

ASTAT-Zusammenfassung

DATA SCIENCE	6
Wozu brauchen wir Data Science?	6
Was ist das Ziel von Data Science?	6
In welchen Bereichen wird Data Science angewendet?	7
Anwendungen nach Art der Tätigkeit	7
Welche Praxisbeispiele für Data Science gibt es?	7
Digitale vs. Produzierende Unternehmen	8
Produktionsoptimierung (Konkretes Beispiel)	8
Beispiele aus der Schulmathematik	9
Alltagsprobleme	10
Vier-Schritte-Problemlösungsstrategie nach Mazur	10
Erster Schritt	10
Zweiter Schritt	10
Dritter Schritt	10
Vierter Schritt	10
Was ist Statistik, die nicht angewandt ist?	11
DESKRIPTIVE STATISTIK (EINDIMENSIONAL)	11
Daten	11
Datensatz (eindimensional)	11
Datensätze (zweidimensional)	11
Deskriptive Statistik	12
Ziele der deskriptiven Statistik	12
Bezeichnungen	12
Kennzahlen	12
Arithmetisches Mittel	12
Streuung graphisch	13
Empirische Varianz und Standardabweichung	13
Median	13
Median vs. Arithmetisches Mittel	13
Quartile	14
Quantile	14
Boxplot	15
Vergleich von Datensätzen	16

ZWEIDIMENSIONALE DESKRIPTIVE STATISTIK	16
Histogramm	16
Beispiel IQ-Test	17
Wahl der Klassen	17
Beispiel Waage	19
Bimodales Verhalten	19
Schiefe von Histogrammen	20
Normiertes Histogramm	20
Vorteil von normierten Diagrammen	21
Schiefe im Boxplot	22
Boxplot Bemerkungen	22
Deskriptive Statistik zweidimensionaler Daten	23
Beispiel Mortalität und Weinkonsum in Ländern	23
Beispiel Old Faithful	24
Erkenntnisse	24
Abhängigkeit und Kausalität	24
Lineare Regression	25
Streudiagramm und Regressionsgrade	25
Residuum	26
Methode der kleinsten Quadrate	27
KORRELATION WAHRSCHEINLICHKEITSMODELL	29
Wie gut passt die Regressionsgerade?	29
Empirische Korrelation	29
Begründung	29
Beispiel	30
Berechnung in R	33
Bemerkungen	33
Wahrscheinlichkeit	34
Wahrscheinlichkeitsmodell	34
Ereignis	35
Neue Mengen aus Bekannten	36
Operationen der Mengenlehre für Ereignisse	36
Graphisch	36
Beispiel: Würfelwurf	36
Axiome der Wahrscheinlichkeit	38
Rechenregeln aus Axiomen	39

Knobelaufgabe	40
Diskrete Wahrscheinlichkeitsmodelle	41
Modell von Laplace	42
Stochastische Unabhängigkeit	43
ZUFALLSVARIABLE WAHRSCHEINLICHKEITSVERTEILUNG	45
Definition	45
Beispiel Zufallsvariable	45
Wahrscheinlichkeitsverteilung eine Zufallsvariablen	49
Wahrscheinlichkeitsverteilung	50
Definition	50
Definition	52
Erwartungswert	52
Varianz und Standardabweichung	52
BEDINGTE WAHRSCHEINLICHKEIT	57
Bedingte Wahrscheinlichkeit	60
Bayes Theorem	66
Gesetz der totalen Wahrscheinlichkeit	67
Gesetz	67
NORMALVERTEILUNG	82
Definitionen	82
Intervalle	82
Beispiel	82
Punktwahrscheinlichkeit 0	82
Eigenschaften Wahrscheinlichkeitsdichte	83
Quantile	84
Normalverteilung (Gaussverteilung): $X \sim N(\mu, \sigma^2)$	85
GESETZ DER GROSSEN ZAHLEN, ZENTRALER GRENZWERTSATZ	89
Funktion von mehreren Zufallsvariablen	89
Beispiel	89
Summe und Durchschnitt	89
Beispiel: Warum ist der Durchschnitt wichtig?	89
Kennzahlen von S_n und X_n	90
Graphisches Beispiel	90
Feststellung	91
Feststellung	91
Allgemein	92
Bemerkungen	92

Standardfehler	92
Zentraler Grenzwertsatz	92
Zentraler Grenzwertsatz	101
Beispiel	101
Beispiel 2	102
HYPOTHESENTEST	104
Beispiel Vorgehen Hypothesentest	110
Graphische Darstellung	113
Bemerkungen	118
Signifikanzniveau	118
Beispiel Abfüllanlage	119
Beispiel: Körpergrösse Frauen	121
Einfluss der Anzahl Messungen auf Verwerfungsbereich	124
P-Wert	126
P-Wert für zweiseitigen Test	128
t-Test	129
VERTRAUENSINTERVALL, ZWEISTICHPROBENTEST, WILCOXON-TEST	136
Vertrauenintervall	136
Interpretation Vertrauensintervall	140
Bemerkungen	145
Nicht-Normalverteilte Daten: Wilcoxon-Test	145
Beispiel: Waage A	146
Wilcoxon-Test vs. t-Test	146
Vergleich von zwei Stichproben: Mögliche Fragestellungen	147
Gepaarte Stichproben	147
Ungepaarte (unabhängige) Stichproben	147
Unterscheidung gepaarte vs. ungepaarte Stichproben	148
Gepaarte versus ungepaarte Stichproben	148
Statistischer t-Test für gepaarte Stichproben mit	149
Statistischer t-Test für ungepaarte Stichproben	150
Mann-Whitney U-Test (aka Wilcoxon Rank-sum Test)	151
LINEARE REGRESSION	151
Beispiel: Einkommen	155
Warum soll f geschätzt werden?	156
Beispiel	156
Rückschlüsse auf f: Fragestellungen	158
Fragen für Beispiel der Werbung	159
Schätzung von f?	160
Beispiel	160
Einfaches Regressionsmodell	162

Schätzung der Parameter	164
MULTIPLE LINEARE REGRESSION	178
Bestimmung der wichtigen erklärenden Variablen	192
Wie gut passt das Modell zu den Daten?	193
QUALITATIVE VARIABLEN	196
Qualitative erklärende Variablen	196
Qualitative erklärende Variable mit nur zwei Levels	199
Qualitative erklärende Variablen mit mehr als zwei Levels	201
Bemerkungen	203

Data Science

Data Science ist per Definition die Schnittmenge zwischen den wissenschaftsbereichen Mathematik, Informatik, sowie dem branchenspezifischen Fachwissen. Der Bereich der Datenwissenschaften befasst sich mit:

- Der Analyse von grossen Datenmengen
- Der Identifizierung von Anomalien in den Daten
- Sowie mit der Vorhersage von zukünftigen Ereignissen

Die im Arbeitsbereich Datenwissenschaft arbeitenden Personen werden als Datenwissenschaftler oder Data Scientist bezeichnet.

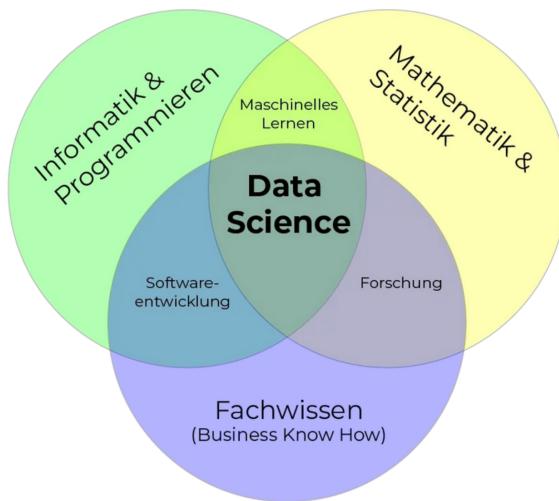


Abbildung 1: Data Science ist die Schnittmenge zwischen den Wissenschaftsbereichen Mathematik, Informatik, sowie dem branchenspezifischen Fachwissen

Wozu brauchen wir Data Science?

Durch die Globalisierung wächst die Konkurrenz auf heimischen und ausländischen Märkten. Neue Trends wie Mass Customization und neue digitale Geschäftsmodelle setzen klassische Unternehmen zunehmend unter Leistungsdruck. Produzierende Unternehmen begegnen den Herausforderungen durch die Optimierung der Produktion oder Erweiterung der Geschäftsfelder.

Was ist das Ziel von Data Science?

Das Ziel der Datenwissenschaft ist es anhand von grossen Datenmengen, die Entscheidungsfindung im Unternehmen zu optimieren. Dies können klassischerweise Kosten- oder Umsatzoptimierungen sein. Oft sollen Entscheidungen auch automatisch vom Machine Learning Algorithmus getroffen werden können (Prozessautomatisierung).

Ziel	Funktion Data Science
Informationen filtern	Datenbestand so analysieren, dass wertvolle Informationen ersichtlich werden.
Handlungsempfehlungen geben	Durch die Analyse des Datenbestands sollen Handlungsempfehlungen gegeben werden.
Verbesserte Entscheidungsfindung	Datenanalyse und Handlungsempfehlungen sollen als Grundlage für die optimierte Entscheidungsfindung dienen
Automatisierung und Optimierung	Prozesse im Unternehmen werden automatisiert und optimiert durch die Analyse der Daten

In welchen Bereichen wird Data Science angewendet?

Data Science lässt sich in allen Unternehmensbereichen einsetzen, sodass fast keinerlei Grenzen gesetzt sind. Demnach spielt Datenwissenschaft überall dort eine entscheidende Rolle wo:

- Eine Vielzahl an Daten aufkommt
- Auf Basis dieser Datenmengen unterschiedliche Ziele für das Unternehmen wie beispielsweise das Treffen von Prognosen erreicht werden sollen

Dennoch gibt es typische Bereiche, in denen Datenwissenschaft vermehrt vorkommt. Dazu gehören vor allem der Online-Handel (E-Commerce), die Logistik, Gesundheits- und Finanzwesen, Industrie und Produktion.

Anwendungen nach Art der Tätigkeit

Tätigkeit	Beschreibung der Tätigkeit
Explorative Datenanalyse	Analyse sowie Auswertung vorliegender Daten zur Stützung von Hypothesen.
Vorhersage von Wahrscheinlichkeiten	Kauf- oder Kündigungswahrscheinlichkeiten lassen sich dank Data Science und zugehörigen Verfahren berechnen und vorhersagen
Vorhersage von numerischen Werten	Durch historische Berechnungen ist es möglich, zukünftige Vorhersagen, wie beispielsweise zukünftigen Stromverbrauch oder Umsätze zu treffen.
Erkennung von Anomalien	Data Science bietet die Möglichkeiten Unregelmäßigkeiten und Anomalien in den Daten zu identifizieren.
Analyse von Text und Sprache	Natural Language Processing (NLP) ist eine Möglichkeit, vorhandene Texte sowie gesprochene Sprache auszuwerten.
Analysieren von Bildern und Videos	Bilderkennung, Klassifikation, etc.
Erkennen von Zusammenhängen und Gruppen	Innerhalb riesiger Datenmengen (Big Data) ist es eine der Aufgabe von Data Science, Zusammenhänge sowie Gruppen innerhalb jeweiliger Datenmengen zu erkennen.

Welche Praxisbeispiele für Data Science gibt es?

Grosse Datenmengen werden heute mit den Instrumenten und Methoden der Data Science für Unternehmen aller Branchen ausgewertet.

Marketing: Vor allem zur Personalisierung im Marketing werden Methoden der Data Science eingesetzt, um grosse Datenmengen zu analysieren und das Treffen von Entscheidungen für beispielsweise Marketingstrategien zu verbessern. Dabei werden historische Daten zu Transaktionen, Verhalten und Demographie der Kunden analysiert, sodass ich Handlungsempfehlungen für das Marketing ableiten lassen.

IT-Security: Data Science wird zunehmend zur Überwachung von IT-Systemen eingesetzt. Kritische IT-Systeme werden dabei mithilfe von Security Information and Event Management (SIEM) geschützt.

Mobilität: Im Bereich der Mobilität wird autonomes Fahren dank Data Science und maschinellem Lernen weiter vorangetrieben. Dazu werden vorwiegend Sensordaten ausgewertet, um genaue Informationen des Fahrzeugs sowie der Umgebung zu generieren.

Retail- und Handelsunternehmen: Sie profitieren von Data Science durch Analysen des Kaufverhaltens von Kunden. Die Untersuchung möglicher Ursachen für Retouren hilft bei der Verringerung von Warenrücksendungen.

Grosse Datenmengen werden heute mit den Instrumenten und Methoden der Data Science für Unternehmen aller Branchen ausgewertet.

In der Gesundheitsbranche ermöglicht Data Science die Erstellung von Ähnlichkeitsanalysen als Grundlage für eine individualisierte Behandlung von Patienten und die Optimierung der Medikation.

Logistikunternehmen verbessern mithilfe von Data Science ihre Arbeitsprozesse und die Qualität ihrer Transport-Dienstleistungen.

Industriebetriebe steuern und optimieren Fertigungsabläufe durch den Einsatz von Data Science.

Versicherung und Banken schöpfen mithilfe von Data Science das Potenzial der ihnen zur Verfügung stehenden externen und internen Daten aus, um ihre Produktion zu verbessern und die Vertriebserfolge zu steigern.

Digitale vs. Produzierende Unternehmen

Digitale Unternehmen haben ein digitales Geschäftsumfeld und generieren automatisch Daten. Es wird nach Nutzen für die Daten gesucht.

Produzierende Unternehmen haben ein materielles Geschäftsmodell. Es besteht keine automatische Datengenerierung. Es gibt vielleicht Maschinen ohne eine IT-Anbindungen und die Digitalisierung der Industrie steht noch aus.

Produktionsoptimierung (Konkretes Beispiel)

Die automatisierte Stillstandserfassung soll dienen, um die Produktivität einschätzen und kritische Produktionsschritte identifizieren zu können.

Domänenwissen

Der heterogene Maschinenpark besteht aus 250 Maschinen verschiedener Hersteller und unterschiedlicher Grösse mit vielfältigen Aufgaben.

Die manuelle Stillstandserfassung stellt die Integrität der Daten in Frage. Der Detailgrad ist nicht ausreichend, da Stillstände nur pro Arbeitsbereich und nicht pro Maschine erfasst werden. Das Unternehmen ist in der Lebensmittelbranche tätig. Somit existieren starke Regularien für eingesetzte Materialien und Produkte.

Herausforderung Datenbeschaffung

Möglichkeiten der Datenbeschaffung

- Zugriff auf Maschinendaten durch OPC UA Schnittstellen
- Messung der Leistungsaufnahme der Maschine
- Vibrationsmessung durch Beschleunigungssensoren
- Vibrationsmessung durch Richtlaser
- Beobachtung von Status-LEDs mit Kamerasystemen

Anforderung des Unternehmens

- Extrem schnelle Umsetzung, um direkt einen Mehrwert zu generieren
- Kostengünstige Umsetzung, um alle Maschinen anbinden zu können
- Die Technologie soll zertifiziert sein oder keine Zertifizierung benötigen
- Maschinenunabhängige und skalierbare Lösung

In diesem Beispiel hat man sich für die Messung mit Vibrationsmessung durch Beschleunigungssensoren geeinigt.



Abbildung 2: Beschleunigungssensor, welcher mit Solar Modul betrieben wird

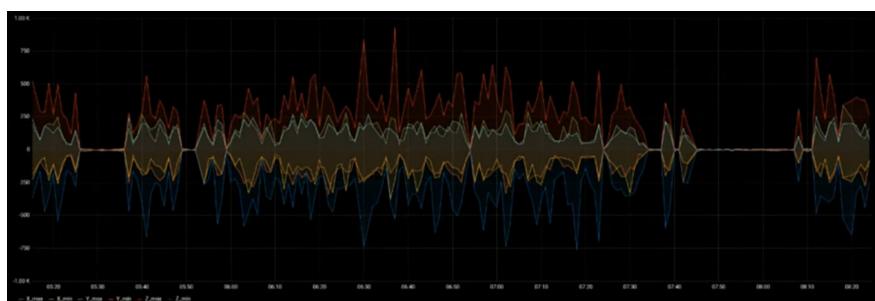


Abbildung 3: Datenauswertung

Beispiele aus der Schulmathematik

Löse folgende Gleichnung nach x auf:

$$2x + 1 = 5$$

➤ Es ist ein Problem, welches klar ist

- **Die Lösung ist klar (oft geübt)**
- **Nichts zu diskutieren oder zu interpretieren über Lösung $x = 2$**

Diese Art von Problemen gibt es in der angewandten Statistik oder Wissenschaft nicht.

Alltagsprobleme

- Sie haben es eilig, irgendwo hinzukommen, aber Sie können den Autoschlüssel nicht finden
- Ihnen geht das Mehl beim Backen eines Geburtstagskuchens aus und der Supermarkt ist geschlossen
- Ihr Flug wurde auf dem Weg zu einem Vorstellungsgespräch gestrichen
- Sie wollen diese tollen Schuhe kaufen, aber es ist kein Geld auf der Bank

Für all diese Probleme gibt es keine Lösungsanweisung. Die Lösung kann nicht mit Formeln gelöst werden. Es gibt ein Lösungsverfahren von Mazur. Die meisten statistischen Problemen sind in Worten formuliert.

Vier-Schritte-Problemlösungsstrategie nach Mazur

Erster Schritt

Es ist nicht klar, welches der effizienteste Weg ist, um auf ein bestimmtes Problem zu reagieren. Der erste Schritt unserer Lösungsstrategie:

- Die gegebenen Informationen organisieren und vergewissern sie sich, dass Ihnen klar ist, was genau in dem Problem erforderlich ist
- Stellen Sie sicher, dass Ihnen klar ist, welche Informationen im Problem enthalten sind
- Das Problem mit den eigenen Worten formulieren
- Feststellen, ob man alle Informationen hat oder nicht, die zur Lösung des Problems notwendig sind

Zweiter Schritt

Der nächste Schritt ist die Ausarbeitung eines Plans zur Lösung des Problems. Ein guter Plan ist es, die Schritte aufzuzeigen, die unternommt werden müssen, um das Problem zu lösen.

Dritter Schritt

Plan ausführen

Vierter Schritt

Das Resultat überdenken und interpretieren. Prüfen, ob das Ergebnis überhaupt möglich ist (z.B. negative Wahrscheinlichkeit ist falsch). Das Ergebnis wird dann in eigenen Worten interpretiert.

Die Daten müssen interpretiert werden. Selbst wenn der 4. Punkt der wichtigste ist, nützt es nichts, ihn richtig zu machen, wenn in den ersten drei Schritten ein Fehler gemacht wurde. Die Anwendung dieser vier Punkte erscheint bei sehr einfachen Aufgaben überflüssig.

Was ist Statistik, die nicht angewandt ist?

Angewandte Statistik: Verfahren und Methoden werden verwendet und beschrieben.

Nicht Angewandte Statistik:

$$S^2 := \frac{1}{m+n-2} \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2 \right\}.$$

Note that \bar{X} has expectation μ and variance σ^2/n , and \bar{Y} has expectation $\mu+\gamma$ and variance σ^2/m . So $\bar{Y} - \bar{X}$ has expectation γ and variance

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \left(\frac{n+m}{nm} \right).$$

The normality assumption implies that

$$\bar{Y} - \bar{X} \text{ is } \mathcal{N}\left(\gamma, \sigma^2 \left(\frac{n+m}{nm} \right)\right) \text{-distributed.}$$

Hence

$$\sqrt{\frac{nm}{n+m}} \left(\frac{\bar{Y} - \bar{X} - \gamma}{\sigma} \right) \text{ is } \mathcal{N}(0, 1) \text{-distributed.}$$

To arrive at a pivot, we now plug in the estimate S for the unknown σ :

$$Z(\mathbf{X}, \mathbf{Y}, \gamma) := \sqrt{\frac{nm}{n+m}} \left(\frac{\bar{Y} - \bar{X} - \gamma}{S} \right).$$

Deskriptive Statistik (eindimensional)

Daten

Daten und Statistiken bestimmen immer mehr unser Leben. Z.B. Zeitungen machen Prognosen durch Befragungen, Google wertet unsere Suchanfragen aus, an Flughäfen werden Gesichter analysiert und vieles mehr.

Datensatz (eindimensional)

Ein eindimensionaler Datensatz ist z.B. eine Liste oder ein Array. Solche Listen heißen eindimensionale Datensätze oder Messreihen.

Bsp: Körpergrösse von 5 Personen: 1.75, 1.80, 1.72, 1.65, 1.54

Datensätze (zweidimensional)

Die häufigste Form von Datensätzen sind Tabellen oder zweidimensionale Datensätze.

Bsp:

Person	Grösse	Gewicht	Geschlecht	Nationalität
A	1.82	72	m	CH
B	1.75	82	w	D
C	1.61	70	w	CH
D	1.80	83	m	A
E	1.89	95	w	FL

Die Grösse und das Gewicht sind Quantitative Daten (also gemessene Zahlen). Sie können theoretisch jeden beliebigen Zahlenwert in einem Bereich annehmen.

Das Geschlecht und die Nationalität sind Qualitative Daten. Es ist nur eine bestimmte Anzahl Werte möglich (z.B. alle Nationalitäten der Welt).

Deskriptive Statistik

Die Deskriptive Statistik befasst sich mit der Darstellung von Datensätzen. Die Datensätze werden durch gewisse Zahlen charakterisiert (z.B. Mittelwert) und graphisch dargestellt. Zunächst wird eine Messgröße an einem Untersuchungsobjekt ermittelt.

Ziele der deskriptiven Statistik

Die Ziele sind das Zusammenfassen durch numerische Kennwerte und die graphische Darstellung der Daten.

Messungen finden nie unter exakt denselben Bedingungen statt. Messungen werden aber mit grösstmöglicher Sorgfalt durchgeführt. Trotzdem können die Messwerte variieren. Es stellt sich nun die Frage gibt es Unterschiede zwischen den Messungen? Falls ja, wie können diese Unterschiede ermittelt werden?

Ziel ist es, verschiedene Messungen / Messreihen zusammenzufassen, um die beiden Messreihen miteinander vergleichen zu können. Die Deskriptive Statistik beschäftigt sich damit, auf welche Weisen die Daten organisiert und zusammengefasst werden können. Ziel ist auch die Interpretation und darauffolgende statistische Analyse dieser Daten zu vereinfachen.

Die Kennzahlen sollen die Daten numerisch zusammenfassen und grob charakterisieren. Bei einer statischen Analyse ist es sehr wichtig, nicht einfach blind ein Modell anzupassen und ein statisches Verfahren anzuwenden. Die Daten sollen immer mit Hilfe von geeigneten graphischen Mitteln und den Kennzahlen dargestellt werden. Nur auf diese Weise kann man Strukturen und Besonderheiten entdecken.

Wann immer ein Datensatz reduziert wird, geht Information verloren! Bspw. Werden aus zehn Schulnoten der Durchschnitt berechnet, gehen die Daten über die einzelnen Noten verloren

Bezeichnungen

Standardbezeichnung von Daten

$$x_1, x_2, x_3, \dots, x_n$$

n stellt dabei der Umfang der Messreihe dar

Kennzahlen

Arithmetisches Mittel

Umgangssprachlich wird das arithmetische Mittel auch als den Durchschnitt bezeichnet. Die Definition ist wie folgt:

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

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02)
mean(waageA)
## [1] 80.02077
```

Das Arithmetische Mittel sagt aus, wo sich die Mitte der Daten befinden.

Streuung graphisch

Das Arithmetische Mittel sagt schon einiges über einen Datensatz aus. Aber das Arithmetische Mittel sagt nicht alles über die Messreihe aus. Es kann also sein, dass mehrere Reihen dasselbe Arithmetische Mittel haben, aber unterschiedliche Daten besitzen. So entstehen verschiedene Streuungen um das Arithmetische Mittel.

Empirische Varianz und Standardabweichung

Man berechnet die Empirische Varianz und die empirische Standardabweichung, also ein Mass für die Variabilität oder Streuung der Messwerte.

$$\text{Var}(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2$$
$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Empirische Varianz

Bei der Varianz wird $x_i - \bar{x}$ quadriert, damit sich die Abweichungen nicht aufheben. In den Nenner kommt dann $n - 1$, dies ist mathematisch begründet. Ist die empirische Varianz gross, so ist die Streuung der Messwerte über das arithmetische Mittel gross. Der Wert hat keine physikalische Bedeutung. Man weiss nur, je grösser der Wert, umso grösser die Streuung. Wichtig: nur die Standardabweichung lässt sich interpretieren.

```
var(waageA)
## [1] 0.000574359
```

Standardabweichung

Für die Standardabweichung wird danach die Wurzel der Varianz gezogen. Durch das Wurzelziehen wird die Varianz wieder zur selben Einheit wie die Daten selbst.

```
sd(waageA)
## [1] 0.02396579
```

Median

Der Median ist ein weiteres Lagemass. Es handelt sich um den Wert, bei dem die Hälfte der Messwerte unter oder gleich diesem Wert sind. Die andere Hälfte ist gleich diesem Messwert oder darüber.

Damit der Median gefunden werden kann, braucht es eine geordnete Stichprobe. Die Runden Klammern geben an, dass der Index geordnet sein muss.

$$x_{(1)} < x_{(2)} < \dots < x_{(n)}$$

```
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)
median(waageB)
## [1] 79.97
```

Median vs. Arithmetisches Mittel

Es handelt sich bei beiden um zwei Laagemasse der Mitte eines Datensatzes. Doch, welches der beiden Masse besser ist, kann man nicht sagen, denn es kommt auf die jeweilige Problemstellung darauf an. Am besten betrachtet man beide Masse gleichzeitig.

Eine spezielle Eigenschaft des Medians ist die Robustheit, sprich es bleibt stabil bei grossen Ausreissern. Das Arithmetische Mittel verändert sich mit den Ausreissern.

Quartile

Der Median sagt aus, wo der Wert ist, wo die Hälfte der Beobachtungen kleiner oder gleich wie dieser Wert sind. Es gibt die analoge Überlegung mit den Quartilen. Es gibt ein oberes und ein unteres Quartil.

Das untere Quartil ist der Wert, wo 25% aller Beobachtungen kleiner oder gleich und 75% grösser oder gleich dieser Wert sind.

Das obere Quartil ist der Wert, wo 75% aller Beobachtungen kleiner oder gleich und 25% grösser oder gleich dieser Wert sind.

Kann nicht genau ein Wert zugeordnet werden, z.B. bei einer ungeraden Anzahl an Beobachtungen, so wird der Mittelwert zwischen den Beobachtungen erhoben und als Quartil verwendet.

```
# Syntax für das untere Quartil: p=0.25
quantile(waageA, p = 0.25, type = 2) ## 25%
## 80.02
quantile(waageB, p = 0.25, type = 2)
## 25%
## 79.96
# Syntax für das obere Quartil: p=0.75
quantile(waageA, p = 0.75, type = 2) ## 75%
## 80.04
```

Quartilsdifferenz

Die Quartilsdifferenz ist ein Mass für die Streuung der Daten.

$$\text{Streuungsdifferenz} = \text{Oberes Quartil} - \text{Unteres Quartil}$$

So misst die Quartilsdifferenz die Länge des Intervalls, das etwa die Hälfte der mittleren Beobachtungen enthalten sollte. Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte um den Median und umso kleiner ist die Streuung. Dieses Streuungsmass ist robust.

```
IQR(waageA, type=2)
## [1] 0.02
```

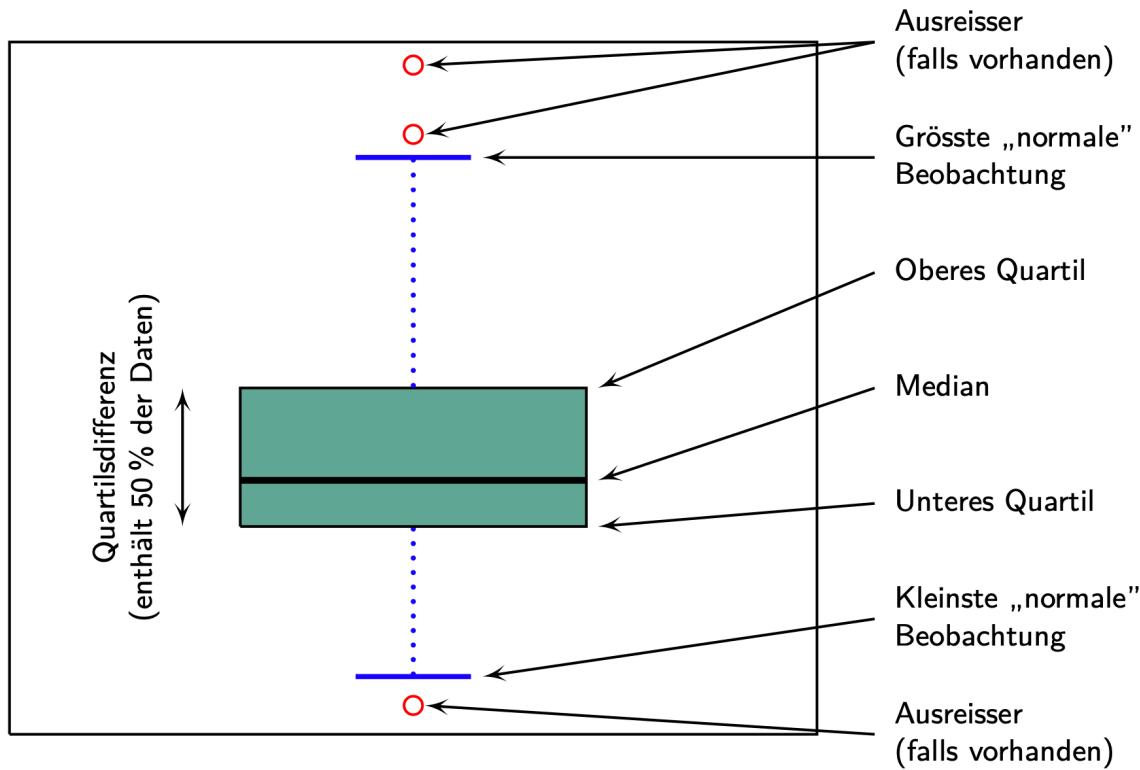
Quantile

Die Quantile können auf jede Prozentzahl angewendet werden. Es gibt also z.B. 10% Quantil, 25% Quantil (unteres Quartil), 50% Quantil (Median), 75% Quantil (oberes Quartil).

```
quantile(waageA, p = .1, type = 2)
## 10%
## 79.98
quantile(waageA, p = .7, type = 2) ## 70%
## 80.04
```

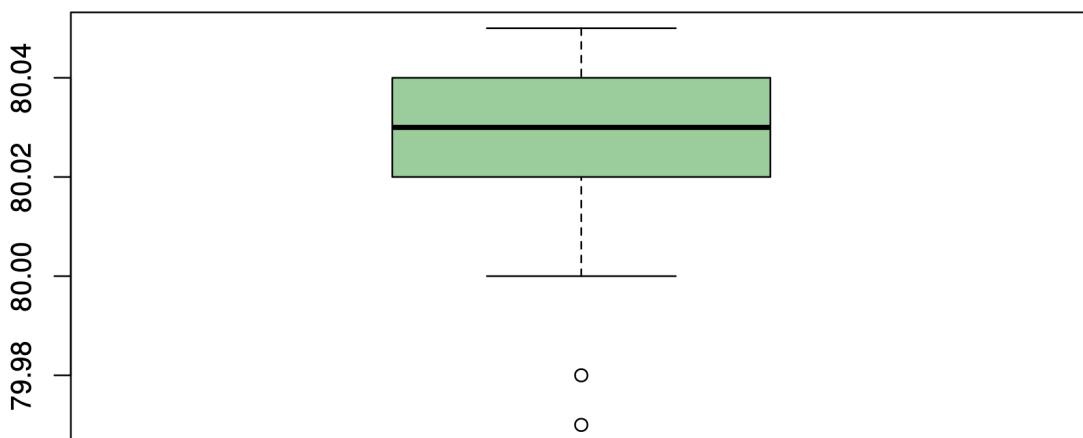
Die Berechnung von oben sagt aus, dass 10% der Messwerte kleiner oder gleich 79.98 sind und entsprechend 70% der Messwerte kleiner oder gleich 80.04 gross sind.

Boxplot



Das Rechteck, dessen Höhe vom empirischen 25% und vom 75% Quantil begrenzt wird, bildet eine Box. Der Horizontale Strich zeichnet den Median ab. Die blauen Linien zeigen den kleinsten, bzw. grössten «normalen» Wert. Der «normale» Wert ist höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt. Ausreisser, also jene Punkte, welche kleiner oder grösser der «normalen» Werte sind, sind rot eingezeichnet.

