Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \tag{16.1}$$

$$= \frac{2 \cdot 11 - 8 \cdot 12}{\sqrt{(2+8)(12+11)(2+12)(8+11)}} = \tag{16.2}$$

$$= \frac{22 - 96}{\sqrt{10 \cdot 23 \cdot 14 \cdot 19}} = \frac{-74}{\sqrt{61180}} = -0.299 \tag{16.3}$$

17 Interpretation von Korrelationen

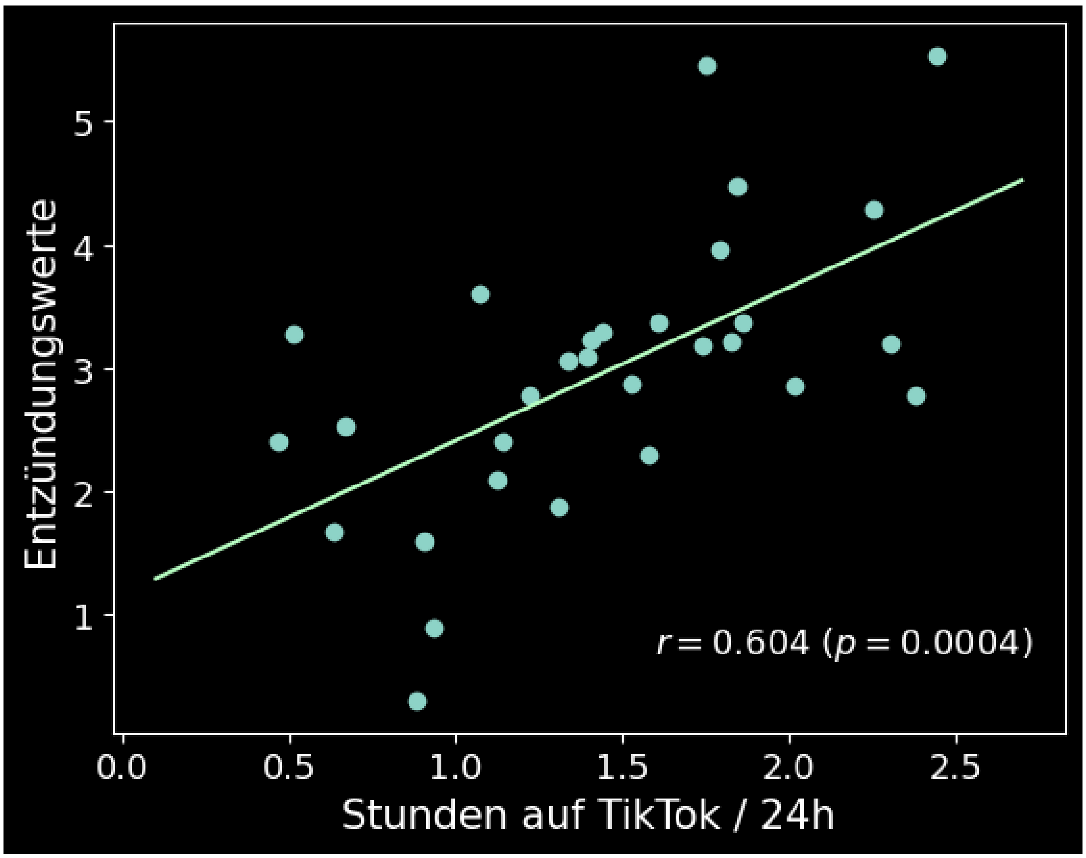
17.1 Korrelation und Kausalität

Die Task Force ist immer noch einigermaßen perblex, ob des Zusammenhangs von TikTok-Onlinezeit und den Entzündungswerten.

- Als erfahrenen Statistiker:innen kennen Sie die **erste Regel der Korrelation**:

Korrelation ist nicht gleich Kausalität

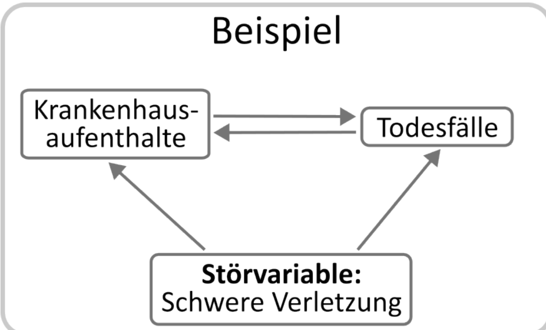
(Oder falls Sie den Lateiner heraushängen lassen wollen: *Cum hoc ergo propter hoc*)