```
boxplot(waageA,
        col = "darkseagreen3"
    )
```



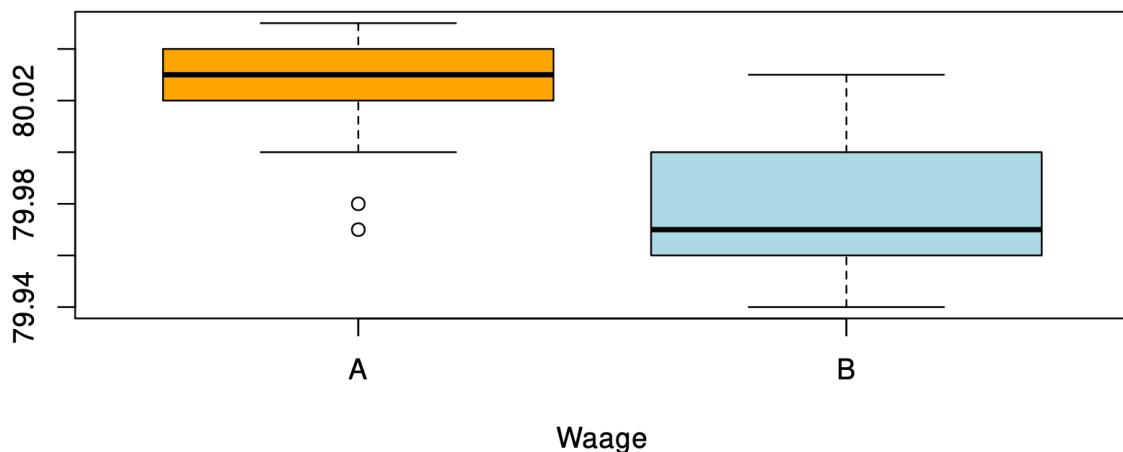
Die Hälfte der Beobachtungen befindet sich zwischen dem oberen Quartil 80.04 und dem unteren Quartil 80.02, die Quartilsdifferenz ist also 0.02. Der Median liegt bei 80.03. Somit liegt der «normalen» Bereich der Werte zwischen 80.00 und 80.05. Es gibt zwei Ausreisser bei 79.97 und 79.98.

Allgemein gesagt ist ein Boxplot also eine Darstellung von Median und Quartilen.

Vergleich von Datensätzen

Es können mittels mehreren Boxplots verschiedene Datensätze verglichen werden.

```
boxplot(waageA, waageB,
       xlab = "Waage",
       col = c("orange", "lightblue")
     )
axis(side = 1, at = c(1, 2), labels = c("A", "B"))
```



Folgendes kann aus dem Diagramm gelesen werden:

- Die Waage A hat grössere Werte als Waage B → Der Median von A ist grösser
- Daten von der Waage A haben weniger Streuung als die Daten von Waage B
→ Das Rechteck ist weniger hoch (Quartilsdifferenz)

Zweidimensionale Deskriptive Statistik

Histogramm

Das Histogramm ist ein graphischer Überblick über die auftretenden Werte. Die Aufteilung des Wertebereichs erfolgt in k Klassen (Intervall). Es gilt folgende Faustregel:

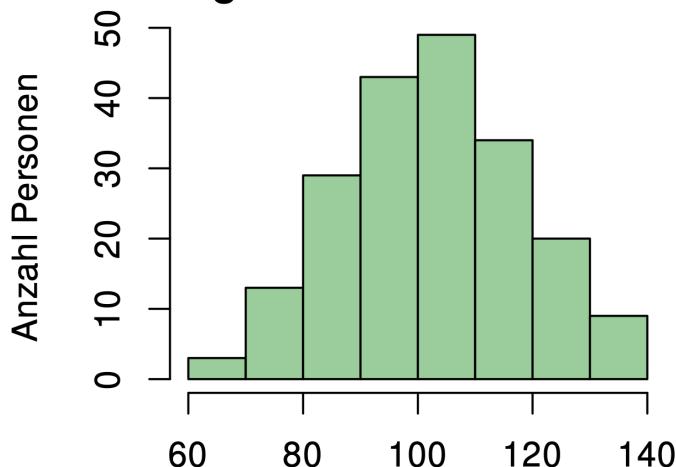
- Bei weniger als 50 Messungen ist die Klassenzahl 5 bis 7
- Bei mehr als 250 Messungen wählt man 10 bis 20 Klassen

Für jede Klasse wird einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist gezeichnet.

Beispiel IQ-Test

Die Abbildung: Ein Histogramm vom Ergebnis eines IQ-Tests von 200 Personen

Verteilung der Punkte in einem IQ-Test



Punkte im IQ-Test

Die Daten wurden simuliert.

Breite der Klassen: 10 IQ-Punkte; für jede Klasse gleich

Höhe der Balken: Anzahl Personen, die in diese Klasse fallen

Beispiel: ca. 20 Personen fallen in die Klasse zwischen 120 und 130 Punkten

Die Form dieses Histogramms ist typisch für viele Histogramme. Es besitzt die Normalverteilung.

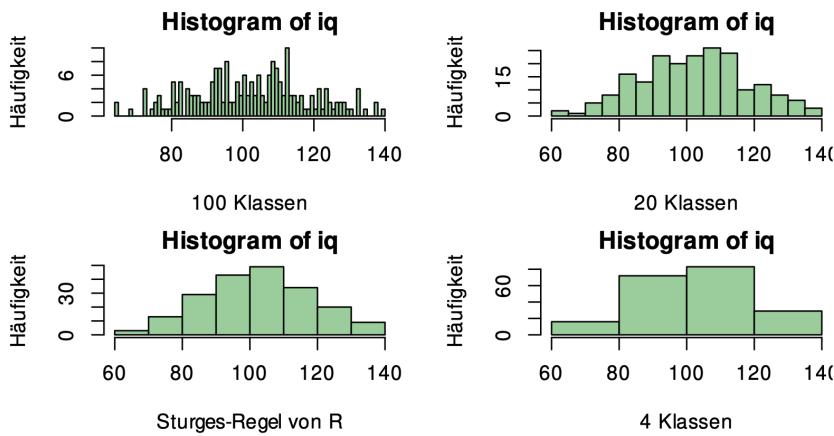
```
iq <- rnorm(n = 200, mean = 100, sd = 15)
hist(iq,
      col = "darkseagreen3",
      xlab = "Punkte im IQ-Test",
      ylab = "Anzahl Personen",
      main = "Verteilung der Punkte in einem IQ-Test"
)
```

Der Befehl `rnorm(n = 200, mean = 100, sd = 15)` wählt zufällig 200 normalverteilte Dateen mit Mittelwert 100 mit Standardabweichung 15 aus.

Der Befehl `hist(...)` zeichnet das Histogramm. Die weiteren Optionen sollten klar sein: `xlab` (x-Label / Beschriftung x-Achse), `ylab` (y-Label / Beschreibung y-Achse), `col` (Farbe), `main` (Haupttitel).

Wahl der Klassen

Die Wahl der Anzahl Klassen ist relevant für die Aussagekraft eines Histogramms. Es gibt keine allgemeingültige Grundregel, wie man die Anzahl Klassen wählt.



In der Abbildung sind die IQ Daten vom Beispiel mit verschiedenen Anzahl Klassen zu sehen.

Oben Links: Das Histogramm ist viel zu detailliert, als dass man ein Muster erkennen könnte

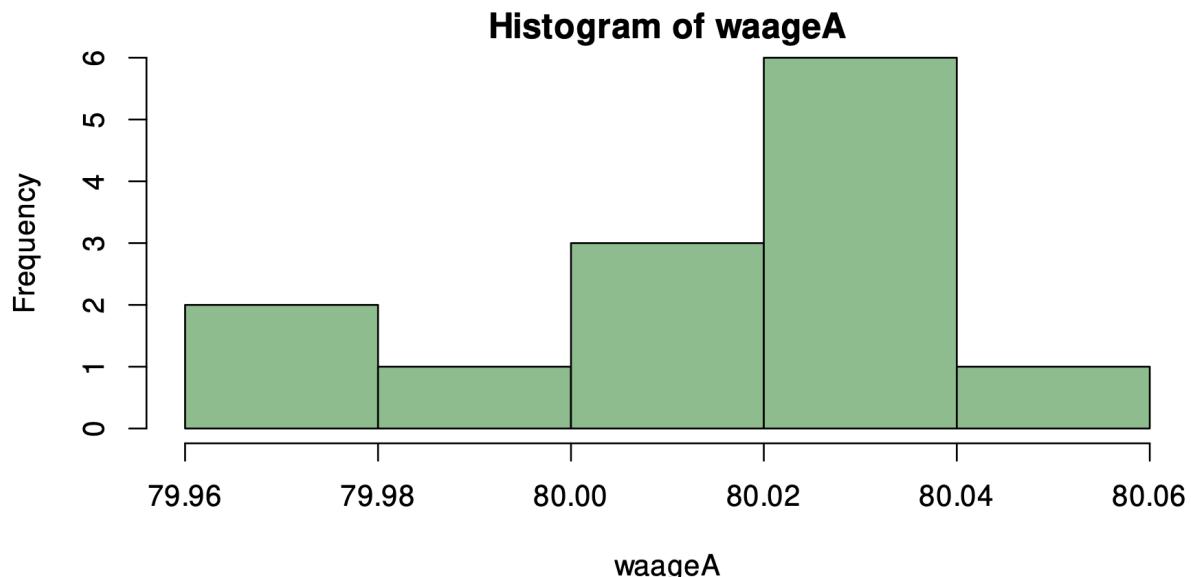
Histogramm rechts unten: Dieses Histogramm ist zu ungenau

```
par(mfrow = c(2, 2))
hist(iq,
      breaks = 100,
      xlab = "100 Klassen",
      ylab = "Häufigkeit",
      col = "darkseagreen3")
)
hist(iq,
      breaks = 20,
      xlab = "20 Klassen",
      ylab = "Häufigkeit",
      col = "darkseagreen3")
)
hist(iq,
      breaks = "sturges", # default R
      xlab = "Sturges-Regel von R",
      ylab = "Häufigkeit",
      col = "darkseagreen3")
)
hist(iq,
      breaks = 3,
      xlab = "4 Klassen",
      ylab = "Häufigkeit",
      col = "darkseagreen3")
```

Der Befehl `par(mfrow = c(2,2))` lässt vier Histogramme in zwei Zeilen und 2 Spalten zeichnen. Die Option «`breaks`» legt die Anzahl Klassen fest. `Breaks` ist jedoch nur ein Vorschlag für R.

Beispiel Waage

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02)  
hist(waageA, col="darkseagreen")
```

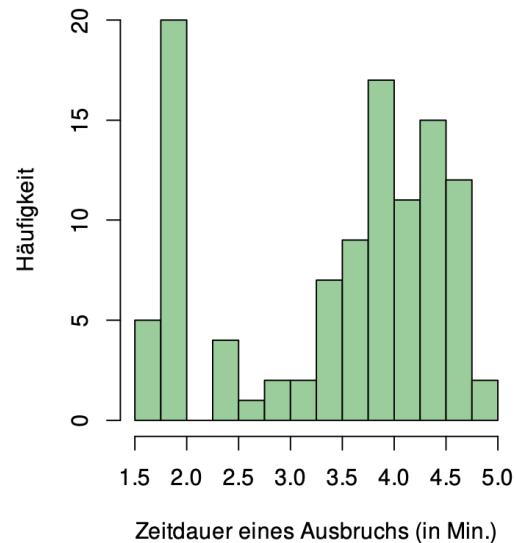
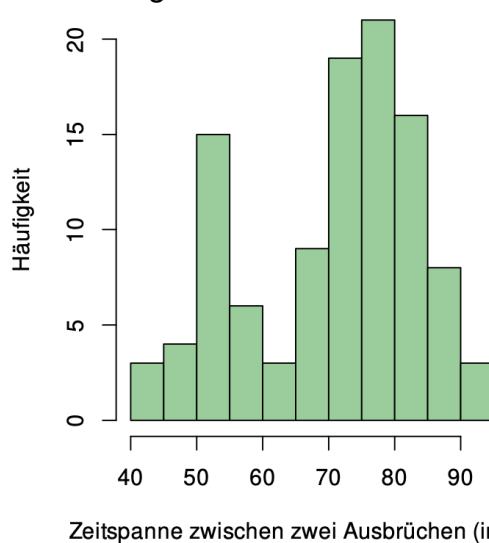


Die Waage hat 13 Messungen → 5 Balken

In der ersten Klasse werden Werte von 79.96 bis 79.98 berücksichtigt. In der zweiten Klasse Werte von 79.99 bis 80.00, usw. Die Linke Grenze wird also nicht berücksichtigt, die rechte schon. Wenn man wünscht, kann man das Ganze auch umkehren, so dass die linken Werte berücksichtigt werden, die rechten aber nicht. Das Histogramm würde minimal anders aussehen, bei grossen Datensätzen spielt dies aber kaum eine Rolle.

Bimodales Verhalten

In den beiden Histogrammen ist ein Bimodales Verhalten sichtbar. Es gibt zwei «Hügel» im Histogramm.



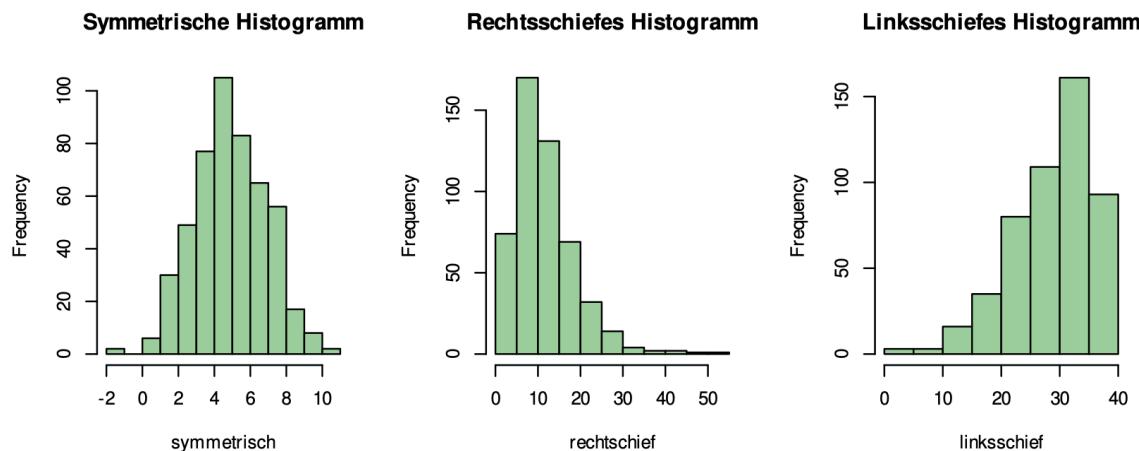
Schiefe von Histogrammen

Das Histogramm links ist symmetrisch bezüglich ungefähr 5. Die Daten sind um 5 auf beiden Seiten ähnlich verteilt.

Beim mittleren Histogramm sind die meisten Daten links im Histogramm, es ist also rechtsschief.

Beim rechten Histogramm sind die meisten Daten rechts im Histogramm, es ist also ein linksschiefer Histogramm.

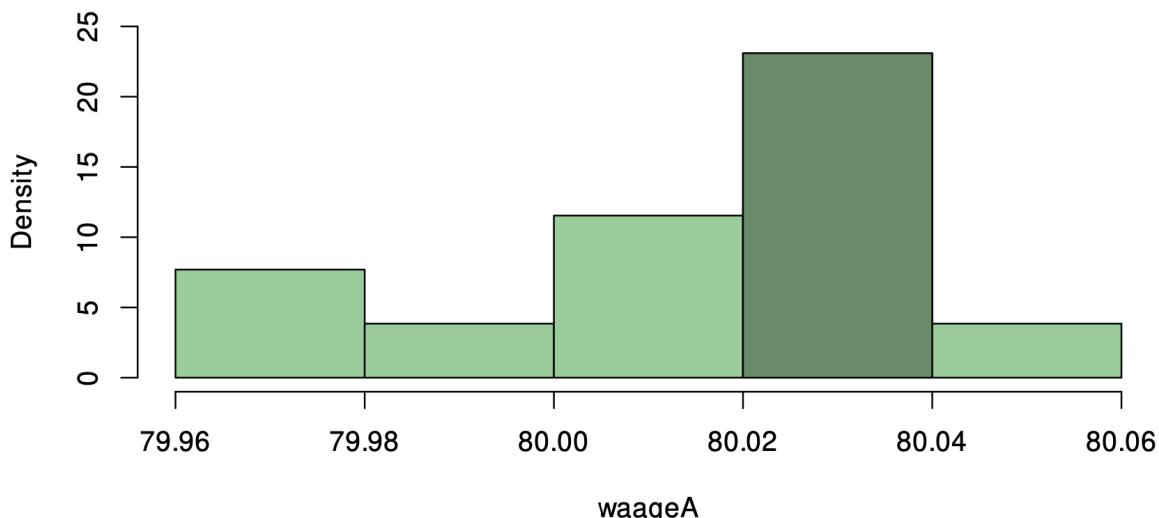
Die Bezeichnung links und rechts bezieht sich also auf die Richtung, wo es weniger Daten hat.



Normiertes Histogramm

Bisher entsprach die Höhe der Balken der Anzahl der Beobachtungen einer Klasse. Oft besser und übersichtlicher wird es, wenn die Balkenhöhe so gewählt wird, dass die Balkenfläche dem prozentualen Anteil der jeweiligen Beobachtungen an der Gesamtzahl Beobachtungen entspricht. Die Gesamtfläche aller Balken muss dann gleich eins sein. Auf der vertikalen Achse wird die Dichte angegeben.

Histogramm von Waage A



Die Dichte der Klasse von 80.02 bis 80.04 ist etwa 23. Die Fläche dieses Balkens wird wie folgt berechnet: $(80.04 - 80.02) * 23 = 0.46$. Die Fläche mit 100 multipliziert ergibt die Prozentzahl der Daten, die in diesem Balken liegen, was also bei etwa 46% der Daten liegt. Also sind 46% der Daten zwischen 80.02 und 80.04 gross.

```

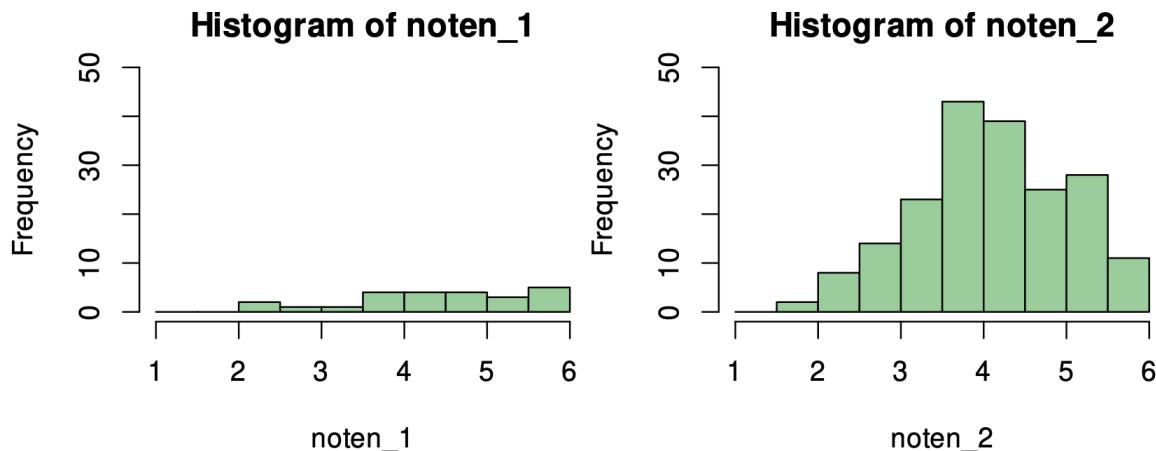
hist(waageA,
freq = F,
main = "Histogramm von Waage A",
col = "darkseagreen3",
ylim = c(0, 25)
)
rect(80.02, 0, 80.04, 23.1, col="darkseagreen4")

```

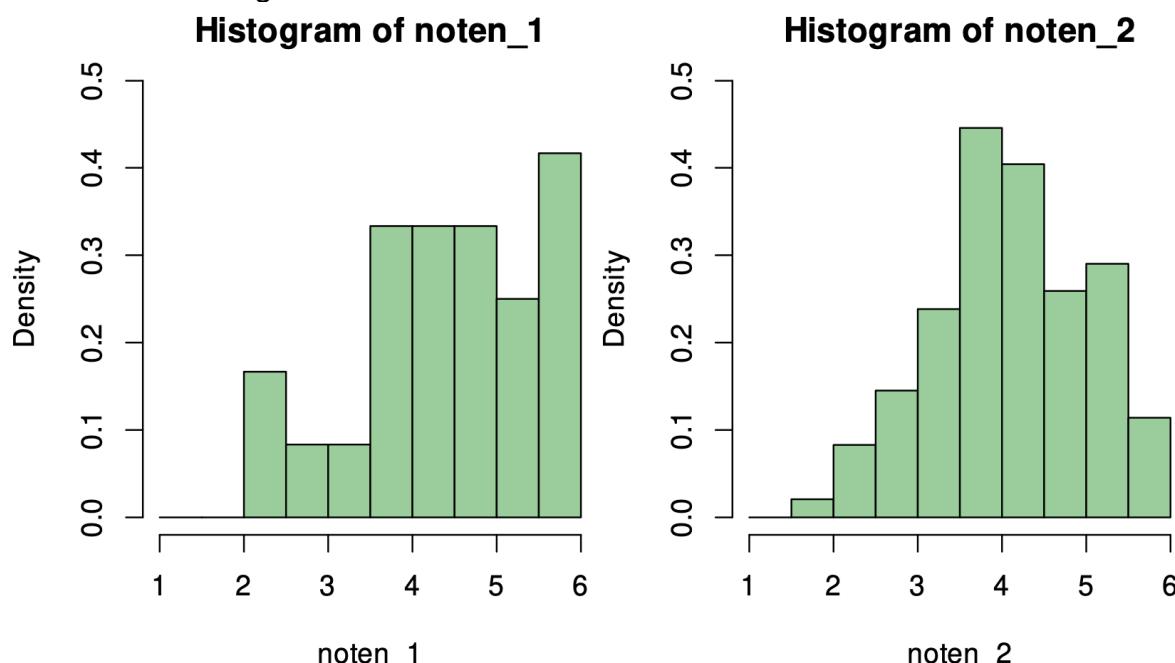
Die Option freq = F gibt an, dass das Histogramm normiert gezeichnet wird. Die Option ylim = C(0,25) limitiert die Y-Achsenhöhe auf 25. rect(80.02, 0, 80.04, 23.1, col = "darkseagreen4") färbt die Fläche zwischen den Punkten (80.02, 0) und (80.04, 23.1) dunkelgrün ein.

Vorteil von normierten Diagrammen

Werden Diagramme normiert, können verschiedene Datensätze miteinander verglichen werden. So kann beispielsweise die Benotung einer Klasse von 24 Lernenden mit einer Benotung einer Klasse mit 194 Lernenden verglichen werden. Ein nicht-Normiertes Diagramm sagt nicht viel aus:



Ein normiertes Diagramm schon viel mehr:



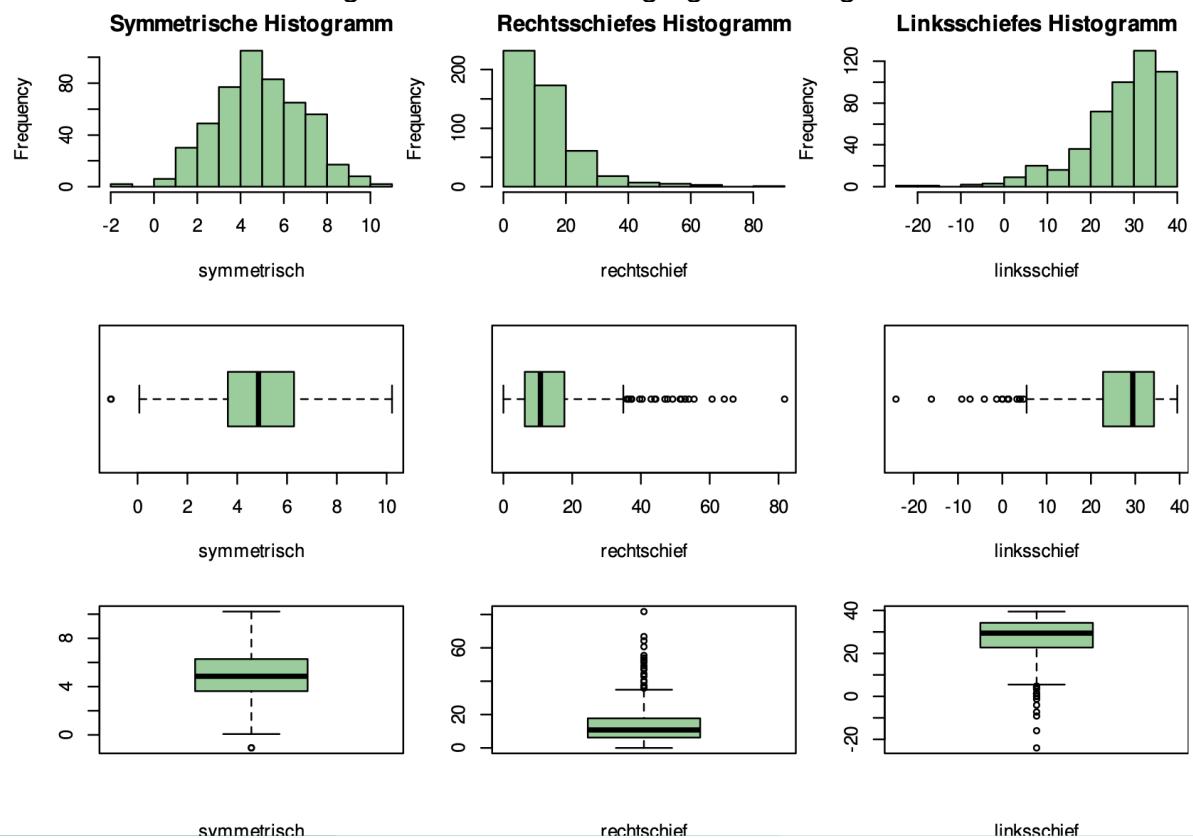
Die Klasse 1 hat im oberen Bereich mehr Anteile als die zweite Klasse. Vor allem der Balken der 5.5-6 der ersten Klasse ist sehr viel höher als der von der zweiten Klasse. Die Klasse 1 hat prozentual als stärkere Lernende als die Klasse 2. Die Klasse 1 hat aber eher mehr schwächere Lernende als Klasse 2. Im Mittleren Bereich der Noten hat Klasse 2 prozentual mehr Lernende als Klasse 1.

Schiefe im Boxplot

Symmetrisches Diagramm links: Median in der Mitte der Box

Rechtsschiefes Histogramm (Mitte): Median nicht mehr in der Mitte der Box, sondern nach links verschoben. Vom unteren Quartil zum Median liegen viele Daten in kleinem Bereich. Vom Median zum oberen Quartil braucht es ein viel größerer Bereich (bis 25% der Daten), die in diesem Intervall liegen.

Beim linksschiefen Histogramm ist die Sachlage gerade umgekehrt.



Boxplot Bemerkungen

Im Boxplot ist ersichtlich:

- Lage
- Streuung
- Schiefe

Man sieht aber z.B. nicht, ob eine Verteilung mehrere Peaks hat.

Deskriptive Statistik zweidimensionaler Daten

Zweidimensionale Daten sind, wenn einem Versuchsobjekt jeweils zwei verschiedene Größen zugewiesen sind. Beispiel: An einer Gruppe von Menschen werden jeweils die Körpergrösse und das Körpergewicht gemessen.

Beispiel Mortalität und Weinkonsum in Ländern

Untersucht durchschnittlicher Weinkonsum (in Liter pro Person und Jahr) und die Sterblichkeit (Mortalität; Anzahl Todesfälle pro 1000 Personen zwischen 55 und 64 Jahren pro Jahr) aufgrund von Herz- und Kreislauferkrankungen in 18 Ländern.

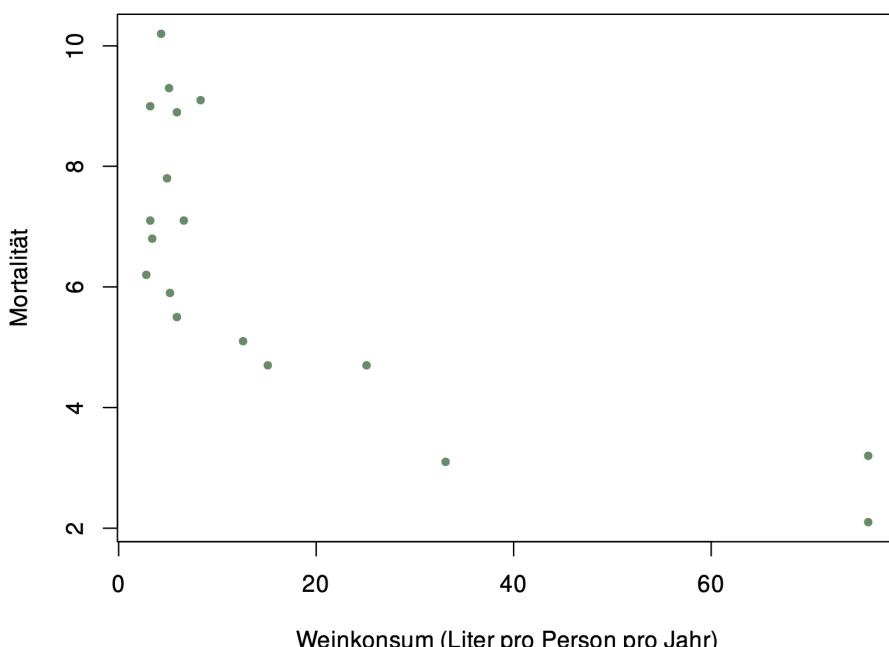
Land	Weinkonsum	Mortalität Herzerkrankung
Norwegen	2.8	6.2
Schottland	3.2	9.0
Grossbritannien	3.2	7.1
Irland	3.4	6.8
Finnland	4.3	10.2
Kanada	4.9	7.8
Vereinigte Staaten	5.1	9.3
Niederlande	5.2	5.9
New Zealand	5.9	8.9
Dänemark	5.9	5.5
Schweden	6.6	7.1
Australien	8.3	9.1
Belgien	12.6	5.1
Deutschland	15.1	4.7
Österreich	25.1	4.7
Schweiz	33.1	3.1
Italien	75.9	3.2
Frankreich	75.9	2.1

Ein wichtiger Schritt in der Untersuchung zweidimensionaler Daten ist die Graphische Darstellung. Meist erfolgt dies über ein sogenanntes Streudiagramm. Zwei Messungen als Koordinaten von Punkten in einem Koordinatensystem interpretiert und dargestellt.

X-Achse: Weinkonsum

Y-Achse: Mortalität

Aus den einzelnen Datenpunkten ergeben sich die Koordinaten für die Punkte (x_1, y_1) (x_2, y_2)

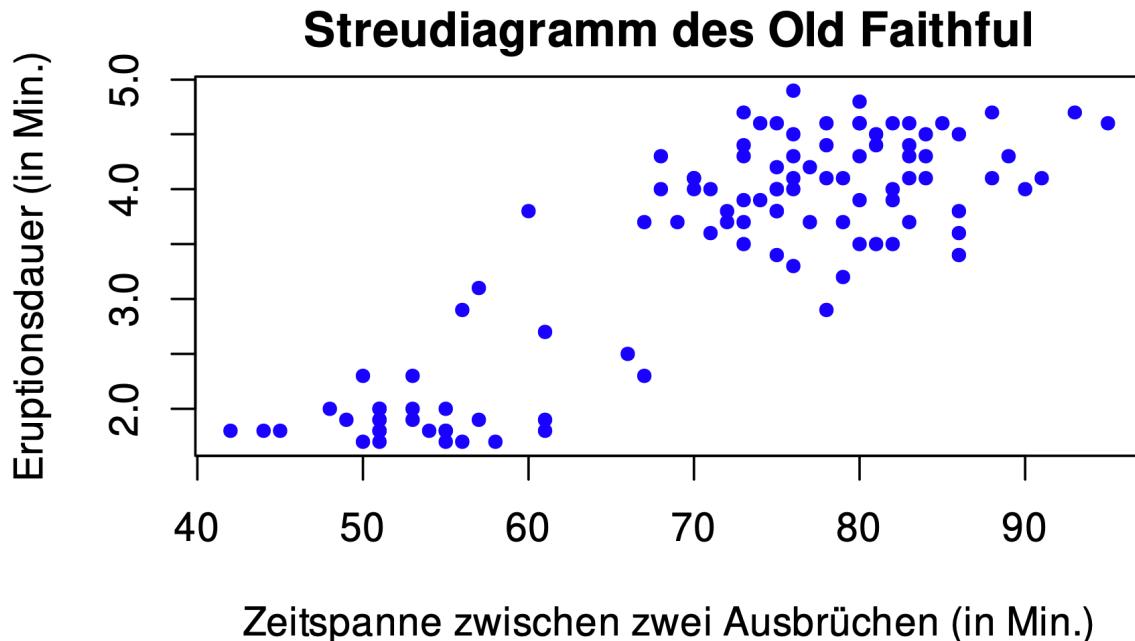


```

wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9, 6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9)
mort <- c(6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5, 7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1)
plot(wein, mort,
     xlab = "Weinkonsum (Liter pro Jahr)", ylab = "Mortalität",
     col = "blue",
     pch = 20
)

```

Beispiel Old Faithful



Erkenntnisse

Zunächst ist die Punktewolke steigend → Je länger die Zeitspanne zwischen den Ausbrüchen, umso länger dauert der Ausbruch

Im Streudiagramm gibt es aber zwei Gruppen → Eine Links unten und eine rechts oben. Ist die Zeitspanne zwischen zwei Ausbrüchen kurz, so ist die nächste Eruptionsdauer kurz. Ist die Zeitspanne lang, so ist die Eruptionsdauer lang. Eine mittlere Zeitspanne mit einer mittleren Ausbruchsdauer gibt es nicht.

Abhängigkeit und Kausalität

Bei Streudiagrammen muss aufgepasst werden, dass die Abhängigkeit nicht mit der Kausalität verwechselt wird. Es kann sein, dass auch Gesetzmäßigkeiten vorhanden sind, es heißt aber noch lange nicht, dass diese Gesetzmäßigkeiten auch kausal erklärt werden können.

Lineare Regression

Ein Kunde kauft in einer Buchhandlung 10 Bücher zu folgenden Preisen

	Seitenzahl	Buchpreis (SFr)
Buch 1	50	6.4
Buch 2	100	9.5
Buch 3	150	15.6
Buch 4	200	15.1
Buch 5	250	17.8
Buch 6	300	23.4
Buch 7	350	23.4
Buch 8	400	22.5
Buch 9	450	26.1
Buch 10	500	29.1

Beobachtung: Je dicker ein Roman ist, desto teurer ist er in der Regel.

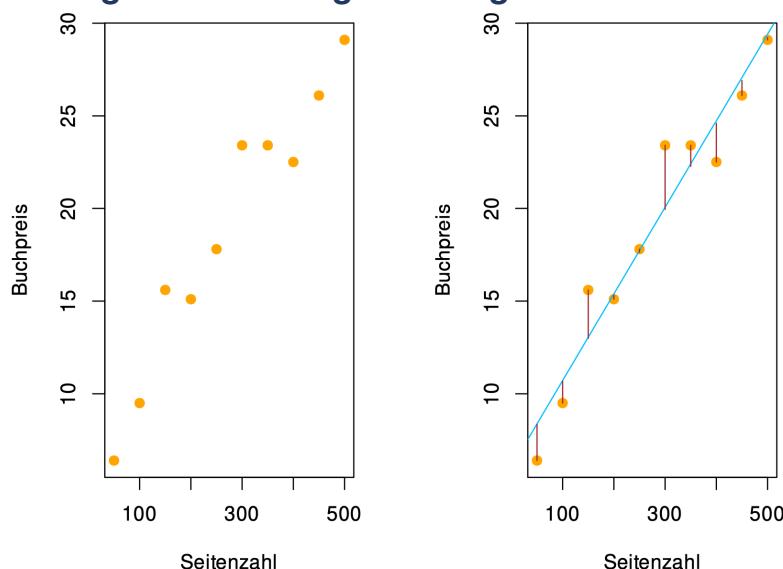
Fragen:

- Wie viel kostet eine Seite?
- Wie teuer ein Buch mit null Seiten wäre? → Grundkosten für ein Buch
- Was würde dann voraussichtlich ein Buch mit 375 Seiten kosten? Diese Seitenzahl kommt in der Tabelle nicht vor

Ziel: Formelmässiger Zusammenhang zwischen Buchpreis und Seitenzahl

Vorhersagen über Buchpreis möglich für Bücher mit Seitenzahl, die in Liste nicht auftauchen

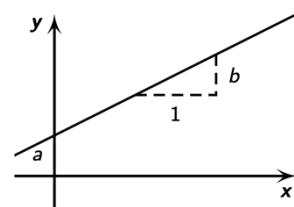
Streudiagramm und Regressionsgrade



Gerade:

a: y-Achsenabschnitt
b: Steigung

$$y = a + bx$$



Nimmt x um eine Einheit zu, so ändert sich y um b.

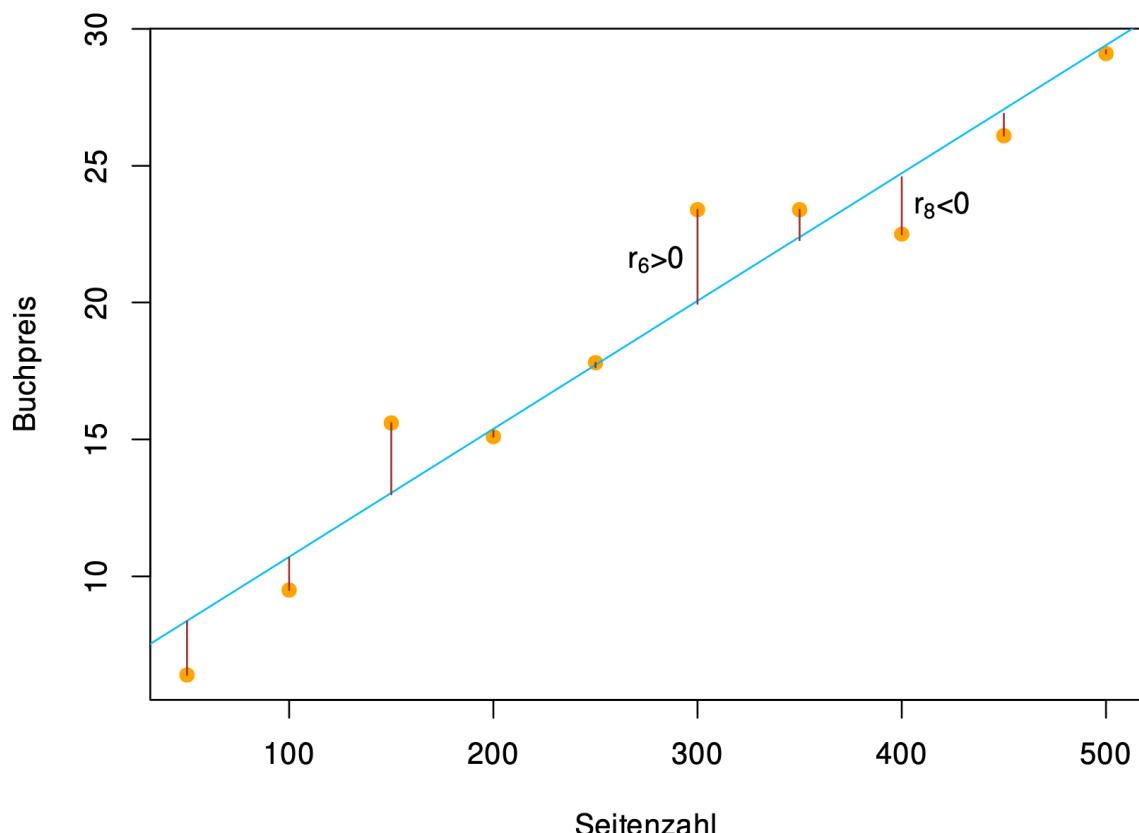
Vermutung: Eine Gerade scheint recht gut zu den Daten zu passen.

Diese Gerade hätte die Form: $y = a + bx$ mit y (Buchpreis), a (Grundkosten), b (Kosten pro Seite). Das Problem liegt gerade dabei, die Gerade zu finden, die möglichst durch alle Punkte geht. Eine Möglichkeit besteht, Vertikale Abstände zwischen den Beobachtungen und Gerade zusammenzählen. Dabei sollte eine kleine Summe der Abstände eine gute Anpassung bedeuten. Der Abstand zwischen dem Messpunkt und der Geraden wird Residuum genannt.

Residuum

Ein Residuum r ist die vertikale Differenz zwischen einem Datenpunkt (x, y) und dem Punkt $(x, a + bx)$ auf der gesuchten Geraden:

$$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$



Bei Residuen r_6 und r_8 für die Geraden der Abbildung. Das Residuum r_6 ist positiv, da der Punkt oberhalb der Geraden ist. Entsprechend ist r_8 kleiner als 0. Die Gerade $y = a + bx$ muss so bestimmt werden, dass die Summe von $r_1 + r_2 + \dots + r_n = \sum_i^n r_i$ möglichst klein ist. Das Ganze hat aber eine gravierende Schwäche. Falls die Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegt, so heben sich die Zahlen auf! Das kann dazu führen, dass die Geraden gar nicht zur Datenpunkte passen.

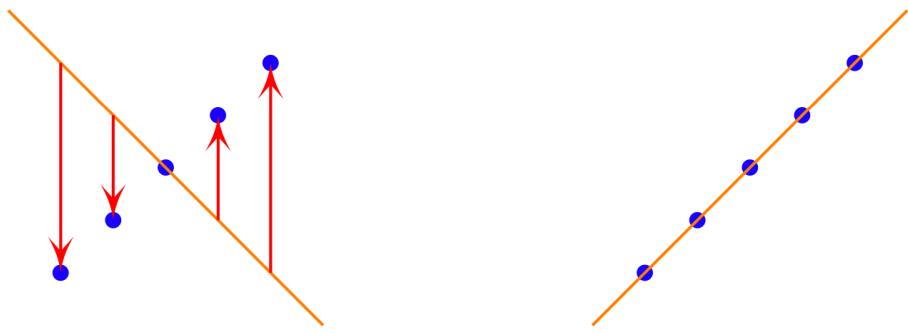


Abbildung links: Summe der Residuen ist 0, aber Gerade passt aber überhaupt nicht
 Abbildung rechts: Summe der Residuen ist 0, Gerade passt perfekt

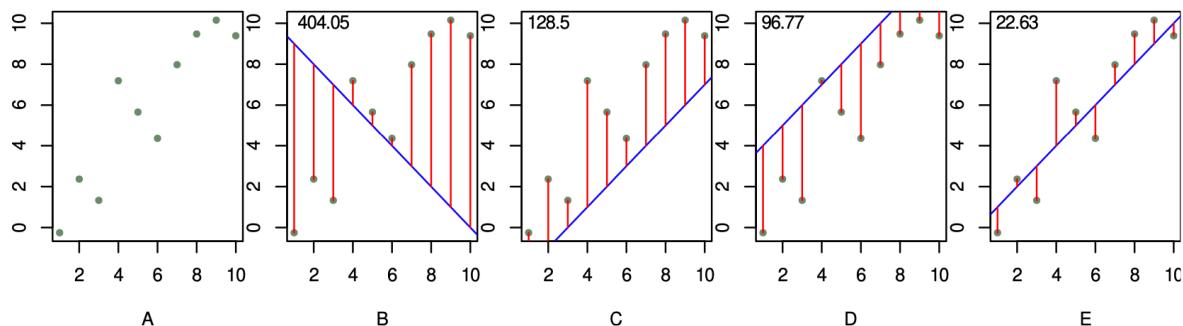
Aber welches ist die „richtige“ Gerade, die am besten zu der Punktewolke passt?
 Verfahren gesucht, das diese Gerade eindeutig festlegt

Methode der kleinsten Quadrate

Eine andere Möglichkeit besteht darin, die Quadrate der Abweichungen aufzusummen, also

$$r_1^2 + r_2^2 + \dots + r_n^2 = \sum_i r_i^2$$

Der Parameter a und b muss also so gewählt werden, dass die Summe minimal wird.



Gesucht ist die gemäss der kleinsten Methode am besten zur Punktewolke passenden Gerade. Die Punktewolke steigt, deshalb eine steigende Gerade. Bei Abbildung B ist die Gerade fallend. Deshalb ist die Gerade nicht gut geeignet und die Residuen werden sehr lange.

Bei Abbildung C steigt die Gerade zwar auf, aber noch zu tief. Die Residuen sind immer noch sehr lang.

Bei Abbildung D ist die Gerade steigend, aber ebenfalls die Residuen noch immer sehr lang, jedoch besser als Abbildung C.

Bei Abbildung E finden wir die Lösung, denn die Gerade passt sehr gut zur Punktewolke und die Residuen sind verglichen mit den anderen Abbildungen klein.

Berechnung der Geraden (Optimierungsproblem)

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

$$a = \bar{y} - b\bar{x}$$

Wobei \bar{x} und \bar{y} die Mittelwerte der jeweiligen Daten sind. Dieser Gerade $y = a + bx$ wird auch Regressionsgerade genannt.

Lineare Regression in R

```
seiten <- seq(50, 500, 50)
preis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5, 26.1, 29.1)
lm(preis ~ Seiten)

##
## Call:
## lm(formula = preis ~ Seiten)
##
## Coefficients:
## (Intercept) Seiten
## 6.04000 0.04673
```

Der Befehl `lm()` steht für «linear model». Mit dem Befehl `lm(x~y)` passt R ein Modell von der Form $y = a + bx$ an die Daten an. R findet uns also $a = 6.04$ und $b = 0.04673$.

Plotten der Regressionsgerade

```
Seiten <- seq(50, 500, 50)
preis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5, 26.1, 29.1)
plot(Seiten, preis,
     col = "orange",
     pch = 19,
     xlab = "Seitenzahl",
     ylab = "Buchpreis")
)
abline(lm(preis ~ Seiten), col = "deepskyblue")
```

Mit diesem Modell können Preise, z.B. für Buchpreise mit Seitenzahlen berechnen, die in der Preistabelle nicht vorkommen.

Ein Beispiel mit 365 Seite kann folgendermassen berechnet werden:

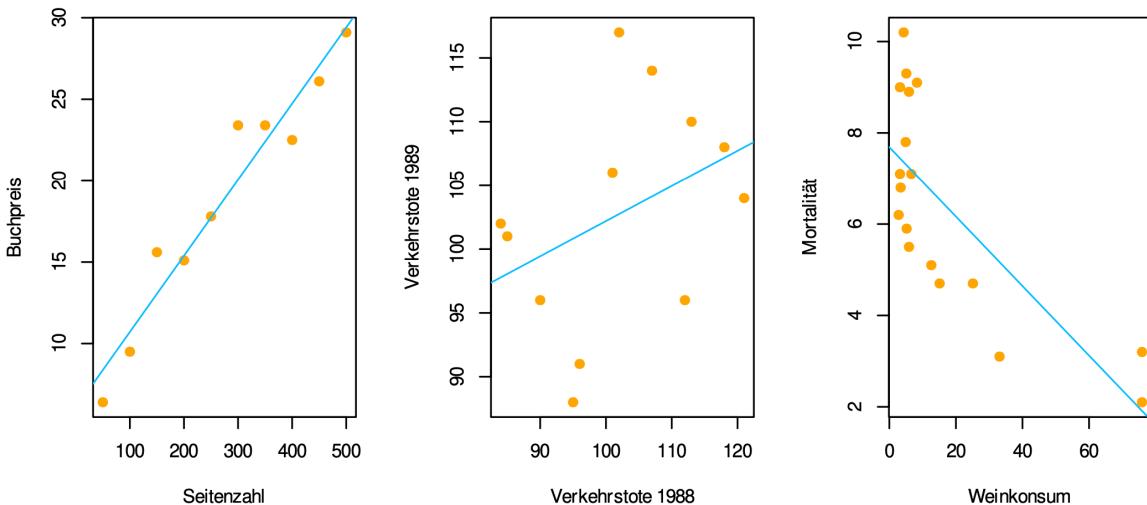
$$y = 6.04 + 0.04673 * 375 \approx 23.60$$

Das Buch dürfe somit ungefähr CHF 23.60 kosten. Jedoch muss man beachten, dass man bei einer Extrapolation sehr vorsichtig sein muss. Denn ein unglaublich dickeres Buch (wir haben Daten bis 500 Seiten), kann es sein, dass noch weitere Artefakte ins Spiel kommen und das Modell somit nicht mehr stimmt.

Korrelation Wahrscheinlichkeitsmodell

Wie gut passt die Regressionsgerade?

Die Regressionsgerade kann fast immer bestimmt werden.



Letzten beiden Beispiele: Die Regressionsgerade sagt sehr wenig über die wirkliche Verteilung der Punkte im Streudiagramm aus. Dafür gibt es zwei Gründe:

- Die Punkte folgen scheinbar keiner Gesetzmässigkeit
- Die Punkte folgen einer nichtlinearen Gesetzmässigkeit

Wie kann man feststellen, ob ein linearer Zusammenhang der Daten besteht oder nicht?

- Möglichkeit: Situation graphisch betrachten
- Ziel: Wert angeben, der den Zusammenhang numerisch beschreibt

Empirische Korrelation

Der numerische Wert der linearen Abhängigkeit von zwei Größen:

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})} * \sqrt{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})}}$$

Die empirische Korrelation ist die Dimensionslose Zahl zwischen -1 und +1. Sie misst die Stärke und Richtung der linearen Abhängigkeit zwischen Daten x und y.

Ist $r = +1$: dann liegen die Punkte auf einer steigenden Geraden:

$$y = a + bx \text{ mit } a \in \mathbb{R} \text{ und ein } b > 0$$

Ist $r = -1$: dann liegen die Punkte auf einer fallenden Geraden:

$$y = a + bx \text{ mit } a \in \mathbb{R} \text{ und ein } b < 0$$

Ist $r = 0$: dann sind x und y unabhängig und haben keinen Zusammenhang

Begründung

Wir wollen die Eigenschaften nicht herleiten, sondern nur graphisch veranschaulichen. Die Kovarianz ist der Zähler der Korrelation.

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})} * \sqrt{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})}}$$

Wenn diese Grösse:

- Grösser als 0 ist, wenn die Punktewolke einer steigenden Geraden folgt
- Kleiner als 0 ist, wenn die Punktewolke einer fallenden Geraden folgt
- 0 ist, wenn kein Zusammenhang vorhanden ist
- 0 sein kann, wenn kein linearer Zusammenhang vorhanden ist

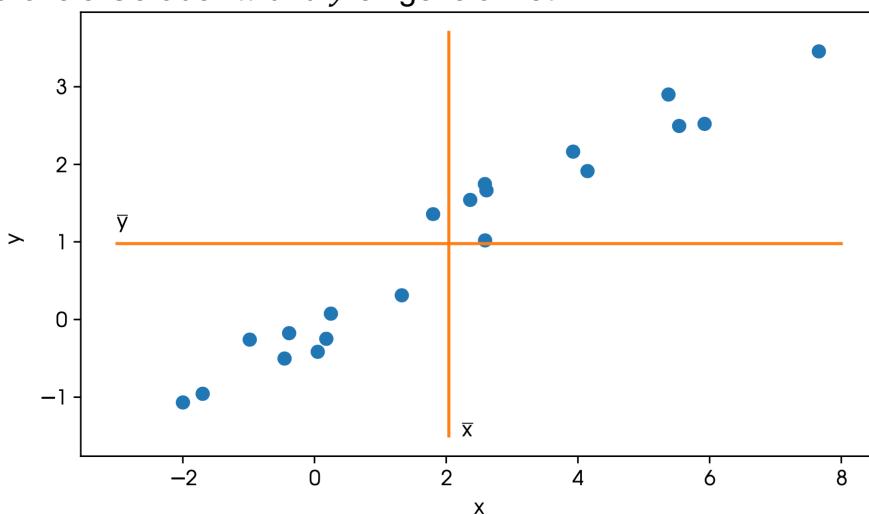
Nenner der Korrelation:

$$r = \frac{\sqrt{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})} * \sqrt{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})}}{\sqrt{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})}}$$

Dieser Wert ist immer positiv und normalisiert die Kovarianz, damit die Werte zwischen -1 und +1 liegen.

Beispiel

Die Datenpunkte folgen mehr oder weniger einer Geraden. Zu den Koordinatenachsen parallele Geraden \bar{x} und \bar{y} eingezeichnet.



- Zähler der Korrelation:

$$\underbrace{(x_1 - \bar{x})}_{x_1^*} \underbrace{(y_1 - \bar{y})}_{y_1^*} + \dots + \underbrace{(x_n - \bar{x})}_{x_n^*} \underbrace{(y_n - \bar{y})}_{y_n^*}$$

- Von x-Koordinaten der Punkte wird der Durchschnitt \bar{x} subtrahiert
- Von y-Koordinaten der Punkte wird der Durchschnitt \bar{y} subtrahiert

Wir haben im Mittelwert von x und y ein neues Koordinatensystem gezeichnet. Das neue Koordinatensystem wird. Der Ursprung ist nun die Mitte der Punktewolke – sprich der Ursprung wird in (\bar{x}, \bar{y}) verschoben.

- Zähler der Korrelation sieht dann wie folgt aus:

$$x_1^*y_1^* + x_2^*y_2^* + \dots + x_n^*y_n^*$$

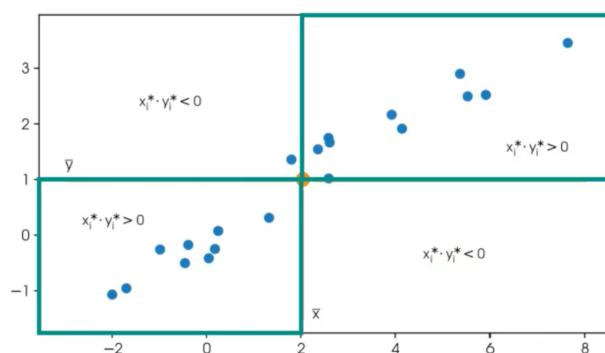
- Was bedeutet dies für die Punkte in Folie ???

- ▶ Punkte im I. Quadranten (rechts oben):
 - ★ Positive Koordinaten
 - ★ $x_i^* \cdot y_i^*$ positiv
- ▶ Punkte im II. Quadranten (links oben):
 - ★ Negative x-Koordinate und eine positive y-Koordinate
 - ★ $x_i^* \cdot y_i^*$ negativ
- ▶ Punkte im III. Quadranten (links unten):
 - ★ Negative x-Koordinate und negative y-Koordinate
 - ★ $x_i^* \cdot y_i^*$ positiv
- ▶ Punkte im IV. Quadranten (rechts unten):
 - ★ Positive x-Koordinate und negative y-Koordinate
 - ★ $x_i^* \cdot y_i^*$ negativ

Die empirische Korrelation misst in welchen Quadranten die Punkte liegen.

Punkte folgen einer steigenden Geraden:

Datenpunkte folgen mehr oder weniger einer Geraden



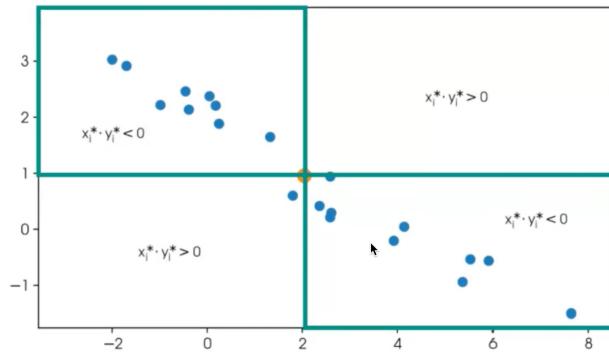
- Fast alle Punkte im I. und III. Quadranten

$$x_i^*y_i^* > 0$$

$$\text{Somit auch } x_1^*y_1^* + x_2^*y_2^* + \dots + x_n^*y_n^* > 0$$

Punkte folgen einer fallenden Gerade:

Datenpunkte folgen mehr oder weniger einer Geraden

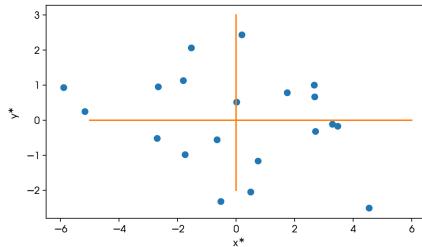


- Fast alle Punkte im II. und IV. Quadranten

$$x_i^* y_i^* < 0$$

$$\text{Somit auch } x_1^* y_1^* + x_2^* y_2^* + \dots + x_n^* y_n^* < 0$$

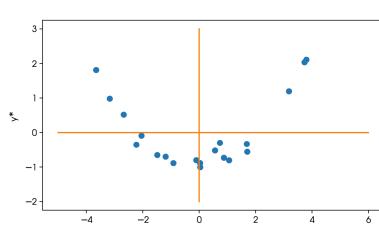
Die Punkte weisen keinen Zusammenhang auf:



- Produkte $x_i^* y_i^*$ über alle Punkte aufaddiert heben sich in etwa auf
 - Hälften aller Punkte im I. und III. Quadranten (Produkte positiv)
 - Andere Hälften im II. und IV. Quadranten (Produkte negativ)
 - Produkte betragsmässig ähnlich
 - Damit gilt

$$x_1^* y_1^* + x_2^* y_2^* + \dots + x_n^* y_n^* \approx 0$$

Die Korrelation liegt bei 0, wenn wir keine Struktur haben. Es kann aber trotzdem nicht erkannt werden, wenn wir z.B. einen quadratischen Zusammenhang haben. Wenn wir also ein 0 bekommen, muss es nicht heißen, dass es keinen Zusammenhang gibt. Es kann sein dass es keinen linearen Zusammenhang gibt.



- Beträge der Produkte links und rechts von der y-Achse auf
- Es gilt

$$x_1^* y_1^* + x_2^* y_2^* + \dots + x_n^* y_n^* \approx 0$$
- Korrelationskoeffizient erkennt nur *lineare* Zusammenhänge

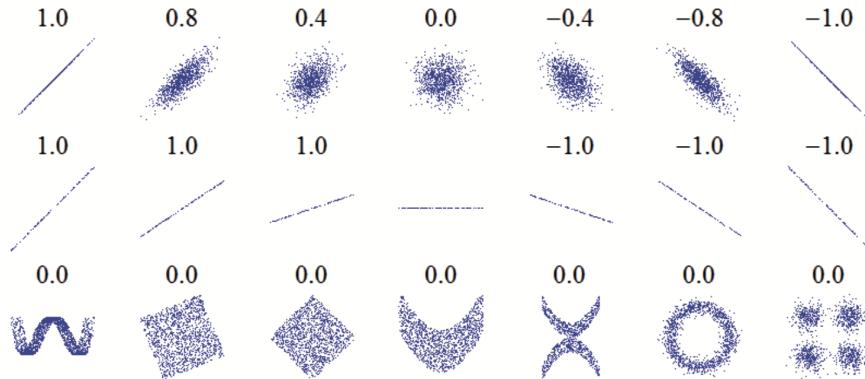
Berechnung in R

```
cor(seiten, preis)  
## [1] 0.9681122
```

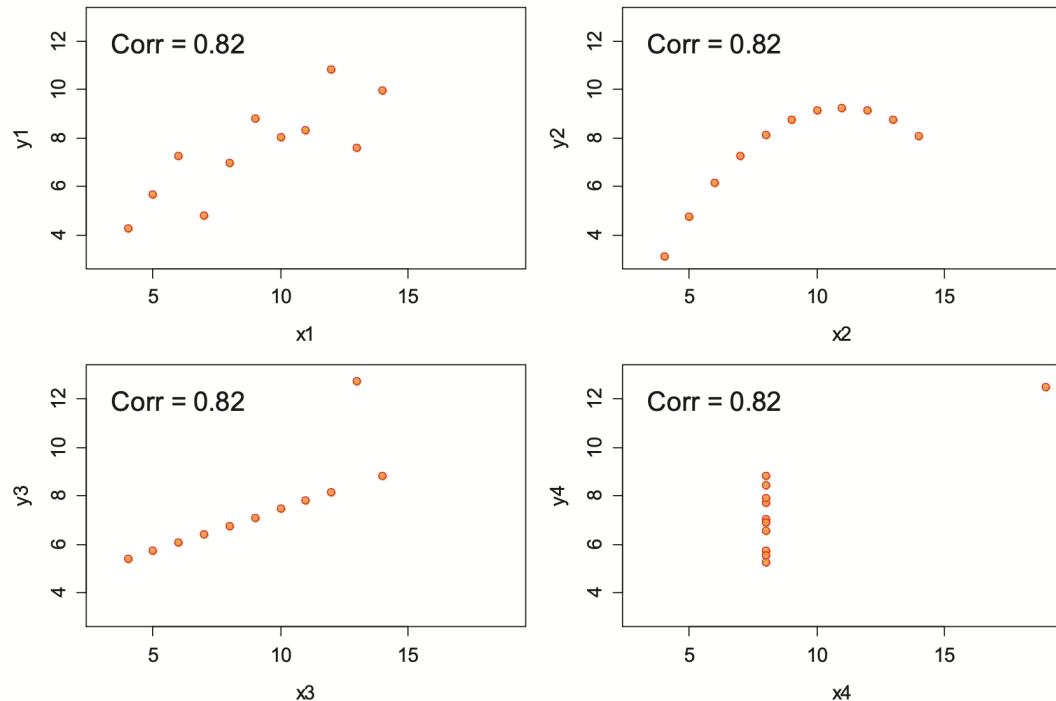
Liegt der Wert sehr nahe bei 1 liegt ein starker linearer Zusammenhang vor. Ist der Wert positiv, «je mehr desto mehr»-Zusammenhang.

Bemerkungen

Es stellt sich die Frage, was «nahe» bei 0 oder «nahe» bei 1 heisst. Das lässt sich allgemein nicht sagen und hängt vom Problem und dem Gebiet ab. Die Korrelation misst «nur» den linearen Zusammenhang. Daher sind die Daten immer auch graphisch zu betrachten, statt sich blind auf Kennzahlen zu verlassen. Symmetrien können beispielsweise nicht erkannt werden.