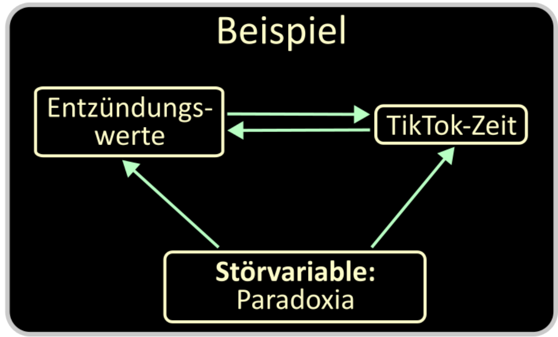


- Es scheint unmöglich, dass virutelle TikTok-Zeit einen **kausalen Einfluss** auf biologische Entzündungswerte hat.
- Sie wissen, dass ein häufiger alternativer Grund für eine Korrelation eine dritte Variable ist, die wiederum Einfluss auf beide Variablen der Korrelation hat— eine **Störvariable**.
- In diesem Fall vermuten Sie, dass die Störvariable eine vorhandene Paradoxie-Erkrankung selbst ist: die Erkrankung führt einerseits zu höheren Entzündungswerten und andererseits zu vermehrter TikTok-Zeit (um Trost bei Leidensgenossen zu erfahren?)

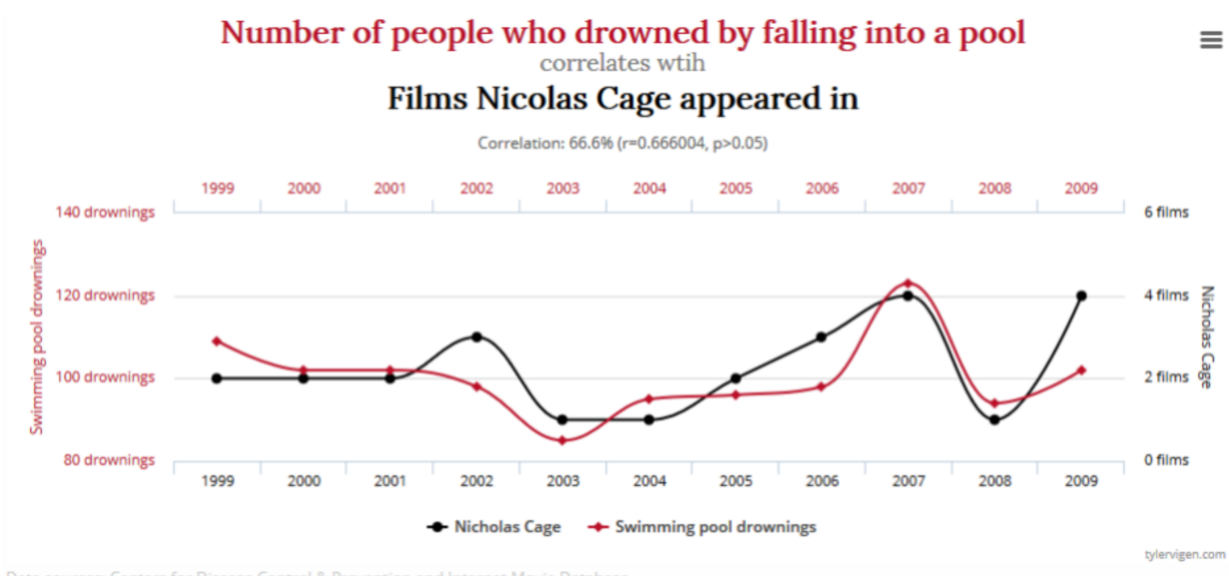
17.2 Korrelation und Kausalität

- Rein rechnerisch bedeutet eine Korrelation einen Zusammenhang zweier Variablen, sogar, dass man die eine Variable aus der anderen statistisch vorhersagen kann
- Dies **bedeutet jedoch nicht, dass sich die Variablen auch kausal bedingen**
- Einerseits **können Korrelation zufällig zustande kommen** (die Wahrscheinlichkeit von solchen irrtümlichen Befunden untersuchen wir noch genauer beim Thema Signifikanztestung)
- .. andererseits könen Korrelationen durch dritte Variablen (**Störvariablen**) verursacht sein.

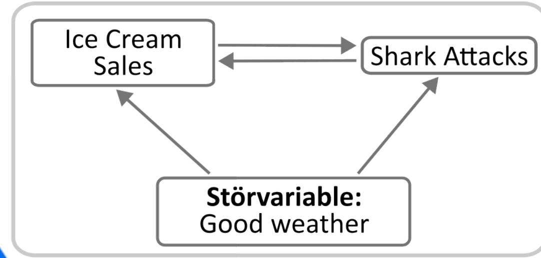


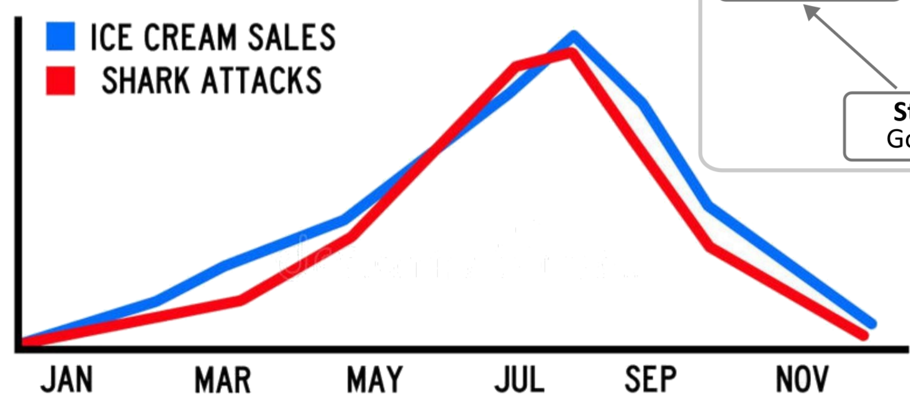


17.3 Korrelation und Kausalität



17.4 Korrelation und Kausalität





Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

17.5 Wann ist ein Korrelationskoeffizient groß oder klein?

- Pauschal schwer zu beantworten
- In der psychologischen Literatur hat sich folgende Nomenklatur nach Jacob Cohen[1] eingebürgert:

Korrelation (r)	Nomenklatur
0.1-0.29	"kleiner" Effekt
0.3-0.49	"mittlerer" Effekt
0.5-1	"großer" Effekt

- Allerdings fügt Cohen im gleichen Artikel hinzu:

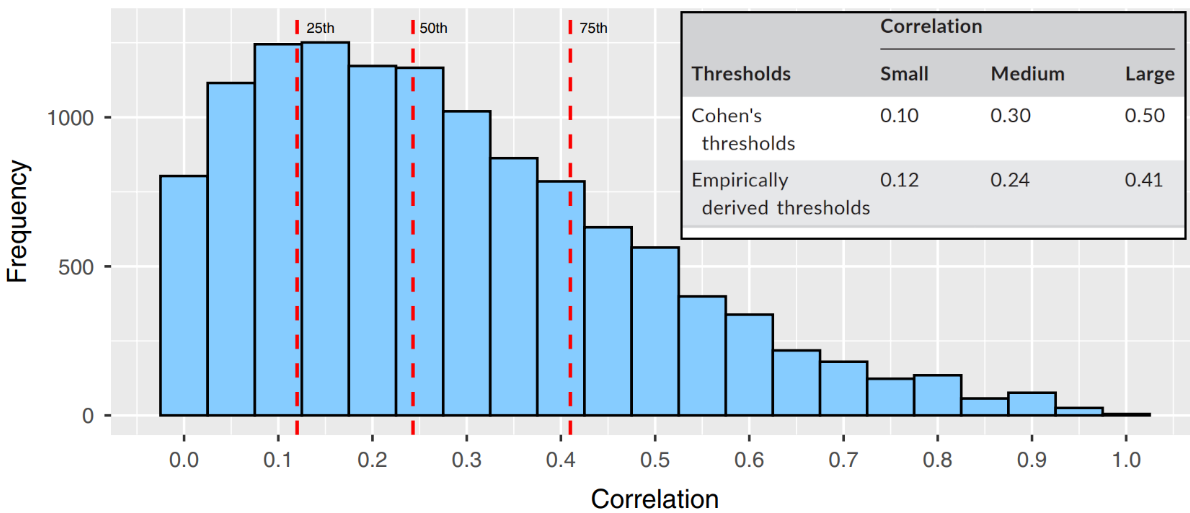
> These proposed conventions were set forth throughout with much diffidence [Zurückhaltung], qualifications [Bedingungen], and invitations **not to employ** them if possible.