Die Korrelation kann gleich sein, aber die Streudiagramme können sehr unterschiedlich aussehen.



Wahrscheinlichkeit

Alle haben ein intuitives Gefühl, was die Wahrscheinlichkeit ist. Die Wahrscheinlichkeit mit einem fairen Würfel eine 4 zu würfeln, ist ein Sechstel. Aber die Interpretation ist überraschend schwierig.

Wahrscheinlichkeitsmodell

Es sind Zufallsexperimente, bei welchem der Ausgang nicht exakt vorhersehbar ist:

- Würfelwurf
- Münzwurf
- Anzahl Anrufe in einem Callcenter

Ein Wahrscheinlichkeitsmodell besteht aus Ereignissen, die in einem solchen Experiment möglich sind und den Wahrscheinlichkeiten, die die verschiedenen Ergebnisse haben.

Beispiel Würfel werfen:

- Mögliche Ergebnisse: 1, 2, 3, 4, 5, 6
- Wahrscheinlichkeit einer dieser Zahl zu werfen: 1/6 (sofern Würfel fair)

Das Wahrscheinlichkeitsmodell hat folgende Komponenten:

- Grundraum Ω : Enthält alle möglichen Elementarereignisse ω
- Ereignisse A, B, C: Teilmengen des Grundraums
- Wahrscheinlichkeiten P, die zu den Ereignissen A, B, C gehören

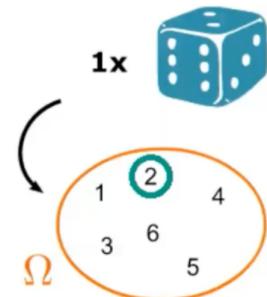
Elementarereignisse sind mögliche Ergebnisse (Ausgänge) des Experiments. Zusammen bilden diese den Grundraum.

$\Omega = \{\text{mögliche Elementarereignisse } \omega\} = \text{mögliche Ausgänge / Resultat}$

Grundraum (die möglichen Ergebnisse)

$$\underline{\Omega = \{1, 2, 3, 4, 5, 6\}}$$

- Element $\underline{\omega = 2}$ ist ein Elementarereignis



Bedeutung: Beim Würfeln wurde die Zahl 2 geworfen

Zahl 7: *Kein* Elementarereignis, da nicht im Grundraum Ω

Beispiel Callcenter

- Anzahl Anrufe in einer Stunde in einem Callcenter
- Grundraum (zumindest theoretisch beliebig viele Anrufe möglich):

$$\Omega = \{0, 1, 2, 3, 4, \dots\}$$

- Elementarereignis $\omega = 6$: 6 Anrufe in einer Stunde

Beispiel Münzwurf

- 2-maliges Werfen einer Münze
- Bezeichnungen K : „Kopf“ und Z : „Zahl“
- Alle möglichen Ergebnisse des Experiments (Grundraum)

$$\Omega = \{KK, KZ, ZK, ZZ\}$$

- Elementarereignis ist z.B. $\omega = KZ$

Ereignis

Ereignisse sind allgemeiner und wichtiger als Elementarereignisse, bestehen aber aus diesem Ereignis A: Teilmenge von Ω : $A \subset \Omega$. «Ein Ereignis A tritt ein» bedeutet, dass Ergebnis ω des Experiments zu A gehört.

Beispiel: 2-maliges Werfen einer Münze

- Ereignis A, wo genau einmal Kopf geworfen wird
- Ereignis A besteht aus Elementarereignissen Kopf-Zahl und Zahl-Kopf
- Ereignis A ist dann die Menge $A = \{KZ, ZK\}$
- Werden Zahl-Zahl: Das Ereignis A ist nicht eingetreten
- Wahrscheinlichkeit, dass A eintritt (falls Münze fair): $P(A) = \frac{2}{4} = \frac{1}{2}$
- In der Statistik werden Wahrscheinlichkeiten oft mit P oder p bezeichnet

Beispiel: Würfeln

Das Ereignis A: «Eine ungerade Zahl würfeln».

- Dann ist $A = \{1, 3, 5\}$
- Ereignis A tritt ein, wenn z.B. Zahl 5 gewürfelt wird
- Wahrscheinlichkeit, dass A eintritt (fairer Würfel): $P(A) = \frac{3}{6} = \frac{1}{2}$

- Ereignis B = «Eine Zahl ist kleiner als 7 Würfeln»
- Das ist immer der Fall und somit ist $B = \Omega$
- B heisst dann sicheres Ereignis
- Die Wahrscheinlichkeit, dass B eintritt (fairer Würfel): $P(B) = \frac{6}{6} = 1$

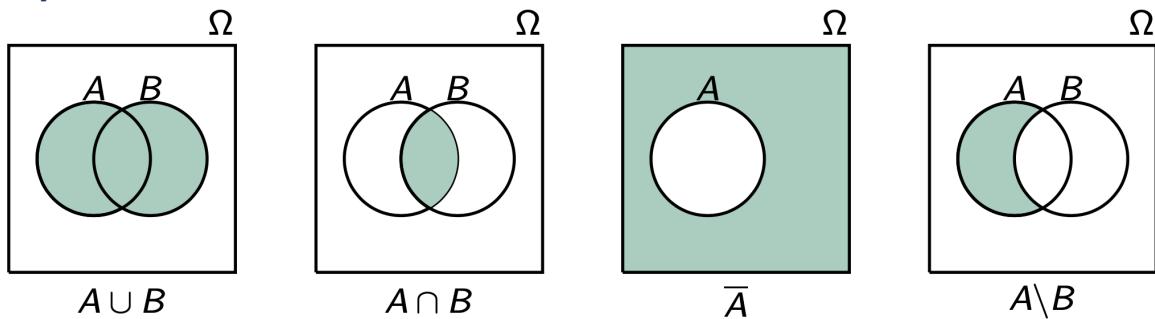
- Ereignis C: «Die Zahl 7 würfeln»
- Dies ist unmöglich, Menge C = {}
- Die Leere Menge {} enthält kein Element
- Heisst es ist ein unmögliches Ereignis
- Die Wahrscheinlichkeit, dass C eintritt (fairer Würfel): $P(C) = \frac{0}{6} = 0$

Neue Mengen aus Bekannten

Operationen der Mengenlehre für Ereignisse

Name	Symbol	Bedeutung
Vereinigung	$A \cup B$	A oder B, nicht-exklusives „oder“
Schnittmenge	$A \cap B$	A und B
Komplement	\bar{A}	nicht A
Differenz	$A \setminus B = A \cap \bar{B}$	A ohne B

Graphisch

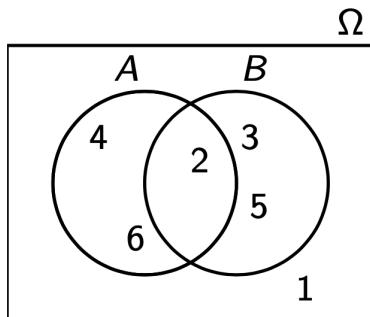


Beispiel: Würfelwurf

Das Ereignis A: Die geworfene Zahl ist gerade: $A = \{2, 4, 6\}$

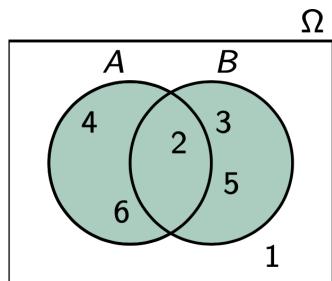
Das Ereignis B: Die geworfene Zahl ist Primzahl: $B = \{2, 3, 5\}$

$\Omega = \{1, 2, 3, 4, 5, 6\}$



Vereinigung

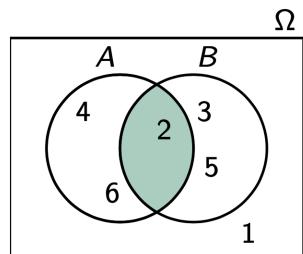
Alle Elemente, die entweder in A oder in B oder in beiden Mengen vorkommen:
 $A \cup B = \{2, 3, 4, 5, 6\}$



Element 2 kommt in Menge A und in der Menge B vor

Schnittmenge

Alle Elemente, die in A und in B vorkommen:
 $A \cap B = \{2\}$

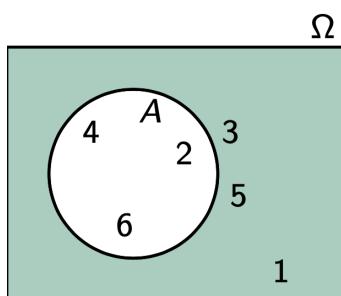


Element 2 einziges Element, das sowohl in Menge A wie auch in Menge B vorkommt.

Komplement

Alle Elemente von Ω , die nicht in der entsprechenden Menge vorkommen:

$$\bar{A} = \{1, 3, 5\}; \bar{B} = \{1, 4, 6\}$$

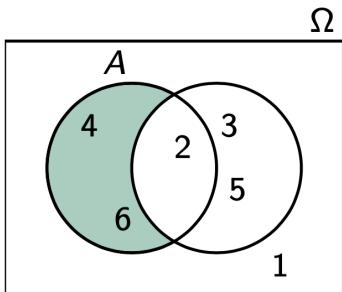


Menge \bar{A} sind die ungeraden Zahlen

Differenz

Alle Elemente der Menge A, die aber nicht in der Menge B vorkommen:

$$A \setminus B = \{4, 6\}$$



Element 2 kommt sowohl in A wie auch in B vor und gehört deswegen nicht zur Differenz.

Axiome der Wahrscheinlichkeit

Eigenschaften der Wahrscheinlichkeit:

Kolmogorov Axiome der Wahrscheinlichkeitsrechnung

Jedem Ereignis A wird eine W'keit $P(A)$ zugeordnet, mit:

A1: $P(A) \geq 0$

A2: $P(\Omega) = 1$

A3: $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \{\}$

Bezeichnung $P(A)$: Die Wahrscheinlichkeit, dass das Ereignis A eintritt

Ereignis A: «ungerade Zahl würfeln» (bei fairem Würfel)

$$P(A) = \frac{1}{2}$$

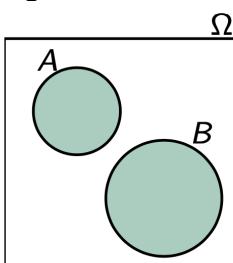
Buchstabe P steht für probabiliy.

Eigenschaften

A1: Wahrscheinlichkeit kann nicht negativ sein

A2: Mit $P(\Omega) = 1$: Wahrscheinlichkeiten eines Ereignisses zwischen 0 und 1

A3: Für zwei disjunkte Ereignisse: Die Wahrscheinlichkeit, dass eines der beiden eintritt, gleich Addition Wahrscheinlichkeiten der beiden Ereignisse. A3 gilt nicht, falls die Ereignisse nicht disjunkt sind.



- Beispiel Würfel fair: $A = \{2, 4, 6\}$, $B = \{2, 3, 5\}$
- Dann $P(A)=P(B)=\frac{1}{2}$
- $A \cup B = \{2, 3, 4, 5, 6\}$
- Wenden A3 an:
 - o $P(A \cup B) = P(A) + P(B) = 1 + 1 = 1 \neq 2$
- Kann nicht sein, da $P(A \cup B) = 5/6$
- Grund: $A \cap B = \{2\} \neq \emptyset$

Beispiel

Wurf zweier Münzen:

$$\Omega = \{KK, KZ, ZK, ZZ\}$$

Plausibel: Alle 4 Elemente gleich wahrscheinlich (faire Münze)

Wegen $P(\Omega) = 1$ müssen sich die Wahrscheinlichkeiten zu eins aufaddieren:

$$P(KK) + P(KZ) + P(ZK) + P(ZZ) = 1$$

Da alle Elementarereignisse gleich wahrscheinlich sind:

$$P(KK) = P(KZ) = P(ZK) = P(ZZ) = \frac{1}{4}$$

Rechenregeln aus Axiomen

Rechenregeln

Sind A, B und A_1, \dots, A_n Ereignisse, dann gilt

$$P(\overline{A}) = 1 - P(A) \quad \text{für jedes } A$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{für beliebige } A \text{ und } B$$

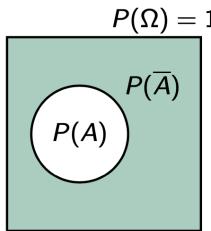
$$P(A_1 \cup \dots \cup A_n) \leq P(A_1) + \dots + P(A_n) \quad \text{für beliebige } A_1, \dots, A_n$$

$$P(B) \leq P(A) \quad \text{für beliebige } A \text{ und } B \text{ mit } B \subseteq A$$

$$P(A \setminus B) = P(A) - P(A \cap B) \quad \text{für beliebige } A \text{ und } B \text{ mit } B \subseteq A$$

Die Wahrscheinlichkeiten als Flächen im Venn-Diagramm vorstellen. Die Totalfläche von Ω gleich 1 oder $P(\Omega) = 1$.

1. Regel



$P(A)$: Flächeninhalt der Fläche A

$P(\bar{A})$: Flächeninhalt der restlichen Fläche in Ω

Es gilt also offensichtlich:

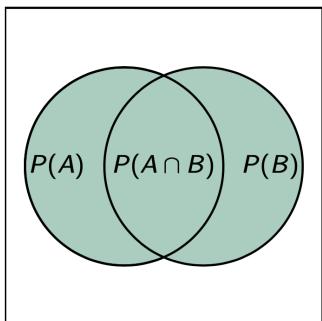
$$P(A) + P(\bar{A}) = P(\Omega) = 1$$

Und somit:

$$P(\bar{A}) = 1 - P(A)$$

2. Regel

$$P(\Omega) = 1$$



Schnittmengen $A \cap B$ mit $P(A) + P(B)$ doppelt gezählt: Einmal abziehen:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

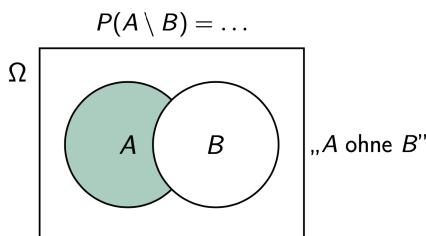
Diese Regel ist die Verallgemeinerung von Axiom A3. Sind die Mengen A und B disjunkt, so gilt:

$$A \cap B = \emptyset$$

Mit der Regel von oben:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(\emptyset) = P(A) + P(B)$$

Knobelaufgabe



1. $P(A) - P(B)$

2. $P(A) + P(B)$

3. $P(A) - P(A \cap B)$

4. $P(A) + P(B) - P(A \cap B)$

Lösung: Nr. 3

Diskrete Wahrscheinlichkeitsmodelle

Der Grundraum ist endlich oder unendlich und diskret. Der Begriff diskret ist eine Endliche Menge, wie z.B. $\Omega = \{0, 1, \dots, 10\}$. Eine Unendliche Menge, aber trotzdem diskrete Menge ist z.B. $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$. Die Menge $\Omega = \mathbb{R}$ (Menge aller Dezimalbrüche, Zahlengerade) ist nicht diskret. Sie wird später für Messdaten eine wichtige Rolle spielen.

Die Berechnung von Wahrscheinlichkeiten für diskrete Modelle ist wie folgt

Im diskreten Fall ist die Wahrscheinlichkeit eines Ereignisses

$$A = \{w_1, w_2, \dots, w_n\}$$

durch die Wahrscheinlichkeiten der zugehörigen Elementarereignisse $P(w)$ festgelegt:

$$P(A) = P(w_1) + P(w_2) + \dots + P(w_n)$$

Alle Wahrscheinlichkeiten der Elementarereignisse aus Ereignis A werden addiert.

Beispiel: Unfairer Würfel

- W'keiten, unterschiedliche Zahlen zu werfen, sind nicht gleich:

ω	1	2	3	4	5	6
$P(\omega)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$

- Es gilt:

$$\begin{aligned} P(\Omega) &= P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{12} + \frac{1}{4} + \frac{1}{12} + \frac{1}{12} \\ &= 1 \end{aligned}$$

- Für Ereignis $A = \{1, 2, 4\}$ gilt:

$$\begin{aligned} P(A) &= P(1) + P(2) + P(4) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{4} \\ &= \frac{3}{4} \end{aligned}$$

- Beachte: Resultat nicht gleich, wenn Würfel fair wäre: $\frac{1}{2}$

- Bsp: Berechne W'keit, eine Zahl kleiner als 6 zu würfeln

- Ereignis B :

$$B = \{1, 2, 3, 4, 5\}$$

- Zugehörige W'keit:

$$\begin{aligned} P(B) &= P(1) + P(2) + P(3) + P(4) + P(5) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{12} + \frac{1}{4} + \frac{1}{12} \\ &= \frac{11}{12} \end{aligned}$$

- Einfacher mit Gegenw'keit: $P(\bar{B})$
- 1. Rechenregel: Komplement \bar{B} von B :

$$\bar{B} = \{6\}$$

- Dann gilt:

$$P(B) = 1 - P(\bar{B}) = 1 - P(6) = 1 - \frac{1}{12} = \frac{11}{12}$$

(Gegenereignis)

Modell von Laplace

Das Modell geht von einer Annahme aus, dass alle Elementarereignisse die gleiche Wahrscheinlichkeit haben. $E = \{w_1, w_2, \dots, w_g\}$

Wir haben also den Grundraum mit m Elementen.

Die Wahrscheinlichkeiten addieren sich zu 1 und deshalb gilt:

$$P(w_k) = \frac{1}{|\Omega|} = \frac{1}{m}$$

Für ein Ereignis E im Laplace Model gilt also:

$$P(E) = \frac{g}{m} = \sum_{k: w_k \in E} P(\{w_k\})$$

Man teilt die Anzahl günstigen Elementarereignisse durch die Anzahl der möglichen Elementarereignisse.

Beispiel: Laplace Modell

- Es werden zwei verschiedene (blau und rot) Würfel geworfen
- Wie gross ist die W'keit, dass die Augensumme 7 ergibt?
- Elementarereignis beschreibt die Augenzahlen auf beiden Würfeln
- Ergebnis in der Form **14** schreiben
- Ergebnis **14** ist *nicht* gleich **41**
- Elementarereignisse:

$$\Omega = \{11, 12, \dots, 16, 21, \dots, 65, 66\}$$

- Anzahl Elementarereignisse:

$$|\Omega| = 36$$

- Ereignis E : Augensumme 7 wird gewürfelt
- Es gibt davon 6 Elementarereignisse:

$$E = \{16, 25, 34, 43, 52, 61\}$$

- Alle Elementarereignisse gleich wahrscheinlich: W'keit für Ereignis E :

$$P(E) = \frac{|E|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

Stochastische Unabhängigkeit

Wir haben gesehen, dass man die Wahrscheinlichkeit von zwei Unabhängigen Ereignissen berechnen können.

$$P(A \cup B) = P(A) + P(B) - (P(A \cap B))$$

Wie berechnet man $(P(A \cap B))$?

Dafür gibt es keine allgemeine Regel. Wenn die Wahrscheinlichkeit $P(A)$ und $P(B)$ bekannt ist, so ist die Wahrscheinlichkeit $(P(A \cap B))$ im Allgemeinen nicht aus $P(A)$ und $P(B)$ berechenbar.

Sind die Ereignisse A und B stochastisch unabhängig sind, so gilt:

$$P(A \cap B) = P(A) * P(B)$$

Was heisst stochastisch unabhängig?

Der Ausgang des Ereignisses A hat keinen Einfluss auf den Ausgang des Ereignisses B und umgekehrt.

Beispiele Stochastische Unabhängigkeit

- Ereignis A: Mit fairem Würfel eine eins oder zwei zu werfen
- Ereignis B: Kopf beim Werfen einer fairen Münze
- Werfen einer Münze keinen Einfluss auf das Resultat beim Würfelwurf
- Formel oben verwenden:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

-
- Ereignis E: Tokyo wird an bestimmten Tag durch Erdbeben erschüttert
 - Ereignis F: An diesem Tag fegt ein Taifun über die Stadt
 - Erdbeben wohl kaum Einfluss auf das Entstehen eines Taifuns
 - Beide Ereignisse sind also stochastisch unabhängig

 - Werfen eine Münze zweimal nacheinander
 - Resultat des ersten Wurfes hat kaum Einfluss auf Resultat des zweiten Wurfes
 - Dies allerdings nur richtig, wenn Münze ideal
 - Reale Münze: Durch Aufprall minimalste Veränderungen
 - Diese haben Einfluss auf die Wurfw'keit für Kopf (oder Zahl) beim nächsten Wurf
 - Veränderungen aber so klein, dass sie vernachlässigbar sind

Gegenbeispiel:

Diese beiden Ereignisse sind nicht stochastisch unabhängig (stochastische Abhängigkeit).

- In Topf 20 Lose mit 5 Gewinnen
- Ziehen zweimal hintereinander *ohne Zurücklegen*
- Ereignis A: Gewinn beim ersten Ziehen
- Ereignis B: Gewinn beim zweiten Ziehen
- Diese beiden Ereignisse sind *nicht* stochastisch unabhängig
- Ziehen beim ersten Ziehen ein Gewinnlos: W'keit, dass A eintrifft:

$$P(A) = \frac{5}{20}$$

- Bei 2. Ziehung fehlt ein Gewinn: W'keit dann zu gewinnen:

$$P(B) = \frac{4}{19}$$

- Ziehen ersten Ziehung Niete: W'keit bei der 2. Ziehung zu gewinnen:

$$P(B) = \frac{5}{19}$$

- Je nachdem, ob Ereignis A eintrifft oder nicht, ändert sich die W'keit für das Eintreffen von B
- Ereignisse sind also nicht stochastisch unabhängig
- Ereignis A: Morgen ist schönes Wetter
- Ereignis B: Person hat morgen gute Laune
- Die meisten Menschen sind bei schönem Wetter besser aufgelegt, als bei schlechtem Wetter
- Eintreffen von A hat Einfluss auf Eintreffen von B
- Die Ereignisse sind nicht stochastisch unabhängig

Zusammenfassend also:

- Formel

$$P(A \cap B) = P(A) \cdot P(B)$$

gilt *nur*, falls Ereignisse A und B stochastisch unabhängig sind

- Sind die Ereignisse *nicht* stochastisch unabhängig, so gibt es keine allgemeine Formel für $P(A \cap B)$

Zufallsvariable Wahrscheinlichkeitsverteilung

Zufallsvariablen werden als Reelle Zahlen R angegeben. Dies sind alle Punkte der Zahlengeraden und alle Dezimalbrüche. Die Notation für Zufallsvariablen ist X, Y, Z, \dots

Die Funktionen werden als $y = f(x)$ angegeben.

Die Zufallsvariable wird in Grossbuchstaben (X, Y, Z) angegeben. Als Kleinbuchstaben werden die konkreten Werte angegeben (x, y, z). Ein Ereignis, bei welchem die Zufallsvariable X den Wert x annimmt, wird als $X = x$ angegeben.

Nicht die Funktion X ist zufällig, sondern nur das Argument ω (Input). Je nach Ausgang ω des Zufallsexperiments, entsteht ein anderer Wert $x = X(\omega)$.

Das x heisst auch eine Realisierung der Zufallsvariablen X .

Definition

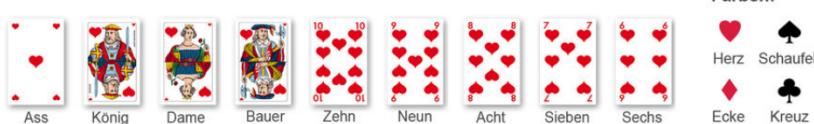
Die Werte einer Zufallsvariablen X (die möglichen Realisierungen von X) treten mit gewissen Wahrscheinlichkeiten auf. Die Wahrscheinlichkeiten, dass X den Wert x annimmt, berechnet sich wie folgt:

$$P(X = x) = P(\{\omega \mid X(\omega) = x\}) = \sum_{\omega; X(\omega)=x} P(\omega)$$

Beispiel Zufallsvariable

Die Zufallsvariable ist in der Statistik ein zentraler Begriff.

- Pack Spielkarten: 36 verschiedene Karten



Ziehen nacheinander drei Karten und legen nach jedem Ziehen zurück
Machen dies zweimal und erhalten folgendes Resultat:

- 1 6, Dame, König
- 2 8, Bube, Ass

Frage: Welcher der Versuche ist „besser“?

- In dieser Form schwer vergleichbar

- Eine Lösung: Einzelnen Spielkarten werden Werte zugeordnet

- ▶ 6, 7, 8, 9 haben Wert 0
- ▶ 10 hat den Wert 10
- ▶ Bube hat den Wert 2
- ▶ Dame hat den Wert 3
- ▶ König hat den Wert 4
- ▶ Ass hat den Wert 11

Jetzt sind Ziehungen miteinander vergleichbar:

- ① 6, Dame, König → $0 + 3 + 4 = 7$
- ② 8, Bube, Ass → $0 + 2 + 11 = 13$

- ▶ Zweite Ziehung mit diesen Werten ist besser als die erste

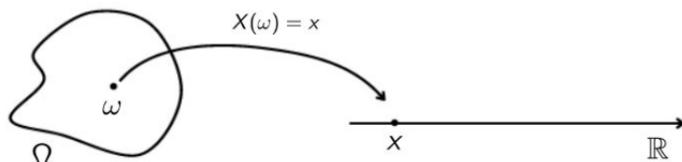
Situation vorher kommt in der Statistik häufig vor

Zufallsexperiment mit dem Grundraum Ω

- Allen Elementarereignissen von Ω wird eine Zahl zugeordnet
Zu jedem Elementarereignis ω gehört demnach eine Zahl
- *Funktion X , die jedem Elementarereignis ω den Zahl x zugeordnet*

$$X(\omega) = x$$

wird Zufallsvariable genannt



Situation vorher kommt in der Statistik häufig vor

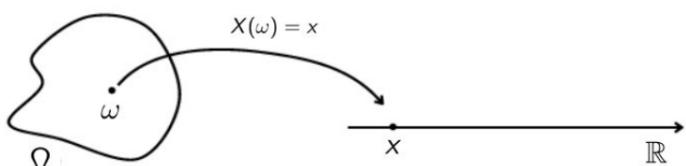
Zufallsexperiment mit dem Grundraum Ω

- Definition:

Zufallsvariable

Eine Zufallsvariable X ist eine Funktion:

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$



Ziehen Karten aus einem Stapel Spielkarten

Jeder Karte wird eine Zahl zugeordnet: $\omega = \begin{array}{|c|} \hline \text{Heart} \\ \hline \end{array} \mapsto X(\omega) = 11$

$\omega = \begin{array}{|c|} \hline \text{King} \\ \hline \end{array} \mapsto X(\omega) = 4$

⋮

$\omega = \begin{array}{|c|} \hline \text{Red} \\ \hline \end{array} \mapsto X(\omega) = 0$

Mit Zahlen $X(\omega)$ kann der Durchschnitt gezogener Karten berechnet werden

- Durchschnitt von „6, Dame, König“ gleich $\frac{7}{3}$

Für Elementarereignisse „6“, „Dame“ und „König“ ohne Zahlen macht das Wort „Durchschnitt“ keinen Sinn

Werfen gemeinsam einen blauen und roten Würfel



Grundraum Ω sind die Augenzahlen der Würfel:

$$\Omega = \{11, 12, \dots, 16, 21, 22, \dots, 26, \dots, 66\}$$

11	12	13	14	15	16	17
12	13	14	15	16	17	18
13	14	15	16	17	18	19
14	15	16	17	18	19	20
15	16	17	18	19	20	21
16	17	18	19	20	21	22
17	18	19	20	21	22	23
18	19	20	21	22	23	24
19	20	21	22	23	24	25
20	21	22	23	24	25	26
21	22	23	24	25	26	27
22	23	24	25	26	27	28
23	24	25	26	27	28	29
24	25	26	27	28	29	30
25	26	27	28	29	30	31
26	27	28	29	30	31	32
27	28	29	30	31	32	33
28	29	30	31	32	33	34
29	30	31	32	33	34	35
30	31	32	33	34	35	36

Auf Ω sind verschiedene Zufallsvariablen definierbar

- Sei X die Zufallsvariable für die Summe der Augenzahlen:

- Dann gilt: $X(16) = 7$ oder $X(31) = 4$
- Die Werte, die die Zufallsvariable annehmen kann, wird als Wertemenge bezeichnet: $\mathbb{W}_X = \{2, 3, 4, \dots, 11, 12\}$

Werfen gemeinsam einen blauen und roten Würfel



Grundraum Ω sind die Augenzahlen der Würfel:

$$\Omega = \{11, 12, \dots, 16, 21, 22, \dots, 26, \dots, 66\}$$

11	12	13	14	15	16	17
12	13	14	15	16	17	18
13	14	15	16	17	18	19
14	15	16	17	18	19	20
15	16	17	18	19	20	21
16	17	18	19	20	21	22
17	18	19	20	21	22	23
18	19	20	21	22	23	24
19	20	21	22	23	24	25
20	21	22	23	24	25	26
21	22	23	24	25	26	27
22	23	24	25	26	27	28
23	24	25	26	27	28	29
24	25	26	27	28	29	30
25	26	27	28	29	30	31
26	27	28	29	30	31	32
27	28	29	30	31	32	33
28	29	30	31	32	33	34
29	30	31	32	33	34	35
30	31	32	33	34	35	36

Auf Ω sind verschiedene Zufallsvariablen definierbar

- Sei Y die Augenzahl des roten Würfels:

- Dann gilt: $Y(16) = 6$ oder $Y(31) = 1$ oder $Y(13) = 3$
- Wertemenge: $\mathbb{W}_Y = \{1, 2, \dots, 6\}$

Werfen gemeinsam einen blauen und roten Würfel



Grundraum Ω sind die Augenzahlen der Würfel:

$$\Omega = \{11, 12, \dots, 16, 21, 22, \dots, 26, \dots, 66\}$$

11	12	13	14	15	16	17
12	13	14	15	16	17	18
13	14	15	16	17	18	19
14	15	16	17	18	19	20
15	16	17	18	19	20	21
16	17	18	19	20	21	22
17	18	19	20	21	22	23
18	19	20	21	22	23	24
19	20	21	22	23	24	25
20	21	22	23	24	25	26
21	22	23	24	25	26	27
22	23	24	25	26	27	28
23	24	25	26	27	28	29
24	25	26	27	28	29	30
25	26	27	28	29	30	31
26	27	28	29	30	31	32
27	28	29	30	31	32	33
28	29	30	31	32	33	34
29	30	31	32	33	34	35
30	31	32	33	34	35	36

Auf Ω sind verschiedene Zufallsvariablen definierbar

- Sei Z gleich 0 für alle Elementarereignisse:

- Dann gilt: $Z(16) = 0$ oder $Z(31) = 0$ oder $Z(13) = 0$
- Wertemenge: $\mathbb{W}_Z = \{0\}$

Z ist eine völlig legitime Zufallsvariable

Wie sinnvoll diese ist, ist eine andere Frage

Wählen zufällig eine Person aus

Grundraum Ω sind die Personen dieses Planeten



Auch hier sind viele Zufallsvariablen denkbar:

- ▶ X : Zufallsvariable, die jeder Person das Einkommen zuordnet
- ▶ Y : Zufallsvariable, die jeder Person die Körpergrösse zuordnet
- ▶ Z : Zufallsvariable, die jeder Person das Alter zuordnet

Folgende Variablen sind *keine* Zufallsvariablen:

- ▶ Variable V ordnet jeder Person das Geschlecht zu
- ▶ Variable W ordnet jeder Person die zugehörige Nationalität zu
- „Resultat“ einer Zufallsvariable muss eine Zahl sein

Realisierungen:

- Spielkartenbeispiel:

Realisierung $X = 11$ entspricht dem Ziehen eines Asses

Für Elementarereignisse $\omega =$ 

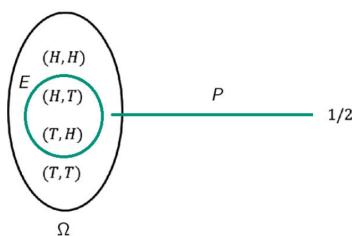
- Würfelbeispiel:

Realisierung $X = 8$ entspricht dem Würfeln der Augensumme 8

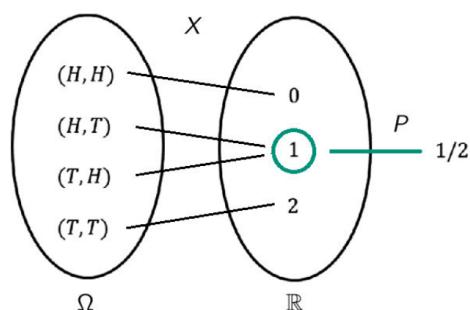
Für Elementarereignisse $\omega =$ 

Wahrscheinlichkeitsverteilung einer Zufallsvariablen

Die Berechnung Wahrscheinlichkeit $P(E)$ eines Ereignisses E :



Entsprechend: Die Wahrscheinlichkeit einer allgemeinen Realisierung x einer Zufallsvariable X



Beispiel Spielkarten

Zufallsvariable X : Wert einer gezogenen Spielkarte

Frage: Wie gross ist W'keit, dass die gezogene Karte den Wert 4 hat?