- Bei der Beurteilung sollte der Korrelationskoeffizient, wie jede Effektgröße, immer in Relation zu typischen Werten im jeweiligen Forschungsfeld gesetzt werden.

[1] [redacted]Cohen

17.6 Wann ist ein Korrelationskoeffizient groß oder klein?

- In einer kürzlichen Metaanalyse von Lovakov Agadullina (2021)[2] wurden typische Effektstärken in der Sozialpsychologie untersucht:



- Erkenntnis: "mittlere" und "starke" Korrelationen (definiert als das 50%- und 75%- Quantil) sind in der Realität kleiner als von Cohen angenommen

[2] [redacted] Lovakov J Soc Psychol 51:485–504.

17.7 [[Zusammenfassung]]

- Zusammenhänge werden mit **Kovarianz** und **Korrelation** untersucht
- Die **Kovarianz** ist die Basis der Pearson-Korrelation – aber sie ist abhängig von den gewählten Einheiten
- Die **Pearson-Korrelation** misst die Linearität eines Zusammenhangs und gilt für intervallskalierte Daten
- Für ordinalskalierte Variablen eignen sich **Rangkorrelationen**: **Spearmans Rho** und **Kendalls Tau** – sie messen die Monotonie eines Zusammenhangs
- Sind beide Variablen dichotom, greift der **Phi-Koeffizient**
- **Correlation does not imply causation, Correlation does not imply causation, Correlation does not imply causation, Correlation does not**

Part V

Vorlesung 05: Regression

18 Vorlesung 05: Regression

18.1 Woher kommt der Ausdruck "Regression"?

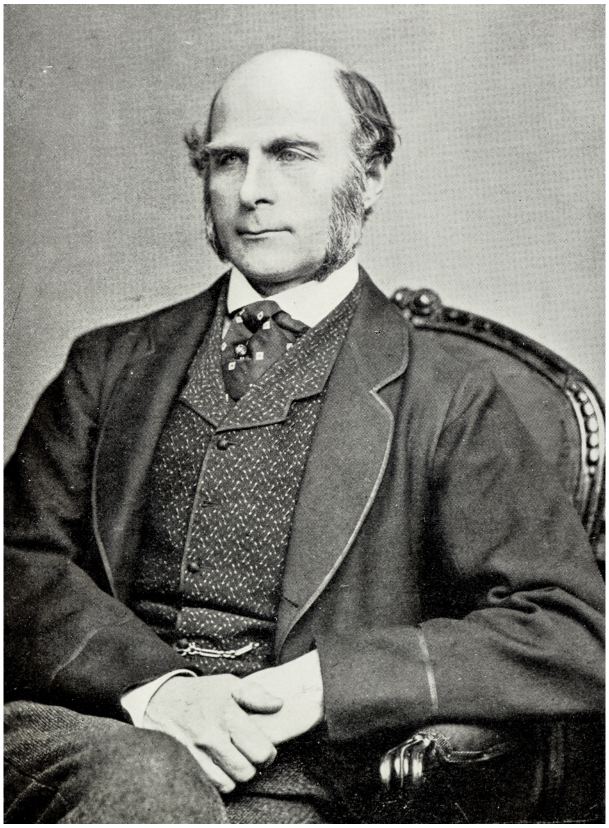
- Lateinisch »regredi« = „umkehren, zurückgehen“
- Psychoanalyse: Regression = Zurückfallen in kindliche Verhaltensmuster

> Wir heißen es **Regression**, wenn sich im Traum die Vorstellung in das sinnliche Bild zurückverwandelt, aus dem sie irgend einmal hervorgegangen ist.

Sigmund Freud (1900). "Traumdeutung".

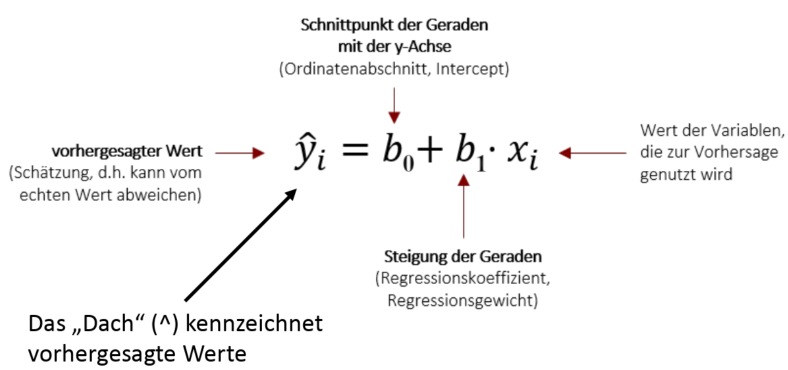
- In die Statistik wird der Ausdruck "Regression" klassischerweise auf **Francis Galton** (Cousin von Charles Darwin) attribuiert, der bereits 1885 ein Phänomen beschrieb, das er *regression toward mediocrity* (**Regression zur Mitte**) taufte
- Das Phänomen bestand darin, dass Nachfahren großer Eltern dazu tendieren, selbst nur durchschnittlich groß zu werden
- Neuere Forschung zeigt allerdings, dass sich Galton selbst wohl noch nicht des statistischen Ursprungs dieses Phänomens bewusst war und eine biologische Erklärung favorisierte[1]

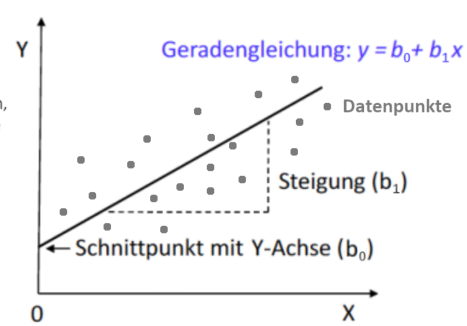
[1] [redacted] Krashniak and Philosophy of Science Part A 86:6–19.



18.2 Regression

- Dem Wortsinn nach ist Ziel der **Regression** eine abhängige Variable auf eine oder mehrere unabhängige Variablen zurückzuführen (auf diese zu *regredieren*)
- Eingängiger ist aber die umgekehrte Formulierung: Ziel der Regression ist es, auf Basis der unabhängigen Variablen die eine abhängige Variable **vorherzusagen** oder **zu erklären**
  - **Abhängige Variable = vorher*sagende* oder erklär*ende* Variable**
  - **Unabhängige Variable(n) = vorher*gesagte* oder erklär*te* Variable(n)**