- Realisierung ist $X = 4$

Zugehörige W'keit ist $P(X = 4)$

- Realisation $X = 4$ entspricht dem Ziehen eines Königs

Gesucht W'keit, dass ein König gezogen wird:

$$P(X = 4) = P(\{\omega \mid \omega = \text{ein König}\}) = \frac{4}{36} = \frac{1}{9}$$

Vorgehen hier ist ausführlicher ausgeführt als unbedingt notwendig, aber es lässt sich auf nicht so einfache Beispiele verallgemeinern

Wahrscheinlichkeitsverteilung

Vorher wollten wir von einer Realisierung die Wahrscheinlichkeit berechnen. Nun möchten wir die Wahrscheinlichkeiten aller Realisierungen berechnen.

Definition

Für jede Realisierung einer Zufallsvariable wird die zugehörige Wahrscheinlichkeit berechnet. → Wahrscheinlichkeitsverteilung dieser Zufallsvariablen.

Die «Liste» von $P(X=x)$ für alle möglichen Werte x_1, x_2, \dots, x_n heisst diskrete Wahrscheinlichkeitsverteilung der diskreten Zufallsvariablen X . Es gilt immer:

$$P(X = x_1) + P(X = x_2) + \dots + P(X = x_n) = 1$$

Mit Summenzeichen:

$$\sum_{\text{aller möglichen } x} P(X = x) = 1$$

Alle Wahrscheinlichkeiten einer Wahrscheinlichkeitsverteilung zusammen addiert ergeben 1.

Beispiel Wahrscheinlichkeitsverteilung

Zufallsvariable X : Wert einer gezogenen Jasskarte

Wahrscheinlichkeit $P(X = 0)$ mit Laplace-Wahrscheinlichkeit:

$$P(X = 0) = \frac{16}{36} = \frac{4}{9}$$

Wahrscheinlichkeit $P(X = 2)$ mit Laplace-Wahrscheinlichkeit:

$$P(X = 2) = \frac{4}{36} = \frac{1}{9}$$

Wahrscheinlichkeit $P(X = 3)$ mit Laplace-Wahrscheinlichkeit:

$$P(X = 3) = \frac{1}{9}$$

Wahrscheinlichkeitsverteilung von X als Tabelle:

Zufallsvariable X : Wert einer gezogenen Jasskarte

x	0	2	3	4	10	11
---	---	---	---	---	----	----

$P(X = x)$	4/9	1/9	1/9	1/9	1/9	1/9
------------	-----	-----	-----	-----	-----	-----

Werte für $P(X = 1)$ oder $P(X = 178)$ sind in der Tabelle nicht aufgeführt. Grund dafür ist, dass diese Werte nicht gezogen werden können. Trotzdem ist eine Wahrscheinlichkeit zuzuordnen, nämlich die Zahl 0:

$$P(X = 1) = 0 \text{ oder } P(X = 178) = 0$$

Die Addition aller Werte der Wahrscheinlichkeitsverteilung ergibt 1. Eine Realisierung muss gezogen werden, es gilt:

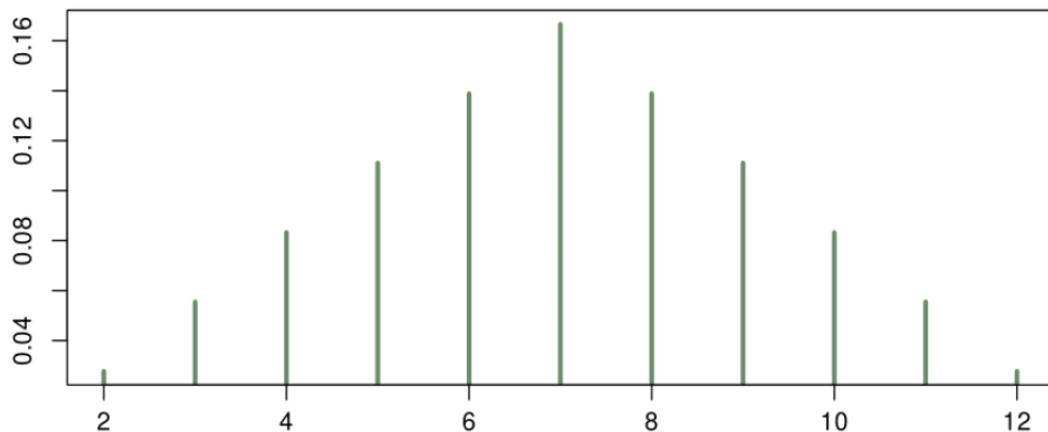
$$P(X = 0) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 10) + P(X = 11) = 1$$

Beispiel Augensumme zweier Würfel

Wahrscheinlichkeitsverteilung für die Zufallsvariable X: (X= Würfelsumme)

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Die Verteilung graphisch dargestellt:



Wie gross ist die Wahrscheinlichkeit, genau die Augensumme 6 zu würfeln? Gesucht ist nun also $P(X = 6) = \frac{5}{36}$.

Wie gross ist die Wahrscheinlichkeit, die Augensumme 6 oder 8 zu würfeln?

$$\text{Gesucht: } P(X = 6) + P(X = 8) = \frac{5}{36} + \frac{5}{36} = \frac{10}{36} = \frac{5}{18}$$

Wie gross ist die Wahrscheinlichkeit, höchstens die Augensumme 3 zu würfeln?

$$\text{Gesucht: } P(X \leq 3) = P(X = 2) + P(X = 3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36} = \frac{1}{12}$$

Wie gross ist die Wahrscheinlichkeit, mindestens die Augensumme 3 zu würfeln?

$$\text{Gesucht: } P(X \geq 3) = P(X = 3) + \dots + P(X = 12)$$

$$\text{Einfacher: } P(X \geq 3) = 1 - P(X = 2) = 1 - \frac{1}{36} = \frac{35}{36}$$

Wie gross ist die Wahrscheinlichkeit, eine Augensumme von 3 bis 5 zu würfeln?

$$\text{Gesucht: } P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5) = \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{9}{36} = \frac{1}{4}$$

Es können nun beliebige (diskrete) Verteilungen durch 2 Kennzahlen zusammengefasst werden:

- Erwartungswert $E(X)$: mittlere Lage der Verteilung
- Standardabweichung: $\sigma(X)$: Streuung der Verteilung

Definition

Diskrete Zufallsvariablen X : Mögliche Werte x_1, x_2, \dots, x_n

Erwartungswert

$$E(X) = x_1 * P(X = x_1) + x_2 * P(X = x_2) + \dots + x_n * P(X = x_n)$$

$$= \sum_{\text{alle möglichen } x} xP(X = x)$$

Varianz und Standardabweichung

$$Var(X) = (x_1 - E(X))^2 * P(X = x_1) + \dots + (x_n - E(X))^2 * P(X = x_n)$$

$$= \sum_{\text{aller möglichen } x} (x - E(X))^2 * P(X = x)$$

$$\sigma(X) = \sqrt{Var(X)}$$

Beispiel Erwartungswert und Standardabweichung

Wurf eines fairen Würfels: Alle 6 möglichen Zahlen gleiche W'keit

Zufallsvariable X sei die geworfene Zahl

- $E(X) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \dots + x_6 \cdot P(X = x_6)$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

Dieser Erwartungswert 3.5 = der Durchschnitt der Augenzahlen

Wie lässt sich dieser Wert nun interpretieren?

- 100 mal würfeln: Durchschnitt meistens *nicht* exakt 3.5, aber in der *Nähe*
- 100 Milliarden mal würfeln: Durchschnitt nicht *exakt* 3.5, aber nahe dran

Interpretation:

Für sehr viele Würfe liegt Durchschnitt sehr nahe beim Erwartungswert

Standardabweichung mir R berechnen:

```
x <- 1:6
p <- 1/6
E_X <- sum(x * p)
var_X <- sum((x - E_X)^2 * p) sd_X <- sqrt(var_X)
sd_X
## [1] 1.707825
```

Das heisst die «durchschnittliche» Abweichung liegt bei 1.7 von 3.5.

Beispiel: Spielkarten

- Verteilung:

x	0	2	3	4	10	11
$P(X = x)$	4/9	1/9	1/9	1/9	1/9	1/9

- Ziehen aus dem Stapel eine Karte
- Welches ist der durchschnittliche Wert der Karte, die gezogen wird?
- Berechnen Erwartungswert $E(X)$:

$$E(X) = 0 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 10 \cdot \frac{1}{9} + 11 \cdot \frac{1}{9} = 3.33$$

- Zahlen untereinander in Tabelle werden multipliziert und dann addiert

```
x <- c(0, 2, 3, 4, 10, 11)
p <- 1 / 9 * c(4, 1, 1, 1, 1, 1) E_X <- sum(x * p)
E_X
## [1] 3.333333
```

Dies ist der durchschnittliche Wert, der zu erwarten ist, wenn die Karte sehr oft gezogen und wieder in den Stapel zurückgelegt wird. Viele Karten mit Wert 0, dann ist der Erwartungswert eher tief.

- Varianz und die Standardabweichung:

$$\begin{aligned} \text{Var}(X) &= (0 - 3.33)^2 \cdot \frac{4}{9} + (2 - 3.33)^2 \cdot \frac{1}{9} + (3 - 3.33)^2 \cdot \frac{1}{9} \\ &\quad + (4 - 3.33)^2 \cdot \frac{1}{9} + (10 - 3.33)^2 \cdot \frac{1}{9} + (11 - 3.33)^2 \cdot \frac{1}{9} \\ &= 16.67 \end{aligned}$$

und

$$\sigma(X) = \sqrt{16.67} = 4.08$$

- R:

```
var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X
## [1] 4.082483
```

- „Mittlere“ Abweichung 4.1: Eher gross wegen Werten 10, 11

Der Erwartungswert einer diskreten Zufallsvariable ist ein gewichtetes arithmetisches Mittel von allen möglichen Werten, wobei die Werte mit ihrer Wahrscheinlichkeit gewichtet werden.

Der Erwartungswert wird oft auch mit μ_X bezeichnet. Index X wird oft weggelassen, falls die Zufallsvariable klar ist.

Wenn die Wahrscheinlichkeiten für alle Werte x_1, x_2, \dots, x_n gleich sind, dann ist der Erwartungswert das arithmetische Mittel der Werte.

Die Varianz (das Quadrat der Abweichung) eines Wertes der Zufallsvariable vom Erwartungswert mit der Wahrscheinlichkeit des Wertes ist gerichtet.

Die Standardabweichung hat dieselbe Einheit wie X: Die Einheit der Varianz deren Quadrat ist:

- Z.B. X in Metern (m) gemessen $\rightarrow \text{Var}(X)$ in Quadratmeter (m^2)
- $\sigma(X)$ wiederum die Dimension Meter (m)

Unterschied empirischer und theoretischer Kennzahlen

- Gesehen: Mittelwert:

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

- Empirische Varianz:

$$\text{Var}(x) = \frac{(x_1 - \bar{x}_n)^2 + (x_2 - \bar{x}_n)^2 + \dots + (x_n - \bar{x}_n)^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

- Empirische Standardabweichung s_x :

$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Wie hängen diese Definitionen und die Definition für die Standardabweichung für eine Zufallsvariable X zusammen? Dies ist sehr genau unterscheiden:

Das Arithmetische Mittel \bar{x} wird aus konkreten Daten berechnet, aus Messdaten x_1, \dots, x_n und wird nach der Formel \bar{x}_n berechnet.

Der Erwartungswert $E(X)$ ist ein theoretischer Wert, der sich aus dem Modell der Wahrscheinlichkeitsverteilung ergibt.

Die Hoffnung, dass sich das Arithmetische Mittel \bar{x} nähert für immer mehr Versuche den theoretischen Wert $\mu_X = E(x)$ immer besser an, sofern die Daten der Wahrscheinlichkeitsverteilung von X folgen. Ist dies nicht der Fall, so stimmt etwas am Modell nicht (z.B. alle Seiten beim Würfel haben gleiche Wahrscheinlichkeit geworfen zu werden).

Wie hängen diese Definitionen und die Kennzahlen für die Verteilung einer Zufallsvariablen X zusammen?

Die Empirische Standardabweichung s_x wird aus konkreten Daten berechnet. Aus Messwerten x_1, x_2, \dots, x_n nach der Formel s_x berechnet.

Die Standardabweichung σ_x ist ein theoretischer Wert, der sich aus dem Modell der Wahrscheinlichkeitsverteilung ergibt.

Die Hoffnung, dass ich die empirische Standardabweichung s_x sich für immer mehr Versuche den theoretischem Wert σ_x immer besser annähert, falls die Daten der Wahrscheinlichkeitsverteilung von X folgen. Ist dies nicht der Fall, so stimmt etwas am Modell nicht (z.B. alle Seiten beim Würfel haben die gleiche Wahrscheinlichkeit geworfen zu werden).

Beispiel: Faire Würfel

- Jede Seite die gleiche W'keit geworfen zu werden
- Gerechtfertigte Annahme aus *Symmetriegründen*: Alle Seiten gleich
- Erwartungswert:

$$E(X) = 3.5$$

- Werfen idealen, fairen Würfel $n = 10$ mal
- Obwohl der Würfel fair ist, wird der Durchschnitt nie genau 3.5 sein
- Idealer fairen Würfel: R:

```
x <- sample(1:6, size = 10, replace = T)
x
## [1] 3 3 3 4 3 6 5 2 3 6
mean(x)
## [1] 3.8
```

- Durchschnitt 3.8: Einiges neben den zu *erwartenden* 3.5
- Würfel nochmals 10mal werfen: Normalerweise ein anderes Resultat
- Simulation mit R
- Simulieren 10mal 10 Würfe: Berechnen entsprechende Durchschnitte

```
for (i in 1:10)
{
  x <- sample(1:6, size = 10, replace = T)
  cat(mean(x), " ")
}
```

```
## 3.3 3.5 4.2 3.4 2.9 3 3.9 2.9 4.1 2.7
```

- Durchschnitte gehen von 2.7 bis 4.2
- Zwar in der „Nähe“ von 3.5 aber nicht sehr genau

- Würfeln $n = 100$ Würfel:

```
x <- sample(1:6, size = 100, replace = T)
x
## [1] 5 6 6 1 5 1 4 5 1 2 3 1 3 6 2 3 1 6 1 4 3 6 1 6 5
## [26] 6 6 3 1 5 5 6 6 2 2 3 4 3 1 1 5 1 2 4 5 6 5 4 2 5
## [51] 6 5 2 6 4 4 4 4 1 2 2 6 6 3 5 3 6 5 5 1 5 6 1 2 1
## [76] 5 4 1 6 1 5 3 1 2 6 5 3 1 4 1 2 1 4 4 1 4 6 1 5 6
mean(x)
## [1] 3.57
```

- Mittelwert 3.57: Schon relativ nahe beim theoretischen Wert von 3.5

- Machen dies 10mal:

```
## 3.39 3.43 3.55 3.5 3.48 3.61 3.46 3.64 3.28 3.41
```

- Durchschnitte: Zwischen 3.28 und 3.64 ($\approx \pm 0.2$ vom Erwartungswert)

- Dasselbe für $n = 1000$ Würfe (auf drei Nachkommastellen):

```
## 3.475 3.49 3.435 3.437 3.407 3.479 3.567 3.474 3.498 3.565
```

- Durchschnitte: Zwischen 3.407 und 3.565 ($\approx \pm 0.1$ vom EW)

- Für $n = 1\,000\,000$: Durchschnitte auf 3 Nachkommastellen gerundet:

```
## 3.498 3.497 3.501 3.496 3.501 3.5 3.497 3.5 3.503 3.499
```

- Durchschnitte: Zwischen 3.498 und 3.503 ($\approx \pm 0.005$ vom EW)

- Durchschnitt: Für immer grössere n immer näher bei 3.5

- Annahme in Beispiel: Fairer Würfel gibt

- Dies ist *nicht* realistisch

- *Kein* realer Würfel ist wirklich symmetrisch, das heisst nicht alle Wurfzahlen sind gleich wahrscheinlich

- Können Würfel zu konstruieren, der *sehr* fair ist und alle Wurfzahlen mit einer W'keit von *fast* einem Sechstel vorkommen

- Aber *exakt* geht das nicht

Bedingte Wahrscheinlichkeit

Beispiel für bedingte Wahrscheinlichkeit

- Betrachten Gruppe von 20 Personen:
 - Einige sind Raucher, die anderen Nichtraucher
 - Einige sind Frauen, der Resten Männer

- Bezeichnungen:

F : Frau, M : Mann, R : Raucher, \bar{R} : Nichtraucher

- Tabelle:

	M	F	
R	3	1	4
\bar{R}	9	7	16
	12	8	20

- Es hat:

- 4 Raucher und 16 Nichtraucher
- 8 Frauen und 12 Männer

- Wert 3 links oben: Anzahl Personen, die Männer sind *und* rauchen

- Schreibweise:

$$|R \cap M| = 3$$

- Für W'keiten: Dividieren alle Werte in Tabelle durch 20

- Tabelle mit W'keiten:

	M	F	
R	0.15	0.05	0.2
\bar{R}	0.45	0.35	0.8
	0.6	0.4	1

- Wert 0.15 links oben: W'keit, dass eine zufällig ausgewählte Person ein Mann ist und raucht

- Berechnung:

$$P(R \cap M) = \frac{|R \cap M|}{|\Omega|} = \frac{3}{20} = 0.15$$

- Wert 0.2 links aussen: W'keit, dass eine zufällig ausgewählte Person ein Raucher ist

- Also:

$$P(R) = \frac{|R|}{|\Omega|} = 0.2$$

- Betrachten nur ein Teil der Tabelle → Raucher

	M	F	
R	0.15	0.05	0.2
R	0.45	0.35	0.8
	0.6	0.4	1

- Können nach W'keit fragen, dass eine zufällig ausgewählte Person *unter den Rauchern* ein Mann ist
- Aus 1. Tabelle (absolute Zahlen) ist diese W'keit:

$$\frac{|R \cap M|}{|R|} = \frac{3}{4} = 0.75$$

- Mit 2. Tabelle (W'keiten):

$$\frac{P(R \cap M)}{P(R)} = \frac{0.15}{0.20} = 0.75$$

- Das heisst, dass 75 % der Raucher sind Männer
- Diese W'keit heisst *bedingte Wahrscheinlichkeit*
- Bezeichnung:

$$P(M | R)$$

- Begriff „bedingt“: Nicht die gesamte Grundmenge betrachten, sondern nur ein Teil davon
- Neue Grundmenge hier: Raucher R
- Dies ist in $P(M | R)$ die Grösse nach dem Längsstrich

- Es gilt dann:

$$P(M | R) = \frac{P(R \cap M)}{P(R)} \quad (*)$$

- Formel wird als Definition der bedingte W'keit verwendet
- Berechnen bedingte W'keit:

$$P(R | M)$$

- W'keit, dass ein zufällig ausgewählter Mann ein Raucher ist
- Tabelle: Nur Männer werden berücksichtigt werden

	M	F	
R	0.15	0.05	0.2
\bar{R}	0.45	0.35	0.8
	0.6	0.4	1

- Berechnung dieser W'keit: Vertauschen in Gleichung (*) die Variable M durch R
- Resultat:

$$P(R | M) = \frac{P(M \cap R)}{P(M)} = \frac{P(R \cap M)}{P(M)} = \frac{0.15}{0.6} = 0.25$$

Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit ist die Wahrscheinlichkeit, dass das Ereignis A eintritt, wenn man schon weiss, dass B eingetreten ist.

Bezeichnung $P(A | B)$

Längsstrich wird als «unter der Bedingung» gelesen.

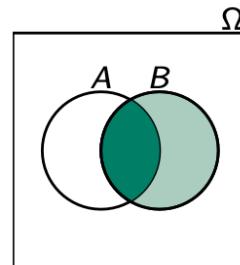
Die bedingte Wahrscheinlichkeit wird definiert durch:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Die Interpretation: $P(A | B)$ ist die Wahrscheinlichkeit für das Ereignis A, wenn man weiss, dass das Ereignis B schon eingetroffen ist.

Verdeutlichung der Formel mit Flächen

- Graphisch:

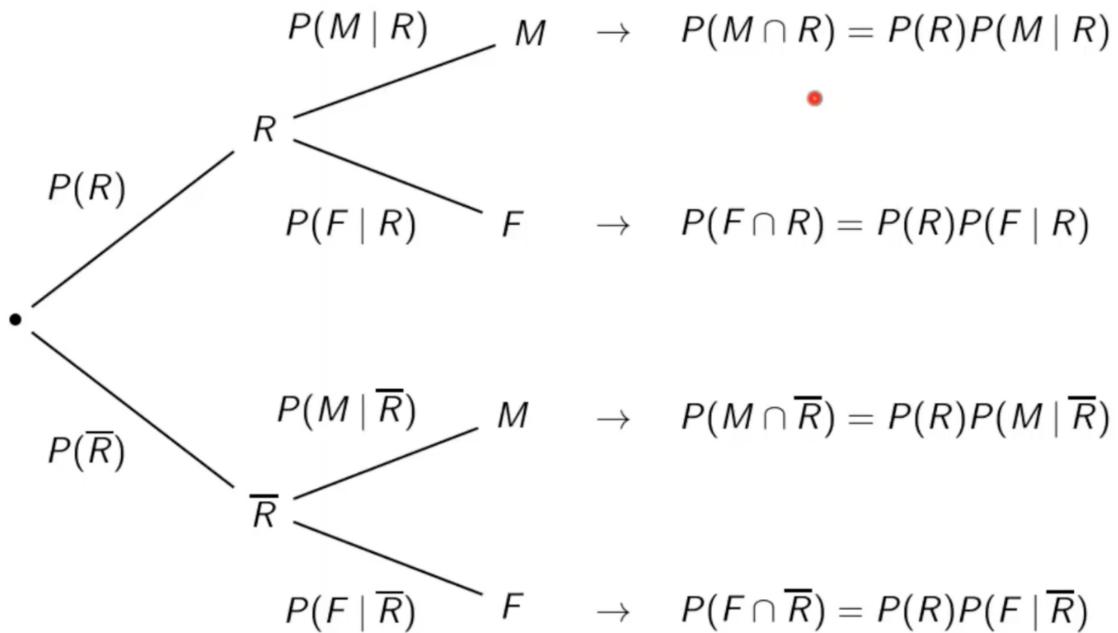
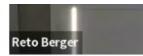


- Es ist $|\Omega| = 1$
- $P(A | B)$ Flächeninhalt der dunkel gefärbten Flächen
- $P(B)$ Flächeninhalt der gesamten gefärbten Fläche B
- Anteil der dunkelgefärbten Fläche zur gefärbten Fläche ist:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

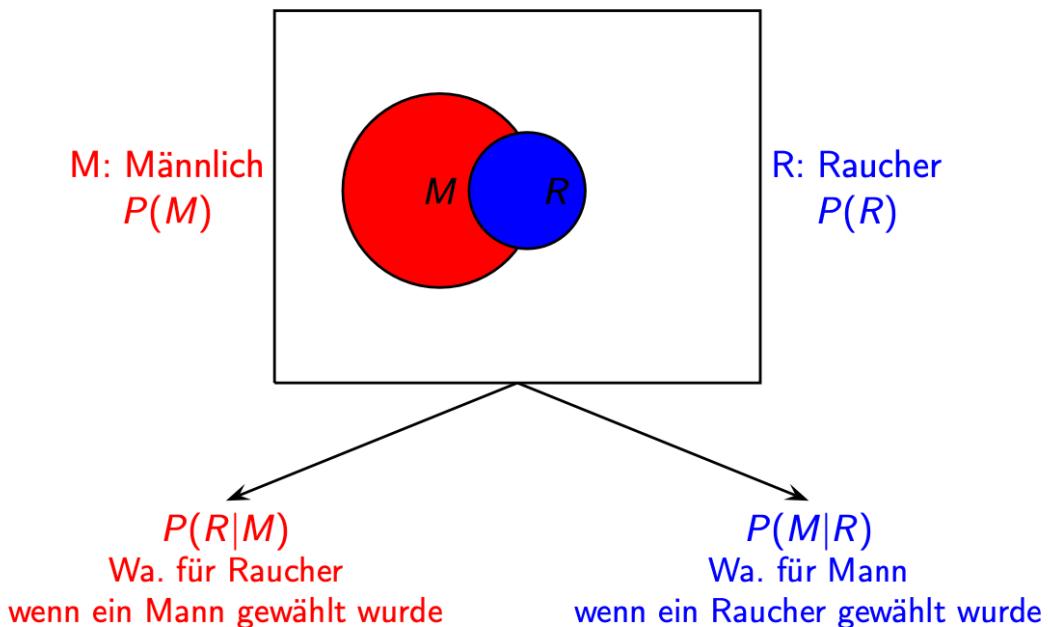
Eine andere Möglichkeit, die bedingte Wahrscheinlichkeit darzustellen:

Baumdiagramm



Bedingte Wahrscheinlichkeit

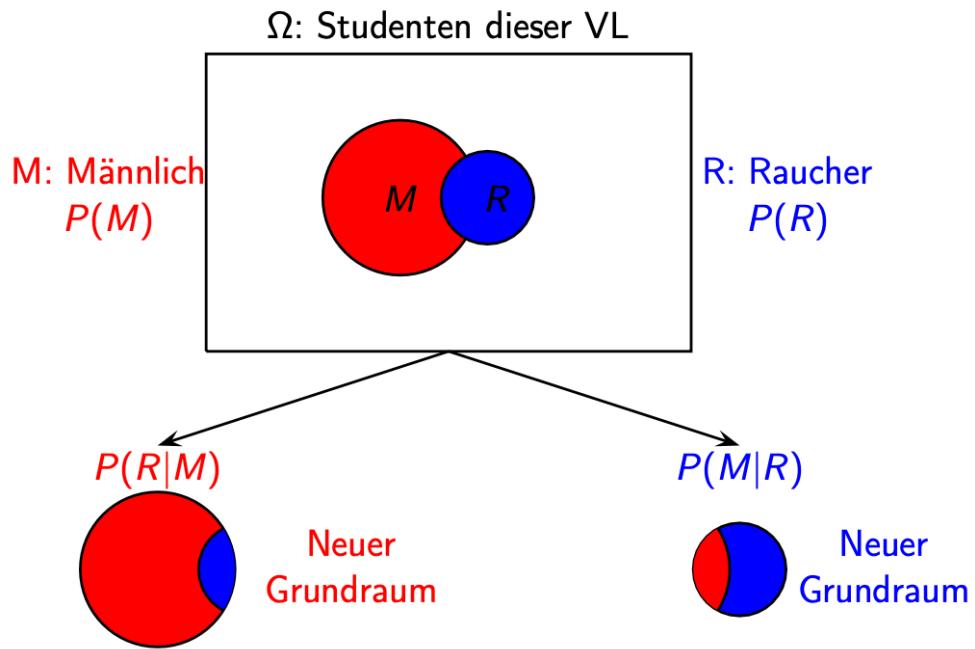
Ω : Studenten dieser VL



Welche Aussagen sind korrekt?

- 1. $P(M|R) = P(R|M)$
- 2. $P(M|R) > P(R|M)$
- 3. $P(M|R) < P(R|M)$

Korrekt ist: Nr. 2



Beispiel: Medizinischer Test

- Medizinischer Test soll für eine Krankheit feststellen, ob eine Person an dieser Krankheit erkrankt ist oder nicht
- Natürlich ist dieser Test nicht ganz genau:
 - ▶ Zeigt manchmal die Krankheit an, obwohl die Person gesund ist
 - ▶ Er zeigt die Krankheit nicht an, obwohl die Person krank ist
- Fragestellung:
 - ▶ Sie gehen zum Arzt und machen diesen Test auf eine tödliche Krankheit
 - ▶ Test ist positiv, d.h. Sie haben gemäss dem Test die Krankheit, müssen aber nicht unbedingt krank sein
 - ▶ Wie gross ist die W'keit, dass Sie wirklich krank sind?

- Bezeichnungen:

- ▶ D : Krankheit ist vorhanden; \bar{D} : Krankheit ist nicht vorhanden
- ▶ $+$: Test zeigt Krankheit an; $-$: Test zeigt Krankheit nicht an

- W'keiten in Tabelle sind durch Versuche bekannt

	D	\bar{D}
$+$	0.009	0.099
$-$	0.001	0.891

- Z.B.: W'keit, dass Krankheit vorhanden *und* Test positiv ausfällt

$$P(D \cap +) = 0.009$$

- Diese W'keit ist recht klein

- Grund: Nur kleiner Prozentsatz der Bevölkerung hat Krankheit

- Verschiedene bedingte W'keiten:

- ▶ $P(+|D)$: W'keit, dass ein Kranker auch wirklich positiv getestet wird
- ▶ $P(-|\bar{D})$: W'keit, dass ein Gesunder richtigerweise negativ getestet wird
- ▶ $P(D|+)$: W'keit, dass positiv Getesteter auch wirklich krank ist
- ▶ etc.

- Berechnen zuerst die W'keit $P(+|D)$:

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

- Bezeichnungen:

- ▶ D : Krankheit ist vorhanden; \bar{D} : Krankheit ist nicht vorhanden
- ▶ $+$: Test zeigt Krankheit an; $-$: Test zeigt Krankheit nicht an

- W'keiten in Tabelle sind durch Versuche bekannt

	D	\bar{D}
$+$	0.009	0.099
$-$	0.001	0.891

- Z.B.: W'keit, dass Krankheit vorhanden *und* Test positiv ausfällt

$$P(D \cap +) = 0.009$$

- Diese W'keit ist recht klein

- Grund: Nur kleiner Prozentsatz der Bevölkerung hat Krankheit

- Verschiedene bedingte W'keiten:

- ▶ $P(+|D)$: W'keit, dass ein Kranker auch wirklich positiv getestet wird
 - ▶ $P(-|\bar{D})$: W'keit, dass ein Gesunder richtigerweise negativ getestet wird
 - ▶ $P(D|+)$: W'keit, dass positiv Getesteter auch wirklich krank ist
 - ▶ etc.
- Berechnen zuerst die W'keit $P(+|D)$:

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

- Für $P(D)$ wurde folgende Tatsache benutzt

$$P(D) = P(D \cap +) + P(D \cap -) = 0.009 + 0.001$$

- Summe der Einträge in der Tabelle in der Spalte unter D
- Die Kranken sind entweder positiv oder negativ getestet
- Bedingte W'keit $P(-|\bar{D})$:

$$P(-|\bar{D}) = \frac{P(- \cap \bar{D})}{P(\bar{D})} = \frac{0.891}{0.891 + 0.099} = 0.9$$

- Scheinbar ist dieser Test recht genau
- Kranke Personen werden zu 90 % als positiv eingestuft, und gesunde Personen werden zu 90 % als negativ eingestuft
- Fragestellung aber auch umkehren
- Angenommen, Sie gehen zu einem Test und dieser wird als positiv eingestuft
- Wie gross ist die W'keit, dass Sie die Krankheit wirklich haben?
- Die meisten Leute würden 0.9 antworten
- Müssen Sie sich also grosse Sorgen machen und das Testament schreiben oder einer Sterbehilfeorganisation beitreten?

- Die *richtige* Antwort ist die bedingte W'keit $P(D|+)$:

$$P(D|+) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.08$$

- Was bedeutet nun dieses Resultat?
- Die bedingte W'keit $P(D|+)$ ist die W'keit, dass man bei einem positiven Test auch wirklich krank ist
- Diese beträgt aber nur 8 %
- Bei positivem Test haben Sie also nur zu 8 % auch wirklich die Krankheit
- Ein positiver Test sagt hier also sehr wenig darüber aus, ob man die Krankheit hat oder nicht
- Die Frage ist nun, warum ist dies so
- Grund: Krankheit *selten*
- Numerisches Beispiel: Untersuchen 100 000 Personen
 - 1000 Personen haben die Krankheit (1 %)
 - 90 % dieser Personen werden positiv getestet: 900 Personen
 - 99 000 haben die Krankheit nicht
 - 10 % dieser Personen werden positiv getestet: 9900 Personen
 - Anzahl positiv Getesteter:
$$900 + 9900 = 10\ 800$$
 - Unter diesen positiv getesteten sind aber bei weitem mehr Gesunde, die fälschlicherweise positiv getestet wurden
 - W'keit, dass eine positiv getestete Person auch wirklich krank ist:

$$\frac{900}{10\ 800} = 0.0833$$

Bayes Theorem

Nützlicher Zusammenhang zwischen $P(A | B)$ und $P(B | A)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$$

Beispiel: Bayes Theorem liefert die gleiche Lösung wie vorher:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.9 \cdot (0.009 + 0.001)}{0.009 + 0.099} = \frac{0.009}{0.009 + 0.099} = 0.08$$

Herleitung

- Zweimalige Anwendung der Definition der bedingten W'keit

► Es gilt:

$$\text{P}(B | A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B \cap A) = P(B | A)P(A)$$

► und:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A | B)P(B)$$

- Da $A \cap B = B \cap A$, gilt:

$$P(A \cap B) = P(B \cap A)$$

- Und somit gilt auch

$$P(B | A)P(A) = P(A | B)P(B)$$

- Dividieren beide Seiten durch $P(B)$ und erhalten Bayes Theorem:

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Gesetz der totalen Wahrscheinlichkeit

Der Begriff der totalen Wahrscheinlichkeit ist ein weiter nützlicher Begriff. Die Menge A wird in Mengen A_1, \dots, A_k unterteilt, die miteinander keine Schnittmenge haben und zusammen (Vereinigung) die ganze Menge A bilden. Eine solche Aufteilung heisst Partitionierung. Für den Würfelwurf ist folgende Partitionierung möglich:

$$A_1 = \{1\}, \quad A_2 = \{2, 4\}, \quad A_3 = \{3, 5, 6\}$$

- Es gilt also:

$$A_1 \cap A_2 = \{\}; \quad A_1 \cap A_3 = \{\}; \quad A_2 \cap A_3 = \{\}$$

und

$$A_1 \cup A_2 \cup A_3 = A$$

Gesetz

Für die Partitionierung A_1, \dots, A_k und jedes beliebige Ereignis B gilt:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k) = \sum_{i=1}^k P(B|A_i)P(A_i)$$

- $k = 2$:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$$

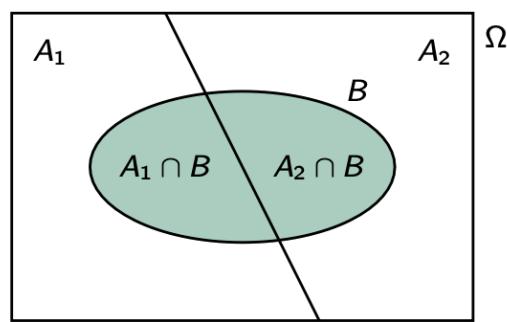
- $k = 3$:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

Herleitung

- Fall $k = 2$

- Graphische Darstellung:



- Mengen A_1 und A_2 bilden eine Partition von Ω

- Es gilt also:

$$A_1 \cup A_2 = \Omega \quad \text{und} \quad A_1 \cap A_2 = \{\}$$

- Menge B in zwei Teile aufteilt: $A_1 \cap B$ und $A_2 \cap B$

- Es gilt also:

$$B = (A_1 \cap B) \cup (A_2 \cap B)$$

- Für W'keit gilt:

$$P(B) = P((A_1 \cap B) \cup (A_2 \cap B))$$

- Es gilt:

$$(A_1 \cap B) \cap (A_2 \cap B) = \{\}$$

- Rechenregel:

$$\begin{aligned} P(B) &= P((A_1 \cap B) \cup (A_2 \cap B)) \\ &= P(A_1 \cap B) + P(A_2 \cap B) \end{aligned}$$

- Rechte Seite: Definition der bedingten W'keiten anwenden:

$$P(B | A_1) = \frac{P(A_1 \cap B)}{P(A_1)} \Rightarrow P(A_1 \cap B) = P(B | A_1)P(A_1)$$

- Entsprechende Formel gilt für $P(A_2 \cap B)$
- Oben einsetzen: Gesetz der totalen W'keit für $k = 2$:

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(B | A_1)P(A_1) + P(B | A_2)P(A_2) \end{aligned}$$

Beispiel: Spam-Mail

- Teilen Emails in drei Kategorien ein:

A_1 : „spam“, A_2 : „niedrige Priorität“, A_3 : „hohe Priorität“

- Aus früheren Beobachtungen bekannt:

$$P(A_1) = 0.7, \quad P(A_2) = 0.2, \quad \text{und} \quad P(A_3) = 0.1$$

- Es gilt

$$P(A_1) + P(A_2) + P(A_3) = 1$$

wie es bei einer Partitionierung auch sein sollte

- Ereignis B : Wort „free“ taucht in der Email auf

- Dieses Wort kommt sehr oft in Spam-Mails vor, aber nicht nur

- Von früheren Beobachtungen bekannt:

$$P(B|A_1) = 0.9, \quad P(B|A_2) = 0.01, \quad \text{und} \quad P(B|A_3) = 0.01$$

- Hier ergibt die Summe nicht 1

- Dies sind die W'keiten, mit der das Wort „free“ in den drei Mailkategorien vorkommt

- Angenommen, es kommt eine Email an, die das Wort „free“ enthält

- Wie gross ist die W'keit, dass es sich um Spam handelt?

- Lösung mit Bayes Theorem und Gesetz der totalen W'keit:

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{0.9 \cdot 0.7}{(0.9 \cdot 0.7) + (0.01 \cdot 0.2) + (0.01 \cdot 0.1)} \\ &= 0.995 \end{aligned}$$

- Viele Spamfilter basieren tatsächlich auf diesem Prinzip
- Mails werden nach Wörtern wie „free“, „credit“, etc. durchsucht, die häufig in Spam-Mails vorkommen, in anderen aber eher nicht

Beispiel

- Einige Kinder werden mit Down-Syndrom geboren
- Es gibt Tests, die schwangeren Frauen machen können, ob ihr Baby an dieser Krankheit leiden könnte
- Untersuchung (Universität Liverpool), wie gut die Testergebnisse von den Beteiligten, d.h. den schwangeren Frauen, ihren Lebenspartnern, Hebammen und Gynäkologen, interpretiert werden
- 85 Personen wurde folgendes Szenario gezeigt:
Der Serumtest untersucht schwangere Frauen auf Babys mit Down-Syndrom. Der Test ist ein sehr guter, aber nicht perfekter Test. Ungefähr 1% der Babys haben das Down-Syndrom. Wenn das Baby das Down-Syndrom hat, besteht eine 90-prozentige Wahrscheinlichkeit, dass das Ergebnis positiv ausfällt. Wenn das Baby nicht betroffen ist, besteht immer noch eine 1% Chance, dass das Ergebnis positiv ist. Eine schwangere Frau wurde getestet und das Ergebnis ist positiv. Wie gross ist die Wahrscheinlichkeit, dass ihr Baby tatsächlich das Down-Syndrom hat?
- Resultat, wie gut die 85 Personen abgeschnitten haben:

	richtig	zu hoch	zu niedrig	
Schwangere Frauen	1	15	6	22
Lebenspartner	3	10	7	20
Hebammen	0	10	12	22
Gynäkologen	1	16	4	21
	5	51	29	85

- Nur fünf der 85 gaben die richtige Antwort
- Die Angehörigen der Gesundheitsberufe waren nicht besser als die schwangeren Frauen und ihrer Lebenspartner
- Besonders bemerkenswert: Nur einer von 21 Gynäkologen erhielt die richtige Antwort

- Anderen Gruppe von 81 Personen: Alternatives Szenario gezeigt:
Der Serumtest untersucht schwangere Frauen auf Babys mit Down-Syndrom. Der Test ist ein sehr guter, aber nicht perfekter Test. Etwa 100 von 10 000 Babys haben das Down-Syndrom. Von diesen 100 Babys mit Down-Syndrom werden 90 ein positives Testergebnis haben. Von den verbleibenden 9900 nicht betroffenen Babys werden immer noch 99 ein positives Testergebnis haben. Wie viele schwangere Frauen, die ein positives Testergebnis haben, bekommen tatsächlich ein Baby mit Down-Syndrom?

- Ergebnis:

	richtig	zu hoch	zu niedrig	
Schwangere Frauen	3	3	10	21
Lebenspartner	3	8	9	20
Hebammen	0	7	13	20
Gynäkologen	13	3	4	20
	19	26	36	81

- Eindeutig eine Verbesserung
- Neuformulierung des Szenarios: Absolute Zahlen anstatt Prozenten verwendet
- Macht es zu einem leichteren Problem
- Muss nur die beiden Zahlen 90 und 99 herauslesen:

$$\frac{90}{90 + 99} \approx 48\%$$

- Aber: Immer noch erhielt nur etwa ein Viertel die richtige Antwort
- Immerhin schnitten die Gynäkologen deutlich besser ab

Repetition: Diskrete Wahrscheinlichkeitsverteilung

- Zufallsvariable X : Ordnet jedem Zufallsexperiment genau eine Zahl zu
- Wir können somit X auch als *Funktion* auffassen
- Beispiel: Zufallsvariable X ordnet einer zufällig ausgewählten, in der Schweiz lebenden Person, die Körpergrösse in cm zu
- Hier: Körpergrösse wird auf Zentimeter gerundet
- *Definitionsmenge* dieser Zufallsvariable X : Menge aller in der Schweiz lebenden Personen
- Zufallsvariable X kann nur folgende Werte annehmen (*Wertemenge*)

$$W_X = \{0, 1, 2, \dots, 500\}$$

- Wertebereich absichtlich zu gross gewählt, damit auch sicher alle vorkommenden Werte dabei sind
- Wertemenge besteht also nur aus endlich vielen ganzen Zahlen
- Eine solche Menge heisst *diskret*
- Wichtig: Wir können *keinen* Wert zwischen zwei Werten der Wertemenge auswählen können
- Die Menge ist „löchrig“
- Dies kann man als saloppe Definition von „diskret“ auffassen

- Zufallsvariable X : Misst die Körpergrösse einer zufällig ausgewählten Person
- Wählen nun zufällig (deshalb Zufallsvariable) eine Person aus
- Name der Person: *Tabea*
- Annahme: Jeder Name kommt nur genau einmal vor, was natürlich nicht der Fall ist
- Aber hätten auch die AHV-Nummer wählen können, die eindeutig ist.
- Tabea Einzigartig hat eine Körpergrösse 166 cm (auf cm gerundet)
- Formulierung mit Zufallsvariable:

$$X(\text{Tabea}) = 166$$

- Auswahl einer weiteren Person: *Tadeo* mit Körpergrösse von 176 cm

- Schreiben:

$$X(\text{Tadeo}) = 176$$

- Dies können wir mit jeder in der Schweiz lebenden Person machen
- Ausdruck

$$X = 174$$

beschreibt das *Ereignis* eine Person ausgesucht zu haben, die eine gerundete Körpergrösse von 174 cm hat

- Wir sprechen von einer *Realisierung* $x = 174$ von X
- *Unterschied*: Gross- und Kleinschreibung:
 - ▶ $x = 174$ ist eine Zahl
 - ▶ $X = 174$ ist eine Menge (Personen mit gerundeter Körpergrösse 174 cm)

- Diesem Ereignis kann man eine Wahrscheinlichkeit zuordnen:

$$P(X = 174)$$

- Berechnung: Anzahl Personen mit gerundeter Körpergrösse von 174 cm durch die Anzahl der in der Schweiz lebenden Personen dividieren
- Auf diese Weise können wir alle Wahrscheinlichkeiten

$$P(X = x)$$

berechnen, wobei x jeden Wert im Wertebereich annehmen kann.

- Insbesondere ist

$$P(X = 500) = 0$$

da es keine Person mit so einer Körpergrösse gibt.

- Deswegen spielt es auch keine Rolle, wenn Wertemenge viel zu gross gewählt wird
- Wir können weitere Wahrscheinlichkeiten bestimmen
- So ist

$$P(X \leq 170)$$

die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine gerundete Körpergrösse von 170 cm *oder weniger* hat

- Beachten Sie, dass die *nicht* der Wahrscheinlichkeit

$$P(X < 170)$$

entspricht

- Dies ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine gerundete Körpergrösse *kleiner als* 170 cm hat. Die Körpergrösse 170 cm gehört hier *nicht* dazu.

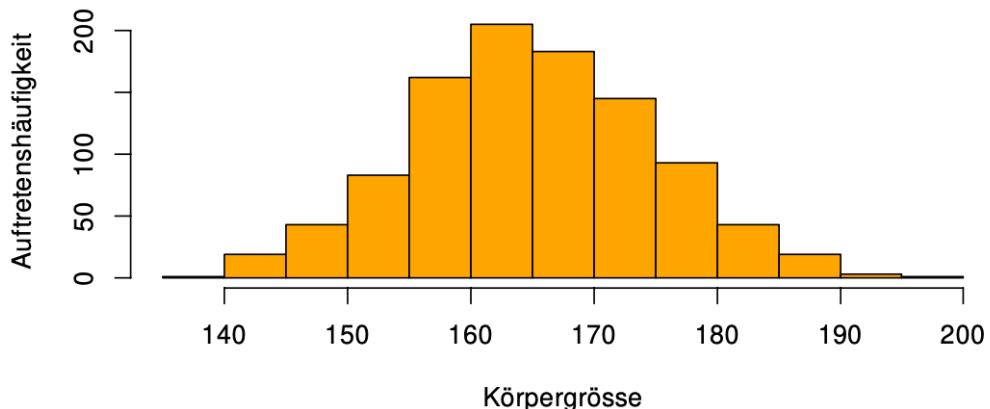
- Wichtig: Es gilt z. B.

$$P(X < 160) \leq P(X < 170)$$

- W'keit eine zufällig eine Person auszuwählen, die kleiner als 160 cm ist, ist kleiner gleich, als dass sie kleiner als 160 cm ist
- Wichtig: Alle W'keiten der Verteilung aufaddiert, ergibt 1:

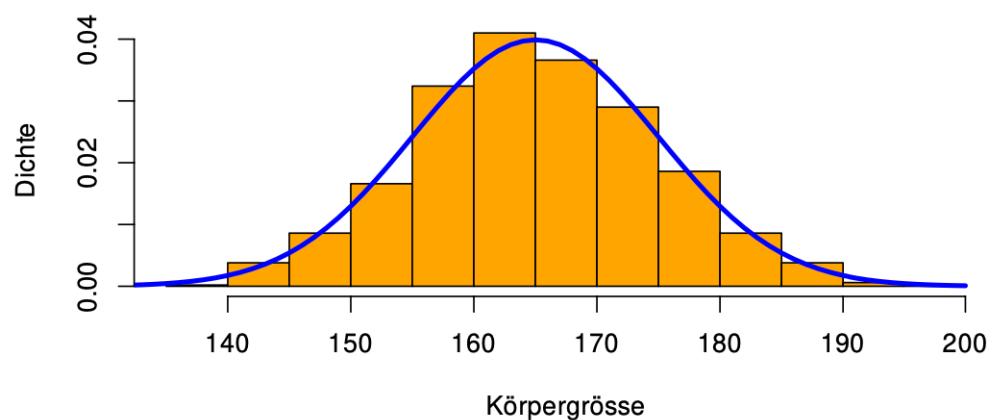
$$P(X = 0) + P(X = 1) + \dots + P(X = 499) + P(X = 500) = 1$$

- Wählen zufällig (siehe DoE) 1000 erwachsene Frauen aus
- Körpergrösse messen und ein Histogramm erstellen



- Form des Histogrammes sehr typisch → kommt recht häufig vor
- In Mitte Balken hoch
- Werden immer kleiner, je weiter sie von Mitte entfernt sind

- Versuchen Kurve einzuzeichnen, die Histogramm möglichst gut folgt

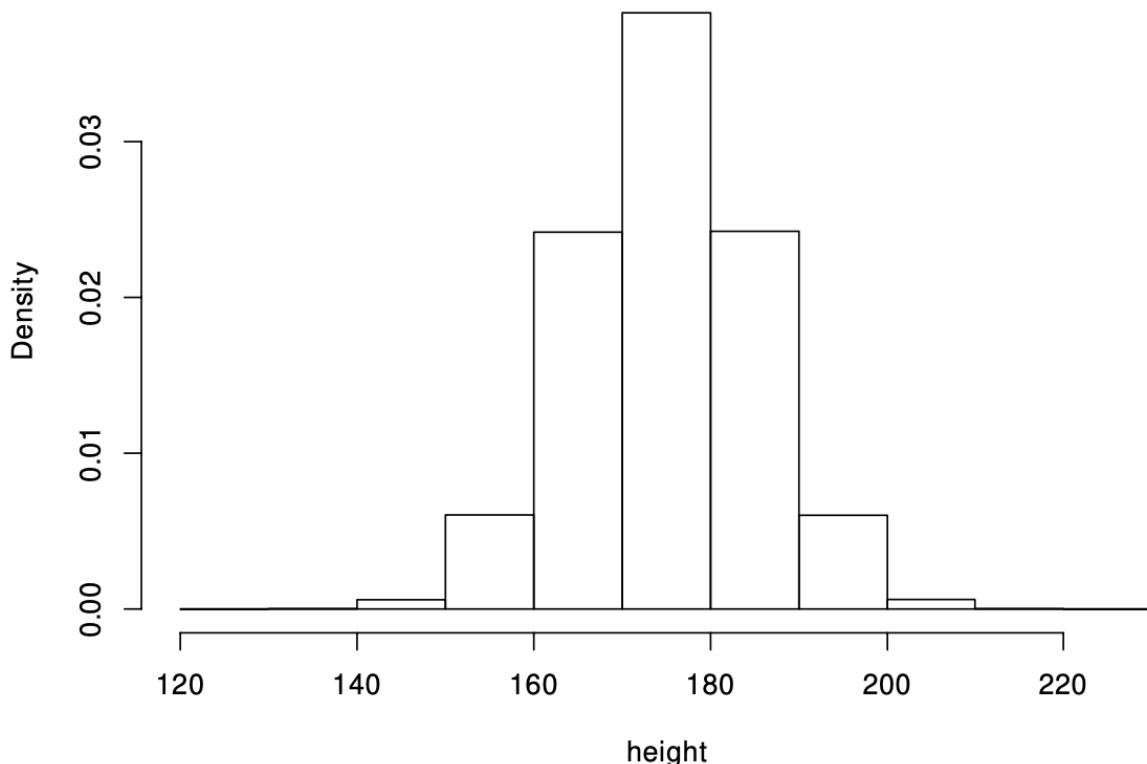


- Auf vertikaler Achse Dichten auftragen → Fläche von Histogramm 1
- Blaue Kurve heisst *Normalverteilungskurve*

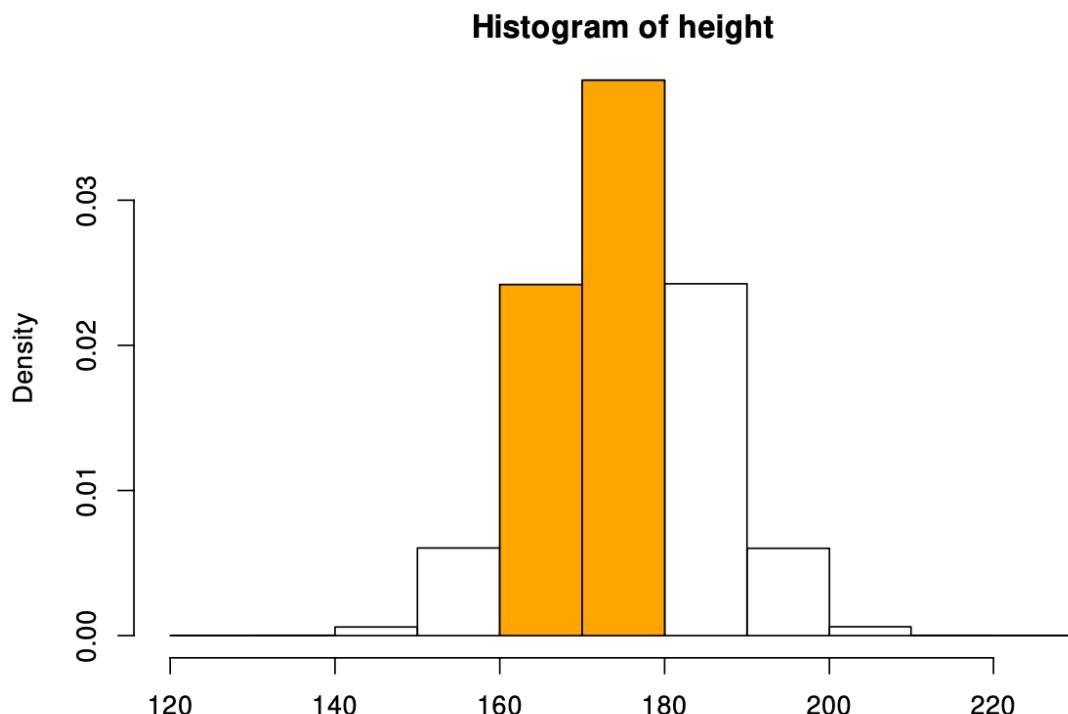
Von diskreter zu kontinuierlicher W'keitsverteilung

- Simulation der Körpergrösse (in cm) von einer Million Personen
- Zeichnen Histogramm dieser Grössen auf
- Annahme: Körpergrösse jeder Person so genau wie möglich bekannt
- Histogramm unten normalisiert: Summe der Fläche aller Balken ist 1
- Beginnen mit Balkenbreite von 10 cm

Histogram of height

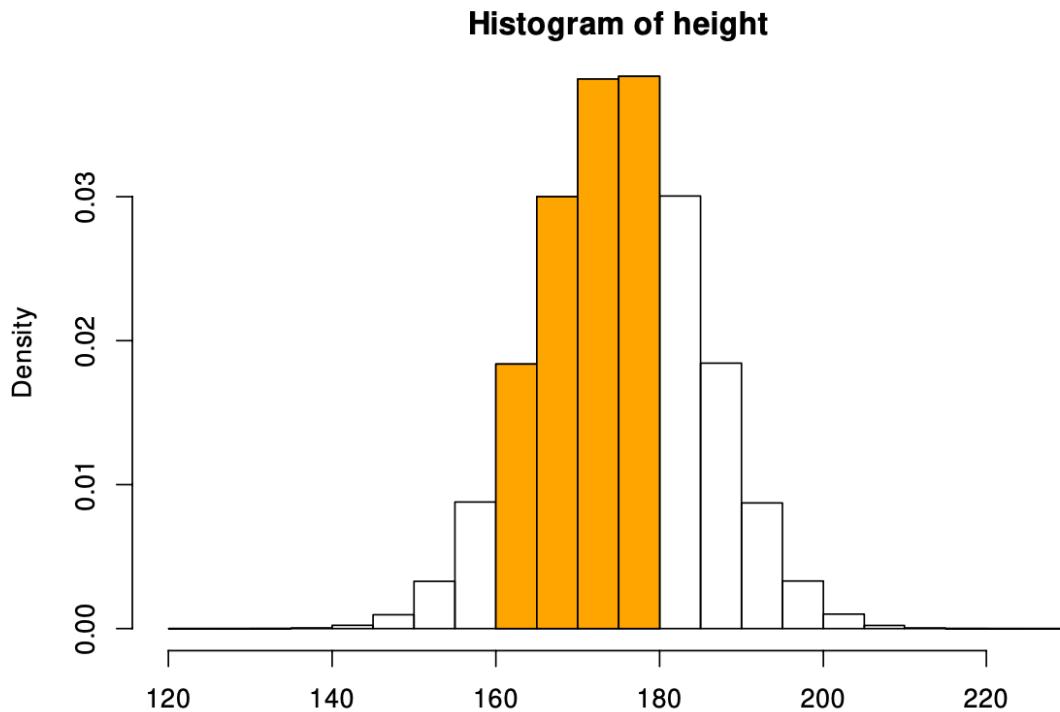


- Im folgendem Histogramm: Zwei Balken von 160 to 180 gefärbt
- Histogramm:

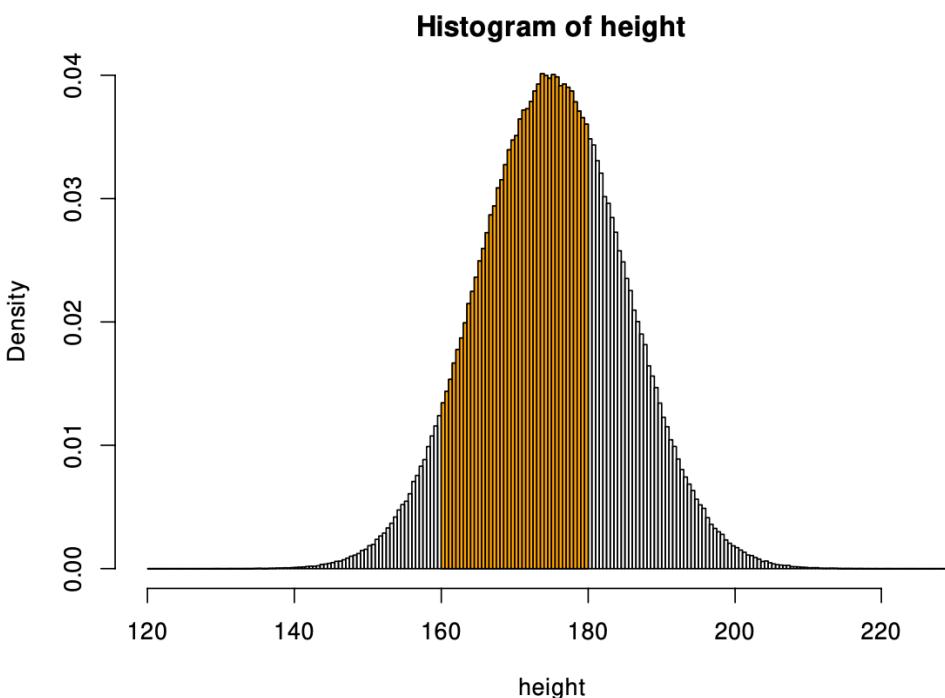


- Da Histogramm normalisiert: Bereich dieser beiden Balken als die W'keit interpretieren, dass eine zufällig ausgewählte Person dieser 1 000 000 Personen eine Grösse zwischen 160 cm und 180 cm
- Begründung: Grösse *jeder* dieser Personen ist im Histogramm enthalten
- W'keit, dass Grösse einer zufällig ausgewählten Person im Histogramm enthalten ist, beträgt 1
- Das ist Fläche der Summe der Flächen aller Balken → 1
- Fläche der beiden Balken als den Anteil aller Personen mit einer in diesen beiden Balken enthaltenen Höhe betrachten
- Dieser Anteil ist nichts anderes als die entsprechende W'keit

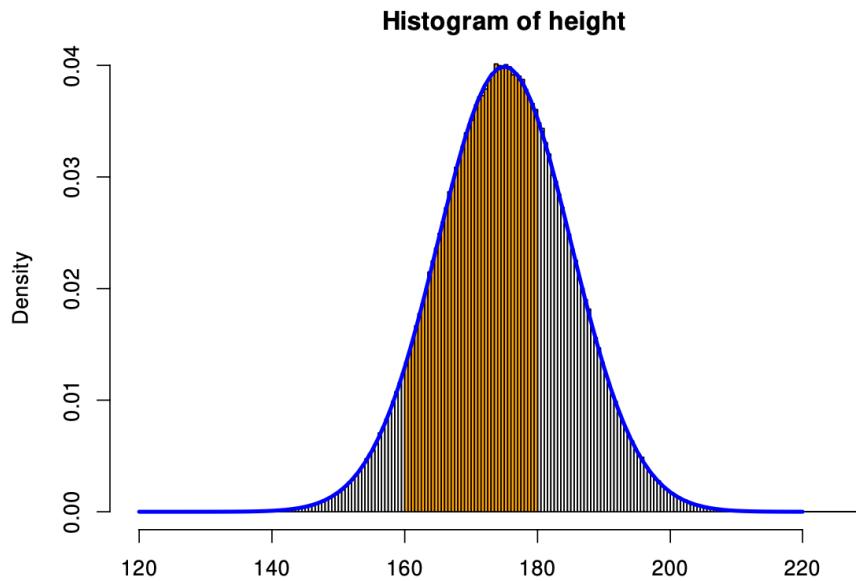
- Histogramm mit Balkenbreite 5 cm



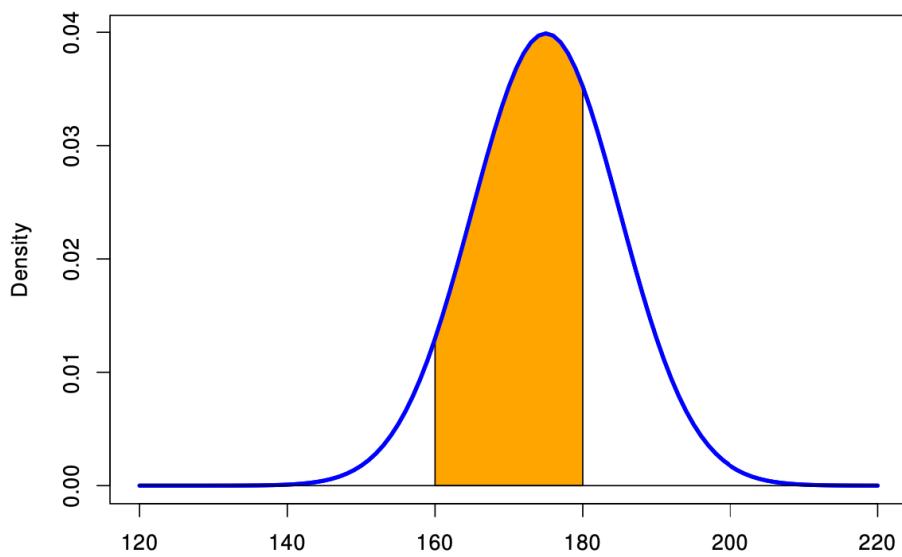
- Fläche der Summe aller Balken immer noch 1
- Interpretation des farbigen Bereichs gleich wie oben
- Beachte: Fläche der einzelnen Balken kleiner ist als die Fläche der einzelnen Balken im Histogramm vor
- Histogramm mit Balkenbreite 0.5 cm



- Kleine Balkenbreite: Histogramm folgt einer glatten Kurve
- Histogramm mit glatter Kurve



- Der letzte Schritt
- Balkenbreite gegen 0 streben (unendlich klein)



- Das „Histogramm“ folgt einer glatten Kurve
- Fläche unter dieser Kurve beträgt 1
- Farbige Fläche immer noch die W'keit, dass eine zufällig ausgewählte Person eine Körpergrösse zwischen 160cm und 180cm hat
- Fläche eines einzelnen „Balken“ ist 0
- Die blaue Kurve wird *Wahrscheinlichkeitsdichtefunktion* genannt

Normalverteilung

In vielen Anwendungen werden keine diskreten Daten, sondern Messdaten gemessen. Messdaten können jeden Wert in einem bestimmten Bereich annehmen. Bsp. Gemesene Körpergrösse in cm können jeden Wert im Intervall [0, 500] annehmen. Also z.B. auch 15.3456543, die Voraussetzung ist lediglich dass eine genaue Messung möglich ist.