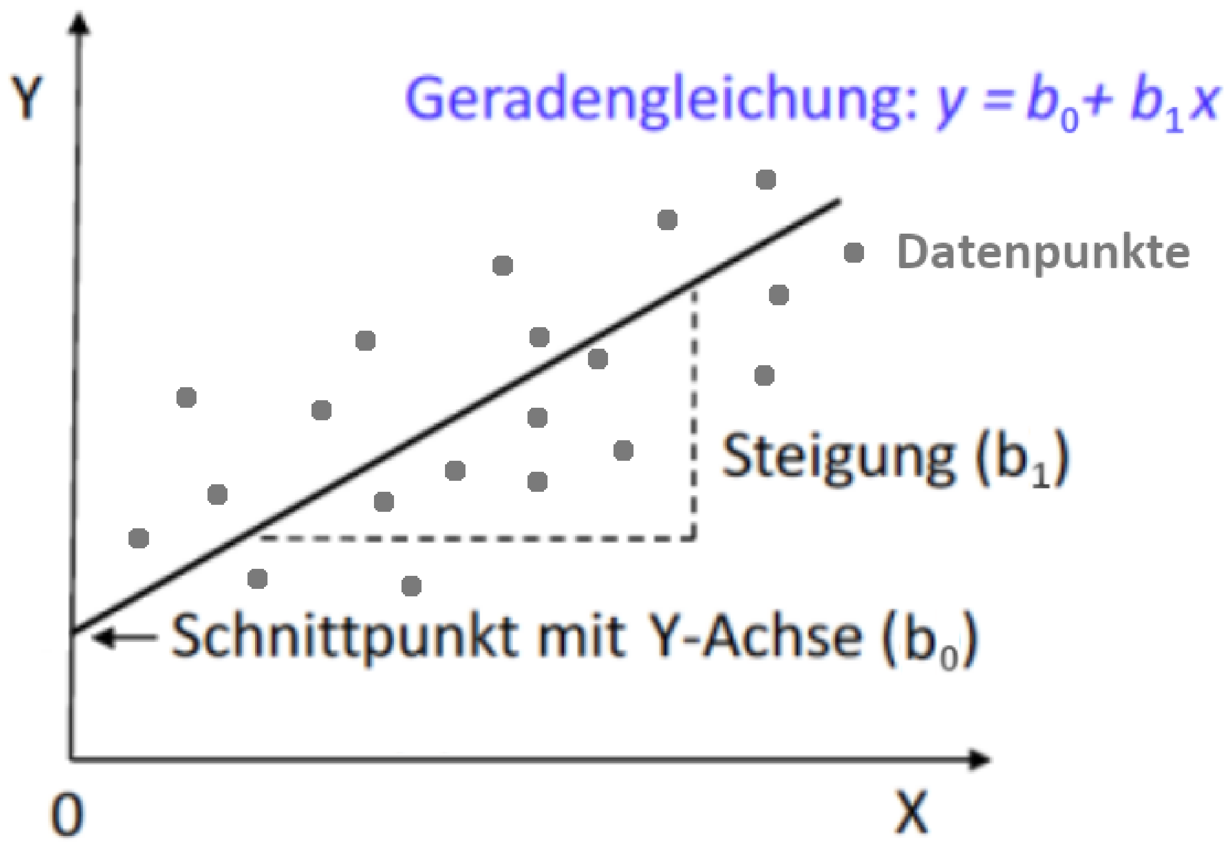




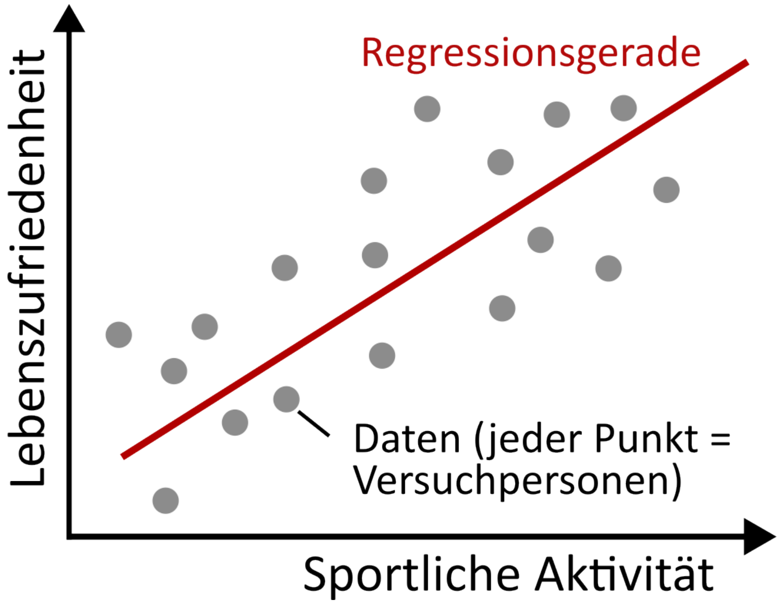
- Beispiel: eine Studie untersucht, ob sich Lebenszufriedenheit auf Basis von sportlicher Aktivität vorhersagen/erklären lässt.
  - Lebenszufriedenheit: unabhängige/vorhersagende/erklärende Variable; — sportliche Aktivität: abhängige/vorhergesagte/erklärte Variable

18.3 Regression

- Im Gegensatz zur Korrelation bestimmt die Regression nicht die Linearität des Zusammenhangs (vielmehr wird dies vorausgesetzt), sondern die **Steigung** des Zusammenhangs
- Aus diesem Grund ist die Regression (wieder im Gegensatz zur Korrelation) nicht symmetrisch – die Steigung ist abhängig davon welche Variable als abhängig und unabhängig deklariert wird.
  - Aus mathematisch etwas komplizierteren Gründen ist es nicht einmal gestattet die Regressionsgleichung zu invertieren ($\hat{x}_i = \frac{1}{b_1} y_i - \frac{b_0}{b_1}$) — im Allgemeinen ist $\frac{1}{b_1}$ *nicht* die Steigung, wenn die Rollen von X und Y vertauscht werden.