Definitionen

Wertebereich W_x einer Zufallsvariable → Menge aller Werte, die X annehmen kann
Zufallsvariable X stetig: Wertebereich W_x kontinuierlich

Kontinuierliche Menge: Hier Ausschnitt aus der Zahlengeraden

Kontinuierlich: «Zusammenhängend» und nicht «löchrig», wie Menge {1,2,3}

Wichtige kontinuierliche Wertebereiche: $W_x = \mathbb{R}$, \mathbb{R}^+ oder $[0, 1]$

Letzter Fall: Zahlen 0 und 1 und alle Zahlen dazwischen

Intervalle

Intervall, wo die Grenzen innerhalb oder ausserhalb des Intervalls sein sollen → Eckige und runde Klammern.

- Runde Klammer: wert ausserhalb des Intervalls
- Eckige Klammer: Wert innerhalb des Intervalls

Intervall $(a, b]$: Alle Punkte x mit $x > a$ und $x \leq b$

Beispiel

Intervall $(1.2, 2.5]$ enthält die Zahl 1.2 nicht, die Zahl 2.5 schon

Unterschied zum Intervall $[1.2, 2.5]$: es enthält nur einen Punkt 1.2 der Zielgeraden mehr. Spielt in der Statistik keine Rolle, ob 1. Oder 2. Intervall verwendet wird.

Punktwahrscheinlichkeit 0

Die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariable: Punkt-Wahrscheinlichkeiten $P(X = x)$ für alle möglichen x im Wertebereich.

Vorher: X Körpergrösse auf cm gerundet → $P(X = 174)$ nicht 0

Stetige Zufallsvariable X : Für alle $x \in W_x$ gilt:

$$P(X = x) = 0$$

Folgerung: Wahrscheinlichkeitsverteilung von X kann nicht mittels der Punkt-Wahrscheinlichkeit beschrieben werden.

Beispiel: Körpergrösse

- Messen Körpergrösse von Personen
- W'keit genau eine Körpergrösse von 182.254 680 895 434 ... cm zu messen ist gleich 0:

$$P(X = 182.254 680 895 434 \dots) = 0$$

- Verwendung der W'keit einen exakten Messwert zu messen, bringt nichts für W'keitsverteilung der Körpergrösse
- Summe aller W'keiten müsste 1 ergeben → Ist nicht der Fall
- Aber möglich: W'keit, dass ein Messwert in einem bestimmten *Bereich* liegt
- Beispiel: zwischen 174 und 175 cm:

$$P(174 < X \leq 175)$$

- Diese W'keit ist dann nicht mehr 0
- Da

$$P(X = 174) = P(X = 175) = 0$$

gilt

$$P(174 < X \leq 175) = P(174 \leq X \leq 175) = P(174 < X < 175)$$

- Neuer Begriff: W'keitsdichte

Eigenschaften Wahrscheinlichkeitsdichte

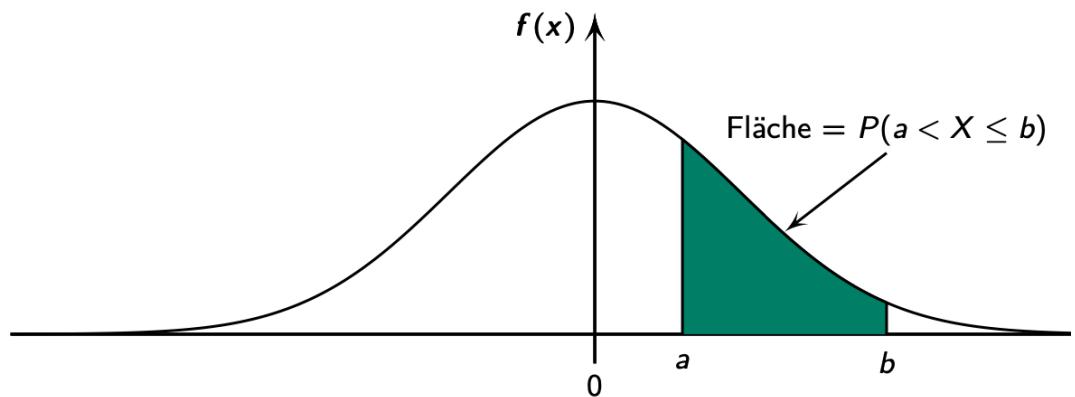
Für eine Wahrscheinlichkeitsdichte $f(x)$ gelten folgende Eigenschaften: Es gilt:
 $f(x) \geq 0$

Das heisst, die Kurve liegt oberhalb der x-Achse

Wahrscheinlichkeit $P(a < X \leq b)$ entspricht der Fläche zwischen a und b unter $f(x)$

Die gesamte Fläche unter der Kurve ist 1:

Dies ist die Wahrscheinlichkeit, dass irgendein Wert gemessen wird

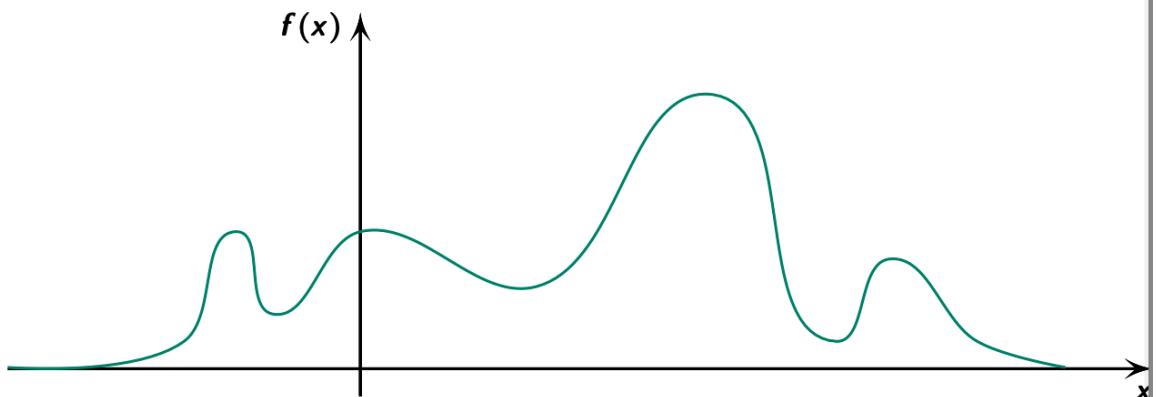


- Wichtig: Zusammenhang zwischen W'keit und Flächen:

Merkregel

Für stetige W'keitsverteilungen entsprechen W'keiten Flächen unter der Dichtefunktion.

- Skizze:

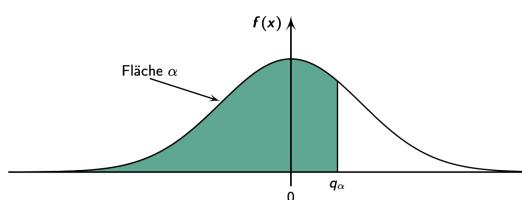


- W'keitsdichtefunktionen müssen keine „schöne“ Form haben
- Normalerweise aber „schöne“ Form vorhanden

Quantile

Stetige Verteilungen: α -Quantil q_α derjenige Wert, wo die Fläche (Wahrscheinlichkeit) unter der Dichtefunktion von $-\infty$ bis q_α gerade α entspricht.

50 %-Quantil: Median



- Messen wieder die Körpergrösse
- Beispiel: Für $\alpha = 0.75$ ist das zugehörige Quantil

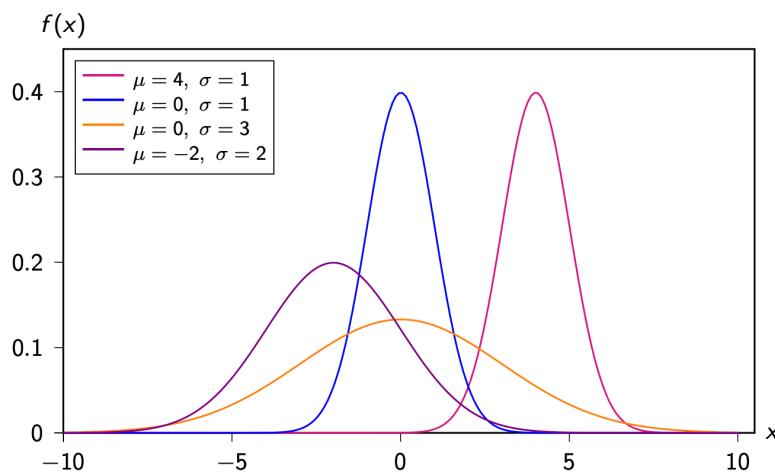
$$q_\alpha = 182.5$$

- D.h.: 75 % der gemessenen Personen kleiner oder gleich 182.5 cm

Normalverteilung (Gaussverteilung): $X \sim N(\mu, \sigma^2)$

- Definition muss man einmal gesehen haben
- Wertebereich $W = (-\infty, \infty)$
- Dichte
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$
- Erwartungswert $E[X] = \mu$
- Varianz $\text{Var}(X) = \sigma^2$

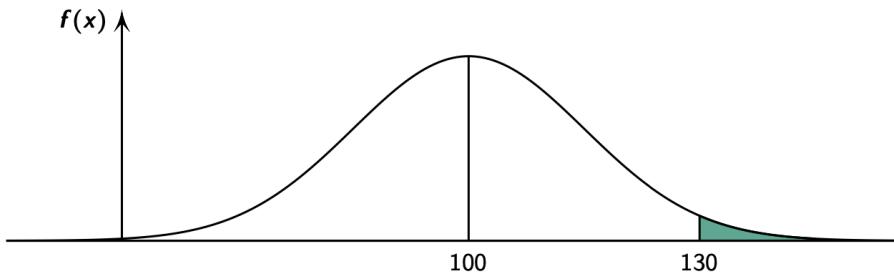
Normalverteilung: Illustration Dichten



- Dichtefunktionen „glockenförmig“
- Durch Parameter μ Verschiebung der Kurve:
 - ▶ Nach rechts, falls μ positiv
 - ▶ Nach links, falls μ negativ
- Durch Parameter σ wird die Kurve
 - ▶ schmal und hoch um μ , falls σ klein (nahe bei 0)
 - ▶ weit und tief um μ , falls σ gross

Beispiel mit R: Verteilung von IQ

- Anwendung: Häufigste Verteilung für Messwerte
- Beispiel: IQ Tests folgen einer Normalverteilung mit Mittelwert 100 und Standardabweichung 15
- X misst den IQ einer zufällig ausgewählten Person
- X normalverteilt mit $\mu = 100$ und $\sigma = 15$
- Notation:
$$X \sim \mathcal{N}(100, 15^2)$$
- Wie gross die W'keit ist, dass jemand einen IQ von mehr als 130 hat, also als hochbegabt gilt?
- $P(X > 130)$, wobei $X \sim \mathcal{N}(100, 15^2)$
- Skizze:



- Berechnung von $P(X > 130)$ mit R-Befehl `pnorm(...)`
- Dieser berechnet die W'keit:

$$P(X \leq 130)$$

- Beachte: Richtung des Ungleichheitszeichens!
- Berechnung:

```
pnorm(q = 130, mean = 100, sd = 15)
## [1] 0.9772499
```

- Befehl `pnorm(...)` berechnet Fläche (W'keit) von $-\infty$ bis $q = 130$ unter der Normalverteilungskurve mit $\mu = 100$ und $\sigma = 15$
- Dies ist aber *nicht* gesuchte W'keit $P(X > 130)$

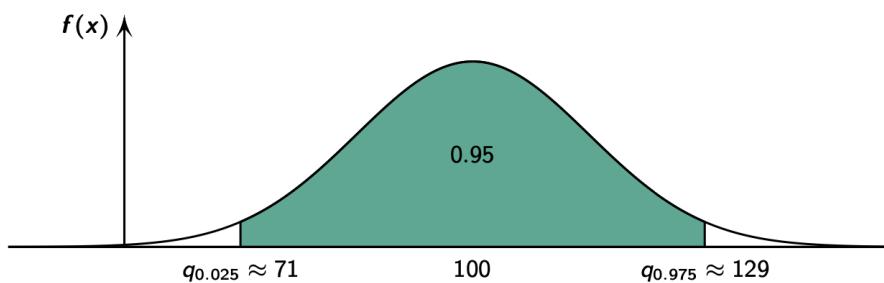
- Aber: Gesamtfläche unter der Kurve 1 → Gesuchte W'keit wie folgt schreiben:

$$P(X > 130) = 1 - P(X \leq 130)$$

- Berechnung mit R:

```
1 - pnorm(q = 130, mean = 100, sd = 15)
## [1] 0.02275013
```

- Also rund 2 % der Bevölkerung ist hochbegabt
- Welches Intervall enthält 95 % der IQ's um den Mittelwert $\mu = 100$?
- W'keit als Fläche:



- Grüne Fläche: 95 % der Gesamtfläche
- Kleine weissen Flächen links und rechts: Jeweils 0.025.
0.025 weil $1 - 0.95 / 2 \rightarrow$ also ist 0.0125 ein weisser Teil
Also ist das untere Quantil bei 0.025 und das obere Quantil bei 0.0975 $\rightarrow 0.025 + 0.95$

- W'keiten gegeben → Suchen die zugehörigen Werte
- Bestimmung der Quartile $q_{0.025}$ und $q_{0.975}$
- R:

```
qnorm(p = 0.025, mean = 100, sd = 15)
## [1] 70.60054
qnorm(p = 0.975, mean = 100, sd = 15)
## [1] 129.3995
```

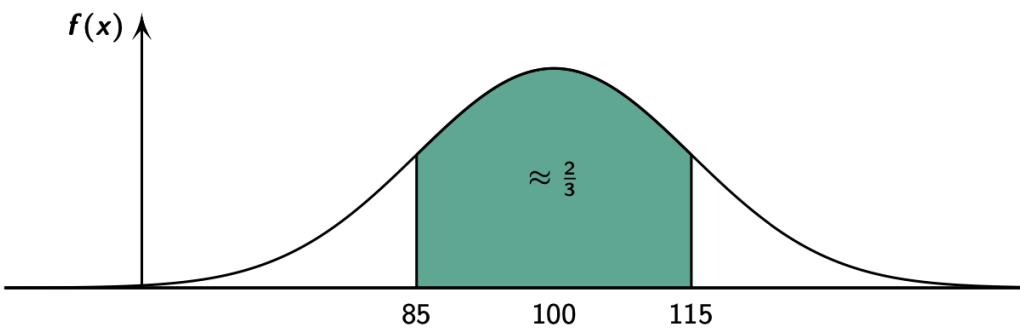
- Oder kürzer:
- ```
qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)
[1] 70.60054 129.39946
```
- 95 % der Menschen haben einen IQ zwischen ungefähr 70 und 130
  - Entspricht Abstand von etwa 2 Standardabweichungen vom Mittelwert  $\mu = 100$ .

- Wieviel Prozent der Bevölkerung liegen innerhalb einer Standardabweichung vom Mittelwert liegen?

- Gesucht W'keit:

$$P(85 \leq X \leq 115)$$

- W'keit als Fläche:



- Mit R:

```
pnorm(q = 115, mean = 100, sd = 15) - pnorm(85, 100, 15)
[1] 0.6826895
```

- D.h.: Etwa  $\frac{2}{3}$  der Bevölkerung hat einen IQ zwischen 85 und 115

## Normalverteilung: Eigenschaften

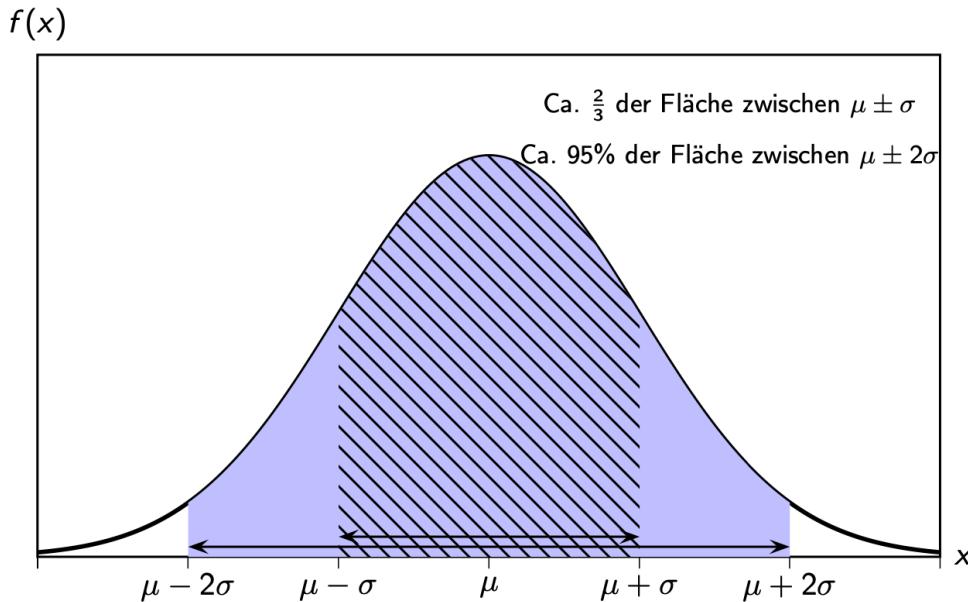
- Letzte Resultat aus Beispiel gilt für alle Normalverteilung  $\mathcal{N}(\mu, \sigma^2)$
- Die W'keit, dass eine Beobachtung eine höchstens Standardabweichung vom Erwartungswert abweicht, ist etwa  $\frac{2}{3}$ :

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx \frac{2}{3}$$

- Normalverteilung: Konkrete Aussage für die Streuung als „mittlere“ Abweichung vom Erwartungswert
- W'keit, dass eine Beobachtung höchstens zwei Standardeinheiten vom Erwartungswert abweicht:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

## Normalverteilung: Eigenschaften



## Gesetz der grossen Zahlen, Zentraler Grenzwertsatz

### Funktion von mehreren Zufallsvariablen

Üblicherweise wird dieselbe Grösse mehrmals gemessen (z.B: ein Gewicht wird mehrmals gemessen). Allgemein: Messungen  $x_1, x_2, \dots, x_n$  als Realisierung einer Zufallsvariablen auffassen:

$$X_1, \dots, X_n$$

$X_i$  ist die i-te Wiederholung vom Zufallsexperiment.

### Beispiel

- 20 Messungen der Wasserverschmutzung in einem See
- Die Messungen sind konkrete Werte:  $x_1, x_2, \dots, x_{20}$
- Die Realisierungen der Zufallsvariable sind:  $X_1, X_2, \dots, X_{20}$
- Annahme: 20 Zufallsvariablen mit gleicher Wahrscheinlichkeitsverteilung
- Die Wasserproben stammen alle aus demselben See und wurden mit einer identischen Methode gemessen. Interessant: Durchschnitt dieser Messungen und die Verteilung der zugehörigen Zufallsvariable

### Summe und Durchschnitt

Gegeben Zufallsvariable:  $X_1, X_2, \dots, X_{20}$

Summe =  $S_n = X_1, X_2, \dots, X_n = \sum_{i=1}^n X_i$

Arithmetisches Mittel:  $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} S_n$

### Beispiel: Warum ist der Durchschnitt wichtig?

- Untersuchen, ob Angabe 500 ml Inhalt einer Petflasche gilt
- Kaufen eine Flasche: Messen Inhalt 495.21ml
- Das ist weniger als 500 ml, aber ist es zu wenig?
- Bei einer Flasche möglich
- Idee: kaufen 100 Flaschen und messen Inhalt

- Durchschnitt 465.21 ml
- Scheint eindeutig zu wenig: Angabe 500 ml kann nicht stimmen
- Genaues Vorgehen: Siehe Hypothesentest

## Kennzahlen von $S_n$ und $\bar{X}_n$

Annahme:  $X_1, \dots, X_n$  i.i.d  $\rightarrow$  independent, identically distributed

Zweites «i» in i.i.d.:  $X_i$  dieselbe Verteilung mit denselben Kennzahlen:

$$E(X_i) = \mu \text{ und } Var(X_i) = \sigma^2 x$$

Gesucht: Erwartungswert und Varianz für

Summe  $S_n = X_1, X_2, \dots, X_n$

Durchschnitt :  $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$

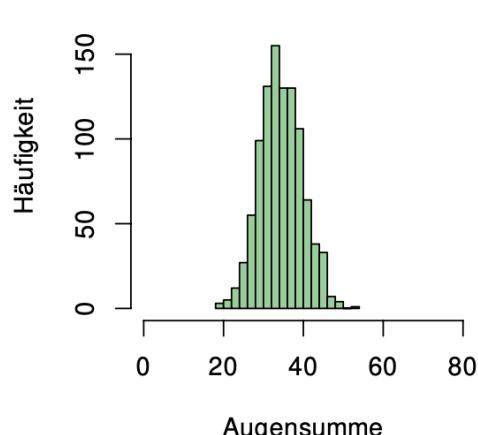
## Graphisches Beispiel

- Werfen eines fairen Würfels
- $X$ : Zufallsvariable für geworfene Augenzahl
- Erwartungswert:  $E(X) = \mu = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$
- Varianz:  $Var(X) = \frac{1}{6}((1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2) = 2.92$

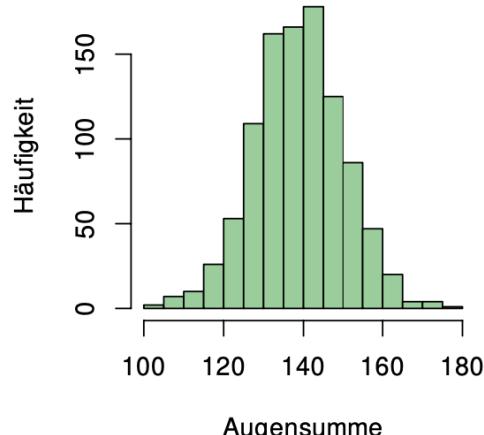
```
x <- c(1, 2, 3, 4, 5, 6)
ave <- mean(x)
var <- mean((x - ave)^2)
var
[1] 2.916667
```

- Würfeln 10mal
- Zufallsvariablen:  $X_1, X_2, \dots, X_{10}$  i.i.d.
- $X_i$ : Augenzahl im  $i$ -ten Wurf
- Erwartungswert und Varianz: Werte der ZV's  $X_i$  oben
- Notieren Augensumme  $s_{10}$  dieser 10 Würfel
- 1000 mal machen: Histogramm aller vorkommenden Augensummen
- Dasselbe mit 40 Würfen
- Simulation mit R

**Augensumme von 10 Würfen**



**Augensumme von 40 Würfen**



### Feststellung

Die Mittlere Augensumme verschiebt sich, wenn mehr Würfe gemacht werden. Abbildung links: Grösste Häufigkeit bei etwa 35, also  $10 \cdot 3.5 = 10 \cdot \mu$   
 $\mu = 3.5$ : Erwartungswert für einen Wurf

Abbildung rechts: Grösste Häufigkeit bei etwa 140  
 $40 \cdot 3.5 = 40 \cdot \mu$

Vermutung:  $E(S_n) = n \mu$

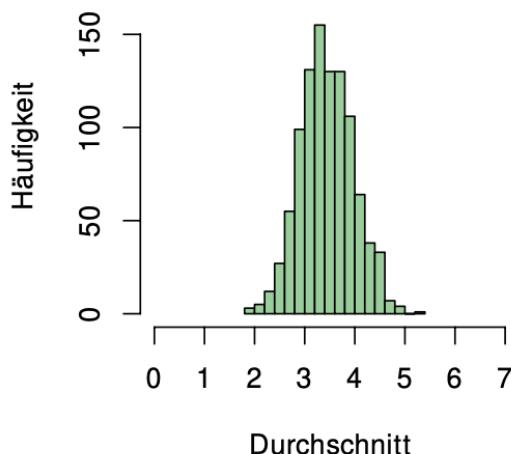
Die Varianz / Standardabweichung nimmt mit zunehmender Anzahl Würfen zu.

$$Var(S_n) = n * Var(X)$$

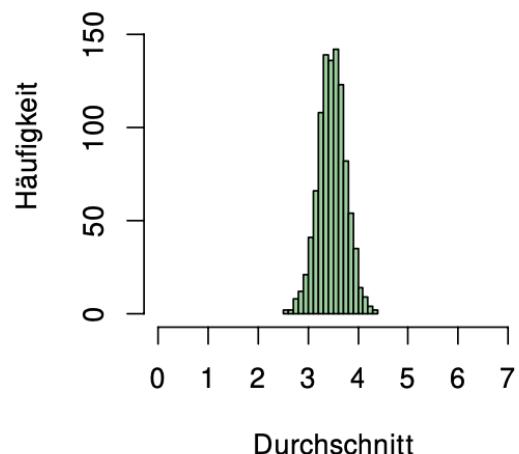
$$\sigma_{S_n} = \sqrt{n} \sigma_x$$

Dasselbe gilt für den Durchschnitt  $\bar{X}_n$

**Durchschnitt von 10 Würfen**



**Durchschnitt von 40 Würfen**



### Feststellung

Beide Histogramme: Grösste Häufigkeit bei 3.5, also  $\mu$

Vermutung  $E(\bar{X}_n) = \mu$

Die Varianz / Standardabweichung nimmt bei zunehmender Anzahl Würfen ab.

Gesetz:

$$Var(\bar{X}_n) = \frac{Var(X)}{n}$$

$$\sigma_{\bar{x}_n} = \frac{\sigma_x}{\sqrt{n}}$$

Oben gemachte Beobachtungen gelten allgemein.

## Allgemein

Annahme:  $X_1, \dots, X_n$  i.i.d.

Es gilt:

Kennzahlen von  $S_n$ :

$$\begin{aligned} E(S_n) &= n\mu \\ Var(S_n) &= n * Var(X_i) \\ \sigma(S_n) &= \sqrt{n}\sigma_x \end{aligned}$$

Kennzahlen von  $\bar{X}_n$ :

$$\begin{aligned} E(\bar{X}_n) &= \mu \\ Var(\bar{X}_n) &= \frac{\sigma_x^2}{n} \\ \sigma(\bar{X}_n) &= \frac{\sigma_x}{\sqrt{n}} \end{aligned}$$

## Bemerkungen

Standardabweichung von  $\bar{X}_n$ : Standardfehler des Arithmetischen Mittels

Standardabweichung der Summe: Wächst mit wachsendem  $n$ , aber langsamer als die Anzahl Beobachtungen  $n$

D. h.: Kleinere Streuung für wachsende  $n$

Erwartungswert von  $\bar{X}_n$ : Gleich demjenigen einer einzelnen Zufallsvariable ZV  $X_i$ , die Streuung nimmt jedoch ab mit wachsendem  $n$

### Gesetz der grossen Zahlen

Falls  $X_1, \dots, X_n$  i.i.d., dann  $\bar{X}_n \rightarrow \mu$  für  $n \rightarrow \infty$

Für  $n \rightarrow \infty$  geht die Streuung gegen null.

## Standardfehler

Standardabweichung des arithmetischen Mittels (Standardfehler) ist nicht proportional zu  $1/n$ , sondern nimmt ab mit dem Faktor  $1/\sqrt{n}$ :

$$\sigma_{\bar{x}_n} = \frac{1}{\sqrt{n}} \sigma_x$$

Um Standardfehler zu halbieren, braucht man also viermal so viele Beobachtungen. Dies nennt man auch das  $\sqrt{n}$ -Gesetz.