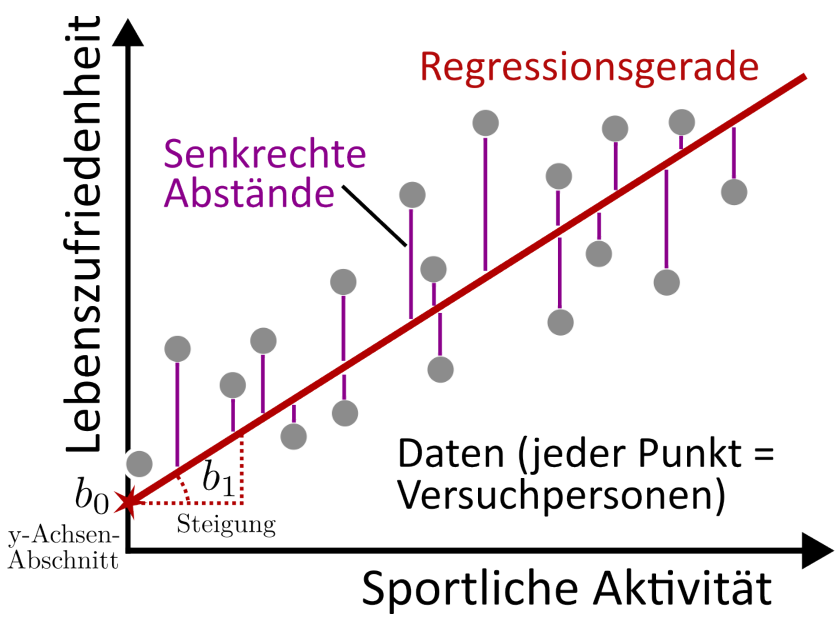


- Die Vorhersage/Erklärung von X durch Y geschieht durch eine Gleichung – die **Regressionsgleichung** – die im Streudiagramm als Gerade eingezeichnet werden kann.



18.4 Bestimmung der Regressionsgerade: Methode der kleinsten Quadrate

- Ziel der Regression ist es, die Gerade zu finden, die die Datenpunkte möglichst gut abbildet — es gibt jedoch verschiedene Definitionen dessen, was "möglichst gut" heißt
- Die häufigste Variante ist die **Methode der kleinsten Quadrate**, bei der die Gerade so gewählt wird, dass die **Summe der quadrierten senkrechten Abstände jedes Datenpunktes zur Geraden** minimal ist
  - Engl. *ordinary least square*



- Die **einfache Regression** mit nur einer unabhängigen Variablen hat zwei freie Parameter, um die Gerade an die Datenpunkte anzupassen (zu "fitten"):

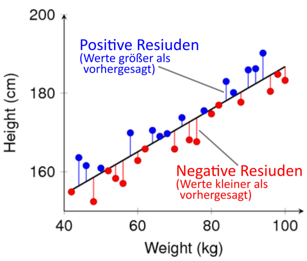
- **y-Achsenabschnitt** b_0 (engl. *intercept*)
- **Steigung** b_1 (engl. *slope*)

- Exakt 0 wären die senkrechten Abstände nur, wenn alle Punkte auf einer perfekten Gerade liegen. Dies ist eigentlich nie der Fall. Die verbleibenden senkrechten Abstände der Datenpunkte von der gefitteten Geraden werden **Residuen** genannt.

18.5 Warum weichen die Datenpunkte überhaupt von einer Geraden ab?

Verschiedene Gründe:

- Variablen korrelieren überhaupt nicht
- Einfluss von Störvariablen
- Messungenauigkeit



In der Psychologie gibt es (bis auf triviale Fälle) keine perfekten linearen Zusammenhänge, d.h. es verbleiben immer **Residuen** $\Delta\hat{y}_i$:

$$\text{Residuum:} \quad \Delta\hat{y}_i = \hat{\epsilon}_i = \hat{y}_i - y_i$$

Residuum = Differenz von vorhergesagtem Wert $\hat{y}_i$ und tatsächlichem Wert y_i

18.6 Totale, erklärte und Residuenquadratsumme

- Die Methode der kleinsten Quadrate minimiert die **Residuenquadratsumme (SQR)**:

$$SQR = \sum (\hat{y}_i - y_i)^2 = \sum \hat{\epsilon}_i^2$$

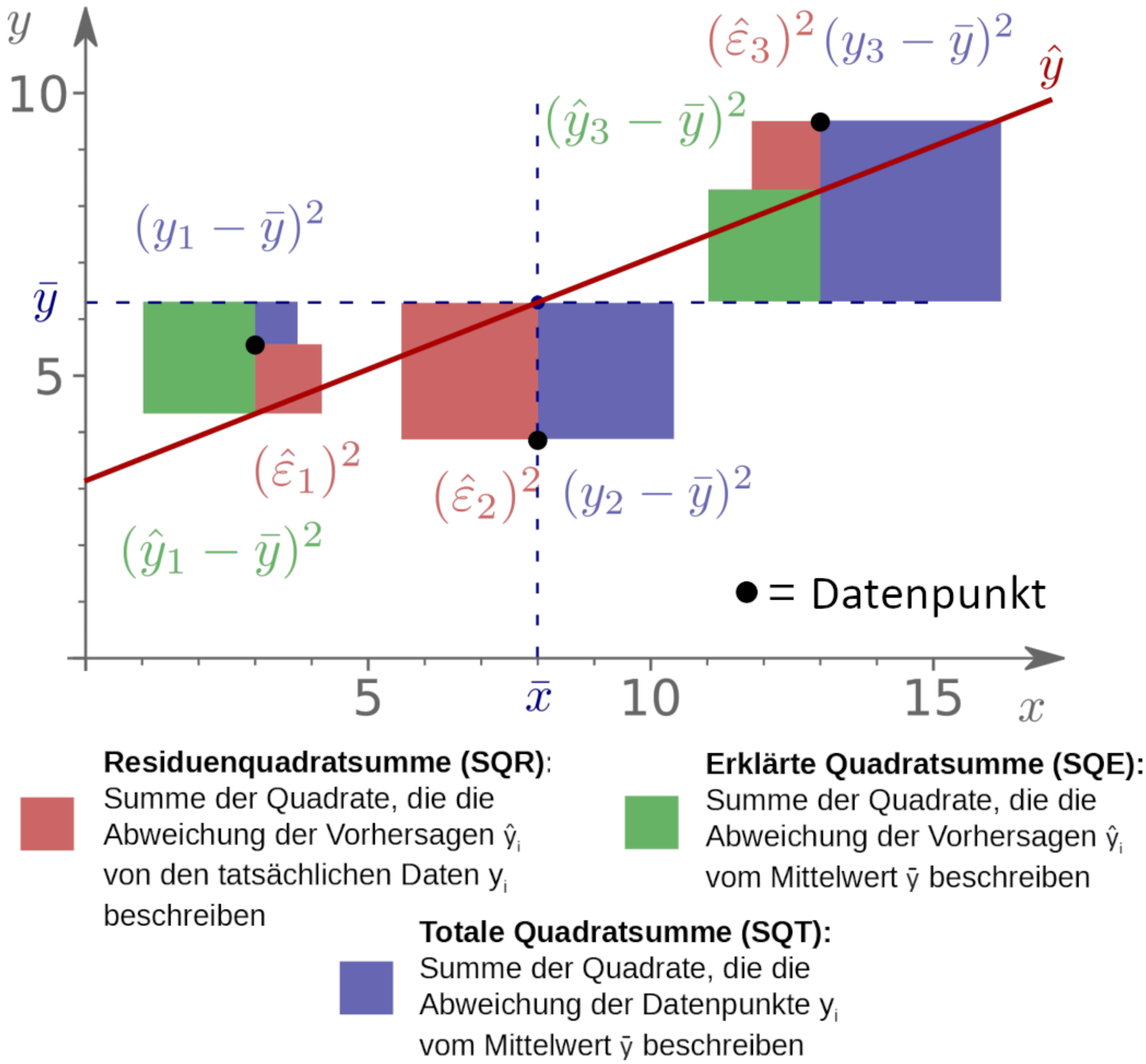
- Diese wiederum lässt sich in Bezug setzen zur **totalen Quadratsumme (SQT)** und zur **erklärten Quadratsumme (SQE)**:

$$SQR = SQT - SQE \qquad \text{bzw.}$$
$$SQT = SQT + SQR$$

- Mit

$$SQE = \sum \left(\hat{y}_i - \bar{Y}\right)^2$$

$$SQT = \sum \left(y_i - \bar{Y}\right)^2$$



- Falls alle Punkte exakt auf der Regressionsgeraden liegen ($\hat{y}_i = y_i$) ist die erklärte Quadratsumme identisch der totalen Quadratsumme ($SQE = SQT$) und die Residuenquadratsumme ist 0

18.7 Bestimmtheitsmaß

- Das **Bestimmtheitsmaß** R^2 gibt an, wie gut die Datenpunkte durch die Regressionsgerade gefittet werden ("Anpassungsgüte")

- Es gibt an, welcher Anteil der Datenvarianz $Var(y)$ durch die Varianz der Vorhersage $Var(\hat{y})$ erklärt wird..

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\sum \left(\hat{y}_i - \bar{Y}\right)^2}{\sum \left(y_i - \bar{Y}\right)^2} = \frac{SQE}{SQT}$$

- .. oder äquivalent, den Anteil der erklärten Quadratsumme an der totalen Quadratsumme.

18.8 Vorhersage versus Erklärung

References