## Zentraler Grenzwertsatz

Bekannt: Kennzahlen von  $S_n$  und  $\bar{X}_n$

Unbekannt: Verteilung von  $S_n$  und  $\bar{X}_n$

- Würfelbeispiel:  $X_i$  gleichverteilt:

|            |  |               |               |               |               |               |               |
|------------|--|---------------|---------------|---------------|---------------|---------------|---------------|
| $x$        |  | 1             | 2             | 3             | 4             | 5             | 6             |
| $P(X = x)$ |  | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

- Wie sind  $S_n$  und  $\bar{X}_n$  verteilt?
- Vermutung wegen Slides 9 und 12: Beide normalverteilt
- Dies ist die Aussage des *Zentralen Grenzwertsatzes*
- Simulation der Aussage, kein Beweis

### Simulation von $\bar{X}_n$

- Ergebnismenge

$$\Omega = \{0, 10, 11\}$$

- Ziehen eine Zahl

- ZV  $X$ : Wert der gezogenen Zahl

- Es gilt:

$$P(X = 0) = P(X = 10) = P(X = 11) = \frac{1}{3}$$

- Erwartungswert von  $X$ :

$$E(X) = \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 10 + \frac{1}{3} \cdot 11 = 7$$

```
werte <- c(0, 10, 11)
ew <- sum(werte * 1/3)
ew
[1] 7
```

- Varianz von  $X$ :

$$\text{Var}(X) = \frac{1}{3} \cdot (0 - 7)^2 + \frac{1}{3} \cdot (10 - 7)^2 + \frac{1}{3} \cdot (11 - 7)^2 = 24.6667$$

```
var.X <- sum((werte - ew)^2 * 1/3)

var.X
[1] 24.66667
```

- Jetzt 10 Ziehungen
- Anzahl Ziehungen zu klein, aber man „sieht“ besser, was passiert
- Ein Versuch (10 Ziehungen):

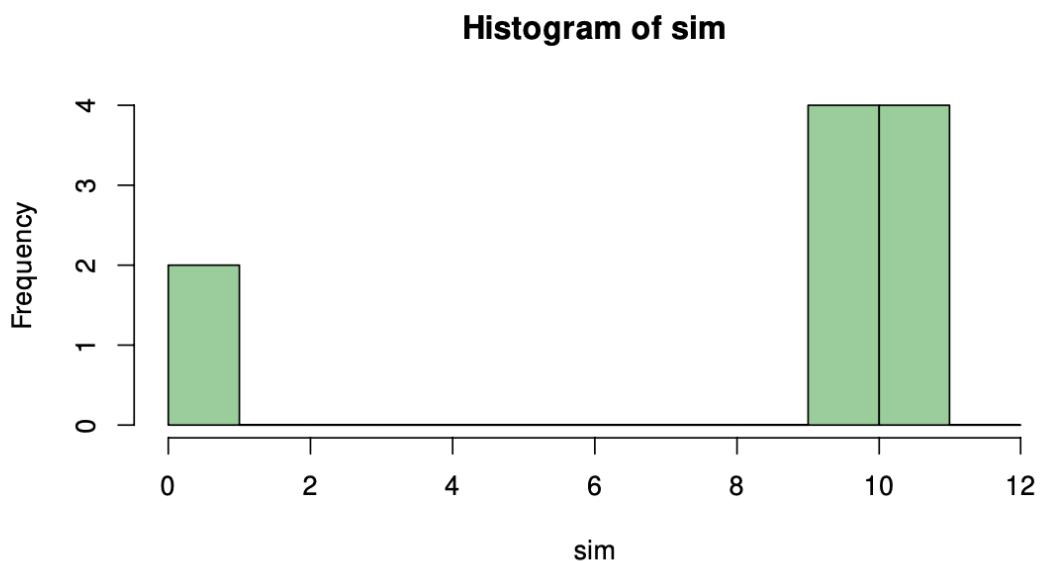
```
zieht 10-mal aus der Menge {0,10,11} einen Wert mit
gleicher W'keit
sim <- sample(werte, 10, replace = T)

Vektor mit 10 Werten
sim
[1] 0 10 11 11 11 11 10 10 0 10
```

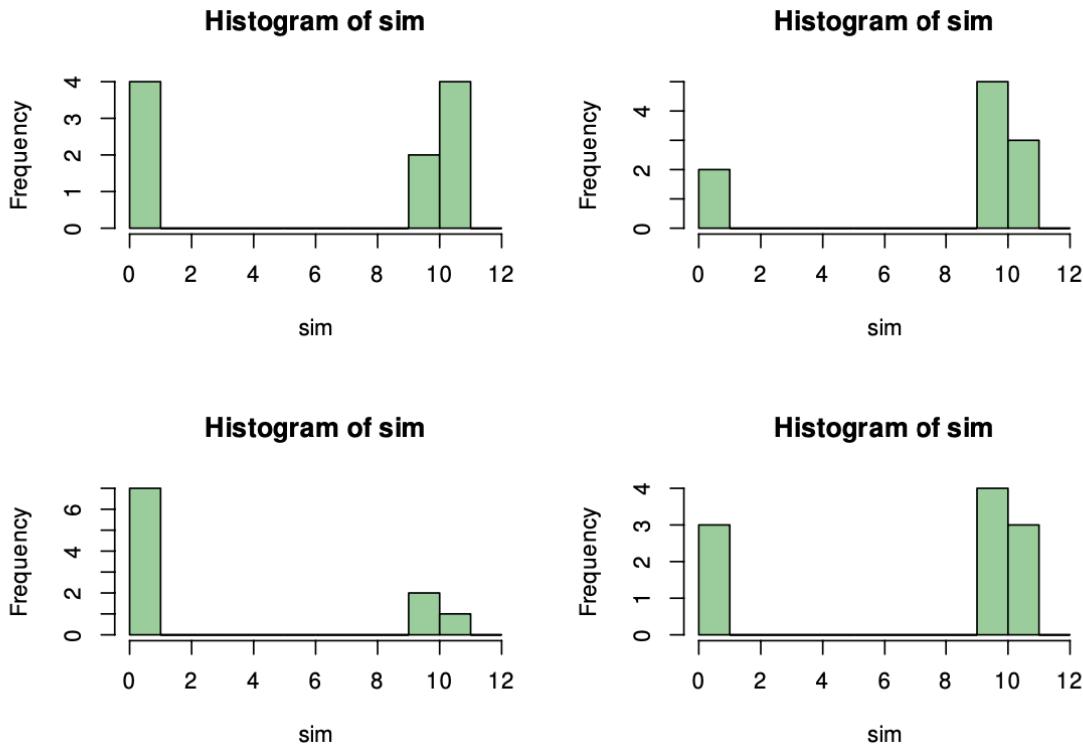
- `sample`: zieht zufällig Zahlen aus `werte`
- `replace = T`: Legt die Zahl nach dem Ziehen wieder zurück

- `hist`: Histogramm mit diesen 10 Werten

```
hist(sim, col = "darkseagreen3", breaks = 0:12)
```



- Bei jedem Versuch: Anderes Histogramm
- Histogramme von 4 Versuchen (10 Ziehungen):

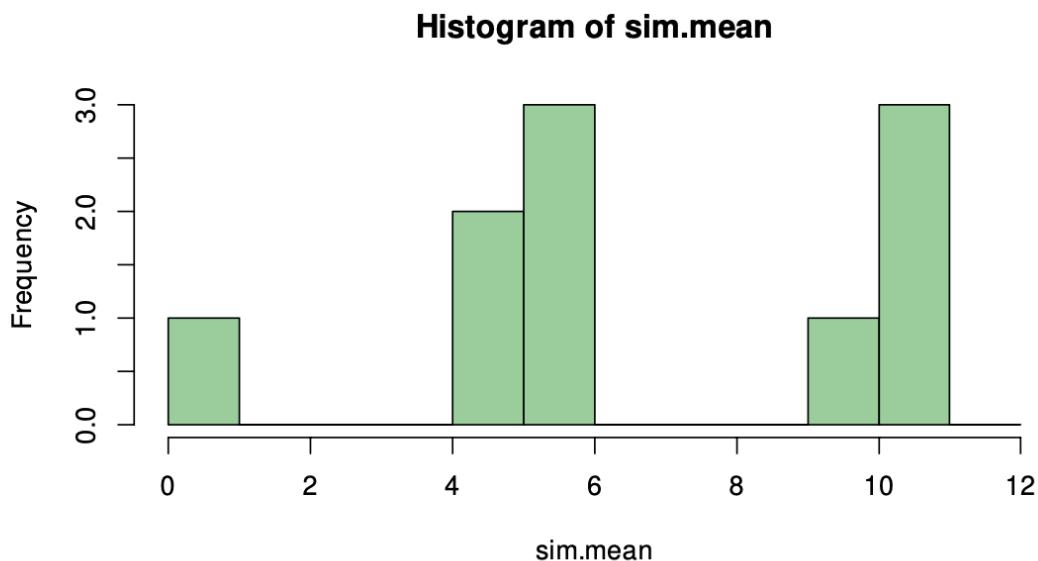


- Offensichtlich keine Normalverteilung
- Bis jetzt: Kommen nur die Zahlen 0, 10, 11 vor
- Nun: Zwei solche Versuche (je 10 Ziehungen) hintereinander ausführen
- Durchschnitt* aus beiden Versuchen berechnen:

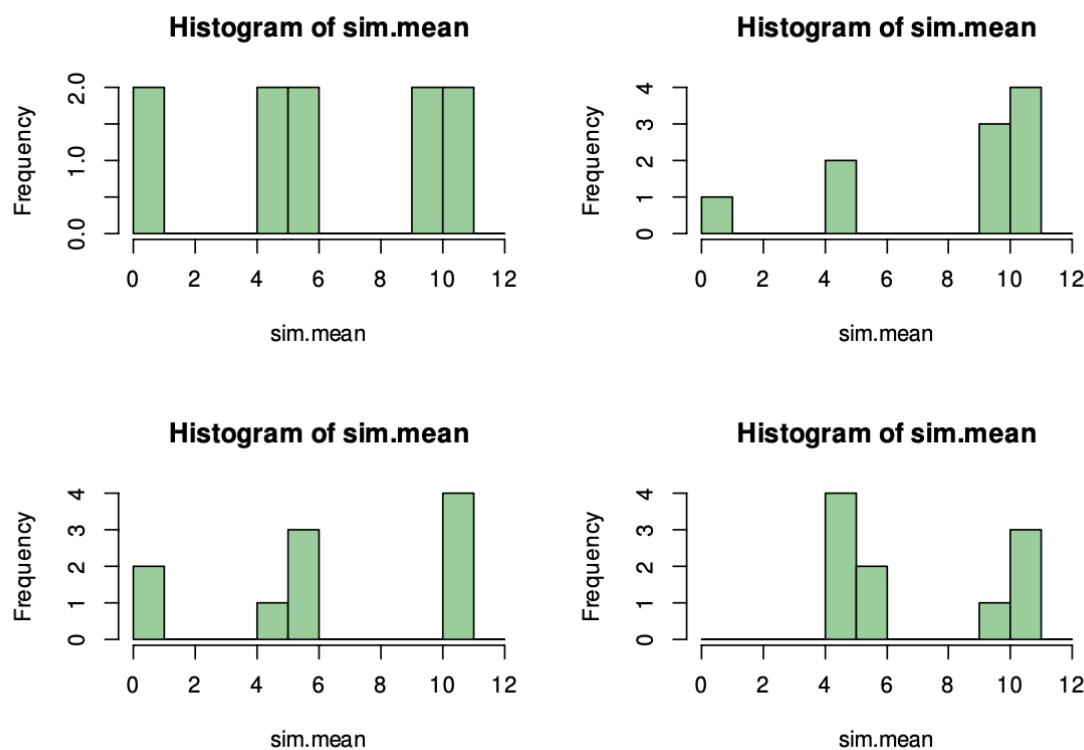
```
sim.1 <- sample(werte, 10, replace = T)
sim.1
[1] 0 11 0 10 0 11 11 10 10 11
sim.2 <- sample(werte, 10, replace = T)
sim.2
[1] 11 0 0 0 10 10 10 10 11 0
sim.mean <- (sim.1 + sim.2)/2
sim.mean
[1] 5.5 5.5 0.0 5.0 5.0 10.5 10.5 10.0 10.5 5.5
```

- Neben Zahlen 0, 10, 11: Auch Zahlen 5, 5.5 und 10.5 können vorkommen
- Histogramm:

```
hist(sim.mean, col = "darkseagreen3", breaks = 0:12)
```



- 4 Histogramme: Alle verschieden



Jeder Versuch sieht anders aus, aber Tendenzen zeichnen sich ab. 0 weniger oft vertreten, da doppelte 0 nur mit Wahrscheinlichkeit 1/9 vorkommt.

- Nun 3 Versuche wiederholen und Durchschnitt nehmen:

```

sim.1 <- sample(werte, 10, replace = T)
sim.1
[1] 10 10 0 11 11 10 11 10 0 10

sim.2 <- sample(werte, 10, replace = T)
sim.2
[1] 0 11 11 0 10 0 11 10 11 11

sim.3 <- sample(werte, 10, replace = T)
sim.3
[1] 0 0 10 10 10 0 0 0 10 0

sim.mean <- (sim.1 + sim.2 + sim.3)/3
round(sim.mean, 2)
[1] 3.33 7.00 7.00 7.00 10.33 3.33 7.33 6.67
[9] 7.00 7.00

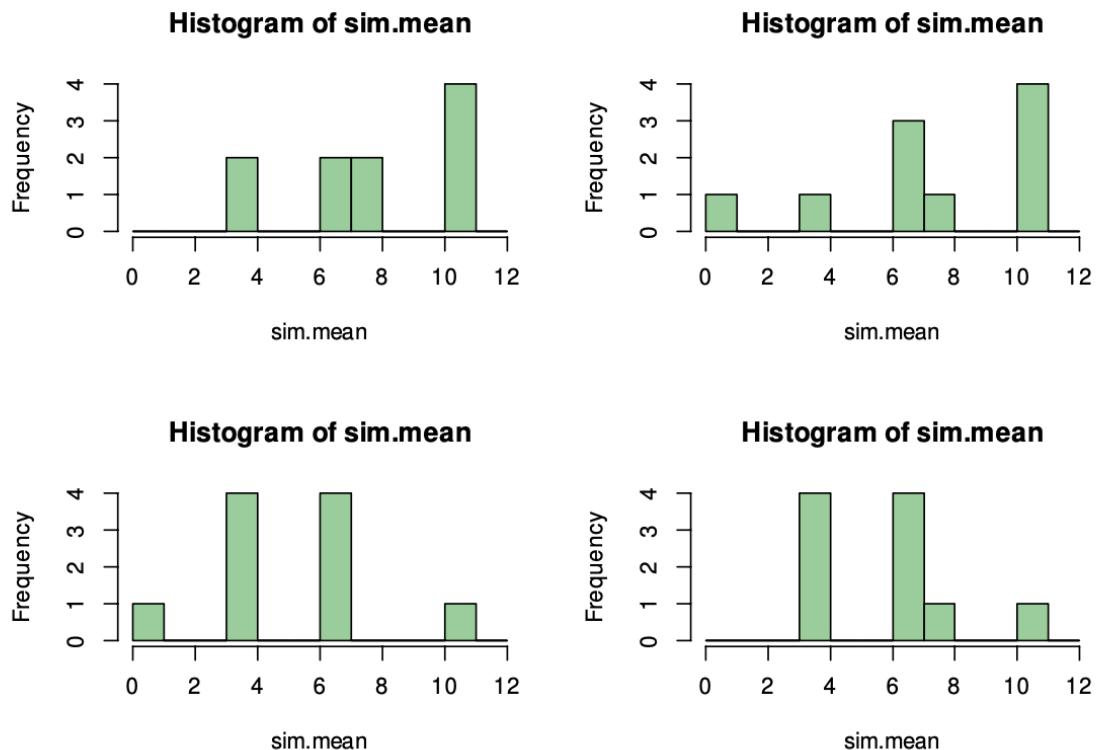
```

- Histogramm:

```
hist(sim.mean, col = "darkseagreen", breaks = 0:12)
```

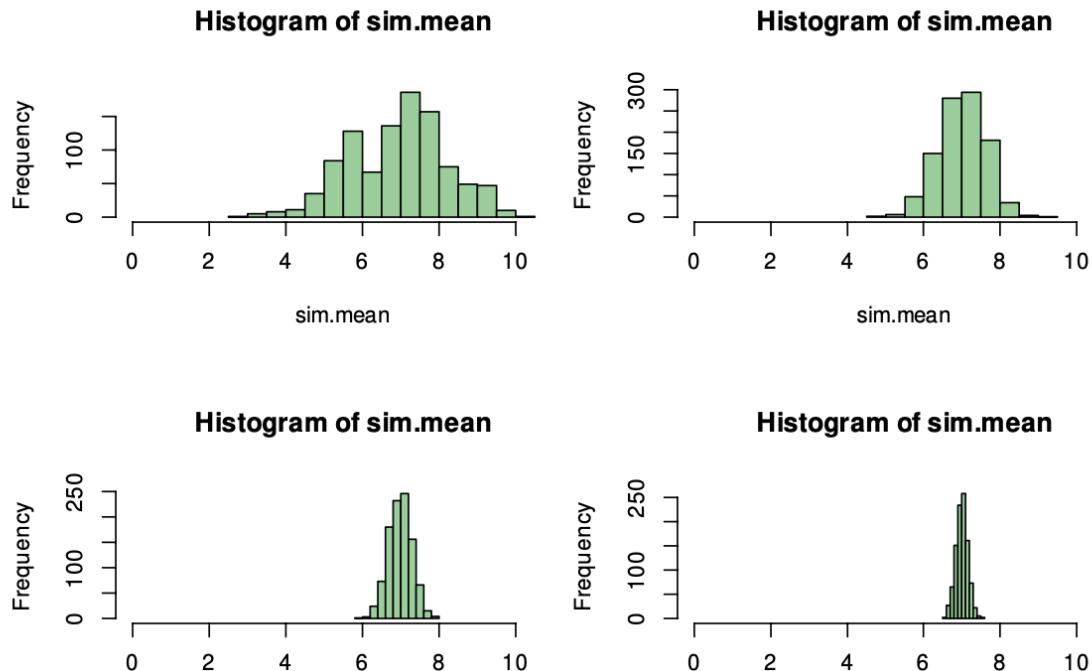


- Mehrere Versuche:



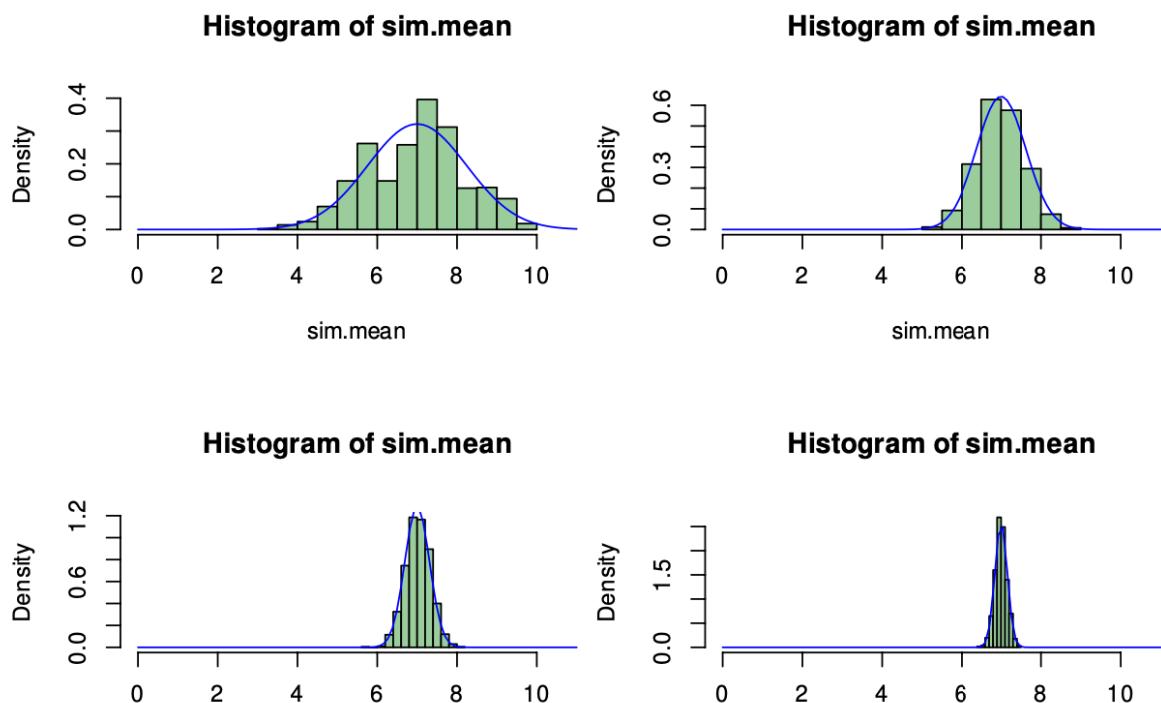
- Tendenz gegen den Erwartungswert 7
- Beim *Durchschnitt* gibt es immer mehr Werte
- Häufung um Erwartungswert 7
- Warum ist dies so?
- Zahl 0 im Durchschnitt kommt praktisch nicht mehr vor: W'keit, dass 3 mal an gleicher Stelle eine 0 vorkommt, ist nur noch  $\frac{1}{27}$
- Dasselbe für Zahl 11

- Nun 16, 64, 256 und 1024 solche Versuche mit jeweils 1000 Ziehungen
- Nehmen jeweils den Durchschnitt wie in den Beispielen vorher
- Histogramme:



- Bei genauerem Hinsehen fällt auf:
  - ▶ Werte häufen sich um den Erwartungswert 7
  - ▶ Standardabweichung wird kleiner: Halbiert sie sich etwa beim Vervierfachen der Anzahl Versuche
  - ▶ Histogramme scheinen einer Normalverteilung zu folgen
- Zeichnen noch die jeweiligen Dichtekurven für

$$\mathcal{N}\left(7, \frac{24.6667}{n}\right)$$



- Fällt auf: Dichtekurven für grössere  $n$  passen immer besser zu den Histogrammen
- Nochmals: Begannen mit Verteilung, die *nichts* mit einer Normalverteilung zu tun hat
- Aber: Verteilung *Mittelwerte*  $\bar{X}_n$  (oder Summen) nähert sich mit wachsendem  $n$  einer Normalverteilung an

## Zentraler Grenzwertsatz

- $X_i$ 's i.i.d. (nicht notwendig normalverteilt), dann gilt der berühmte

### Zentraler Grenzwertsatz

$X_1, \dots, X_n$  i.i.d. mit irgendeiner Verteilung mit Erwartungswert  $\mu$  und Varianz  $\sigma^2$ , dann gilt (ohne Beweis):

$$S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$$

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

- ▶ Approximation wird mit grösserem  $n$  i.A. besser
- ▶ Approximation besser, je näher die Verteilung von  $X_i$  bei der Normalverteilung  $\mathcal{N}(\mu, \sigma_X^2)$  ist

## Beispiel

- Strassenverkehrsamt hat genug Streusalz gelagert, um mit einem Schneefall von insgesamt 80 cm pro Jahr fertigzuwerden
- Täglich fallen im Mittel 1.5 cm mit einer Standardabw. von 0.3 cm
- Wie gross ist W'keit, dass das gelagerte Salz für die nächsten 50 Tage ausreicht?

- $X_i$ : ZV für die gefallene Menge Schnee am Tag  $i$
- Annahme: i.i.d.  $\rightarrow$  gerechtfertigt?
- Es gilt  $\mu = 1.5$  und  $\sigma_X = 0.3$
- Schneemenge (Summe)  $S_{50}$  der nächsten 50 Tage
- Soll 80 nicht übersteigen
- Es gilt annähernd:

$$S_{50} \sim \mathcal{N}(50 \cdot \mu, 50 \cdot \sigma_X^2) = \mathcal{N}(75, 4.5)$$

- Gesucht:

$$P(S_n \leq 80) = 0.991$$

```
pnorm(q = 80, mean = 50 * 1.5, sd = sqrt(50) * 0.3)
[1] 0.9907889
```

## Beispiel 2

- Die Lebensdauer eines bestimmten elektrischen Teils ist durchschnittlich 100 Stunden mit Standardabweichung von 20 Stunden
- Testen 16 solcher Teile
- Wie gross ist W'keit, dass das Stichprobenmittel
  - ▶ unter 104 Stunden oder
  - ▶ zwischen 98 und 104 Stunden liegt?

- $X_i$ : Zufallsvariable für die Lebensdauer des Teils  $i$

- Es gilt  $\mu = 100$  und  $\sigma_X = 20$

- Annahme i.i.d.

- Betrachten durchschnittliche Lebensdauer  $\bar{X}_{16}$

- Annähernd verteilt wie:

$$\bar{X}_{16} \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(100, \frac{20^2}{16}\right) = \mathcal{N}(100, 25)$$

- Gesucht:

$$P(\bar{X}_{16} \leq 104) = 0.788$$

```
pnorm(q = 104, mean = 100, sd = 20/sqrt(16))
[1] 0.7881446
```

- Gesucht:

$$P(98 \leq \bar{X}_{16} \leq 104) = 0.444$$

```
pnorm(q = 104, mean = 100, sd = 20/sqrt(16)) - pnorm(98,
100, 20/sqrt(16))
[1] 0.4435663
```

## Hypothesentest

Hypothesentests sind ein wichtiges statistisches Mittel um zu entscheiden, ob eine Messreihe zu einer gewissen Grösse passt.

Beispiel:

- Brauerei bestellt eine neue Abfüllmaschine für 500ml Büchsen
- Abfüllmaschine füllt nie genau 500ml ab, sondern nur ungefähr
  - o Mal einen Tropen mehr, mal einer weniger
- Für Brauerei wichtig, dass Abfüllmaschine möglichst genau abfüllt
  - o Füllt Maschine zu viel ab, so ist dies schlecht für die Brauerei, da sie zu viel Bier für den angegebene Preis verkauft
  - o Füllt sie zu wenig ab, sind Kunden unzufrieden, da sie für den angegebenen Prei zu wenig Bier bekommen

Herstellerfirma behauptet: Maschine füllt Büchsen normalverteilt mit  $\mu = 500$  ml und  $\sigma = 1$  ml ab

- Brauerei macht 100 Stichproben
- Mittelwert dieser Stichproben ist 499.57ml

Weniger als 500 ml, aber liegt dies noch innerhalb der Angaben  $\mu = 500$  ml und  $\sigma = 1$  ml des Herstellers der Abfüllanlage? Wie können wir dies überprüfen?

- Wäre Mittelwert 421.54 ml, so würden wir reklamieren
- Wo ist die Grenze zwischen „ok“ und „nicht ok“?

Allgemeiner: Sie stellen eine Maschine her und müssen sich auf die Angaben der Spezifikationen der Hersteller für die Bestandteile verlassen können. Wie können wir feststellen, dass die Bestandteile die Spezifikationen auch erfüllen?

(Fiktive) Anfrage beim Bundesamt für Statistik: Durchschnittliche Körpergrösse der erwachsenen Frauen liegt in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm. Angabe ist gefühlsmässig wohl falsch, da viel zu hoch.  
Wie können wir dies aber mathematisch überprüfen und begründen, ohne uns auf unser Gefühl zu verlassen?

Ziel: Standardisiertes, reproduzierbares Verfahren einzuführen, mit dem wir entscheiden können, ob der Mittelwert einer Messreihe zu einem bestimmten „wahren“ Mittelwert  $\mu$  passt oder nicht.

Achtung: Das folgende Verfahren liefert niemals einen Beweis, dass beispielsweise eine Grösse nicht zu einer Messreihe passt

Können mit statistischen Mitteln nur zeigen, dass diese Grösse mit grosser Wahrscheinlichkeit nicht zu dieser Messreihe passt

Lesen Sie in der Zeitung „... mit Statistik bewiesen. . .“, ist das ein Blödsinn!

Waagebeispiel von früher:

|         |       |       |       |       |       |       |       |       |       |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Waage A | 79.98 | 80.04 | 80.02 | 80.04 | 80.03 | 80.03 | 80.04 | 79.97 | 80.05 |
| Waage A | 80.03 | 80.02 | 80.00 | 80.02 |       |       |       |       |       |
| Waage B | 80.02 | 79.94 | 79.98 | 79.97 | 79.97 | 80.03 | 79.95 | 79.97 |       |

Messungen als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_i$  betrachten.

Zweite Messwert  $x_2 = 80.04$  der Waage A eine Realisierung der Zufallsvariable  $X_2$ .

- Betrachten Messdaten  $x_1, \dots, x_n$  als Realisierungen von

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2)$$

- Zwei Kennzahlen der Zufallsvariablen  $X_i$  sind:

$$\mathbb{E}(X_i) = \mu \quad \text{und} \quad \text{Var}(X_i) = \sigma_X^2$$

- Typischerweise sind diese (und andere) Kennzahlen unbekannt
- Ziel: Rückschlüsse darüber aus den Daten
- (Punkt-) Schätzungen für den Erwartungswert und die Varianz sind:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$
$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Der Hut über dem Zeichen bedeutet = Schätzung.

Beispiel Waage A:

- Schätzungen für den Mittelwert  $\mu$  und die Varianz  $\sigma_X^2$ :

$$\hat{\mu} = 80.02 \quad \text{und} \quad \hat{\sigma}_X^2 = 0.024^2$$

- R:

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
 79.97, 80.05, 80.03, 80.02, 80, 80.02)

mean(waageA)
[1] 80.02077
sd(waageA)
[1] 0.02396579
```

- Problem: für *andere* Messreihen lauten diese Schätzwerte praktisch immer anders

- Jetzt: Messreihen simulieren, die „ähnlich aussehen“, wie die Werte in Waage A

- Annahme:* Messwerte in Waage A normalverteilt mit *wahren* Parametern:

$$\mu = 80 \quad \text{und} \quad \sigma_X^2 = 0.02^2$$

- Generieren mit `rnorm` Zufallszahlen, die dieser Verteilung folgen

- Wegen Übersichtlichkeit: Messreihen der Länge 6

- Runden meist auf zwei Nachkommastellen (`round(..., 2)`)

- `set.seed(...)`: bringt immer dieselben Zufallszahlen

- Code:

```
set.seed(1)
waageA.sim1 <- round(rnorm(n = 6, mean = 80, sd = 0.02), 2)

waageA.sim1
[1] 79.99 80.00 79.98 80.03 80.01 79.98
mean(waageA.sim1)
[1] 79.99833
sd(waageA.sim1)
[1] 0.0194079
```

- Geschätzte Werte  $\hat{\mu}$  und  $\hat{\sigma}^2$ : (Leicht) anders, als in Beispiel vorher

- Führen dies fünfmal durch:

```
set.seed(10)
for (i in 1:5) {
 waageA.sim1 <- round(rnorm(n = 6, mean = 80, sd = 0.02),
 4)
 cat(round(mean(waageA.sim1), 2), round(sd(waageA.sim1), 4),
 "\n")
}
80 0.0131
79.99 0.0213
80 0.0142
79.99 0.0257
79.99 0.0087
```

- Mittelwerte sind hier alle nahe bei 80, was auch zu erwarten war
- Keine Zweifel, dass der wahre Mittelwert nicht  $\mu = 80$  sein könnte
- Abweichungen sind durchaus zu erwarten
- Beispiel vorher: geschätzte Mittelwerte alle sehr nahe bei  $\mu = 80$
- Allerdings sind auch folgende Fälle möglich:

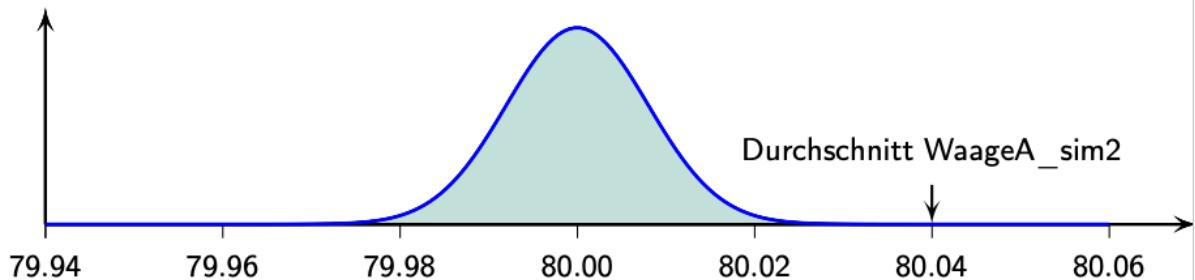
```
set.seed(1450070)
waageA.sim2 <- rnorm(n = 6, mean = 80, sd = 0.02)

waageA.sim2
[1] 80.05403 80.03896 80.03671 80.06336 80.01052 80.04372
mean(waageA.sim2)
[1] 80.04122
sd(waageA.sim2)
[1] 0.01804572
```

- Mittelwert dieser Messreihe verteilt wie (ZGWS):

$$\bar{X}_6 \sim \mathcal{N} \left( 80, \frac{0.02^2}{6} \right) = \mathcal{N} (80, 0.0082^2)$$

- Mittelwert Messreihe fast 5 Standardabweichungen grösser als 80
- Möglich, aber nicht sehr wahrscheinlich
- Abbildung:



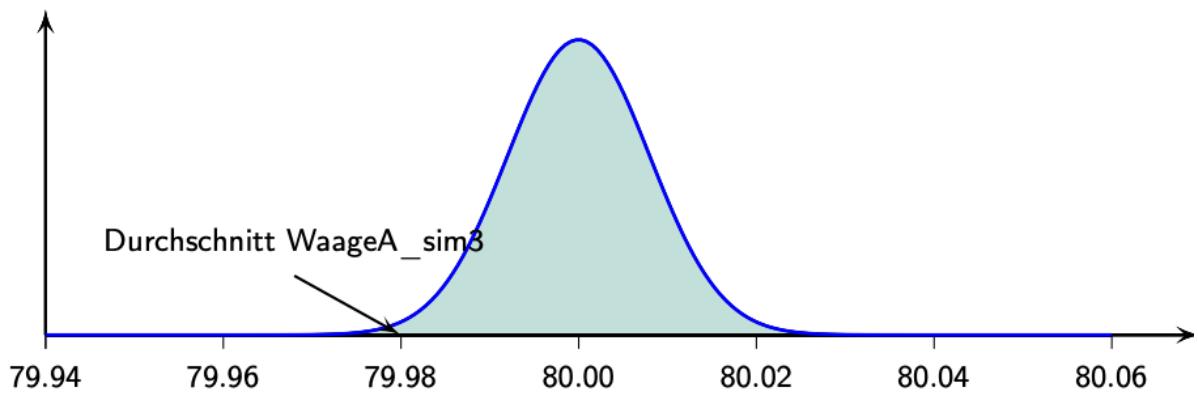
- Aber was heisst hier „nicht sehr wahrscheinlich“?
- Erwarten, dass der Mittelwert in der Nähe von  $\mu = 80$  liegt, sofern der wahre Mittelwert tatsächlich  $\mu = 80$  ist
- Ein weiteres Beispiel:

```
set.seed(384)
waageA.sim3 <- rnorm(n = 6, mean = 80, sd = 0.02)

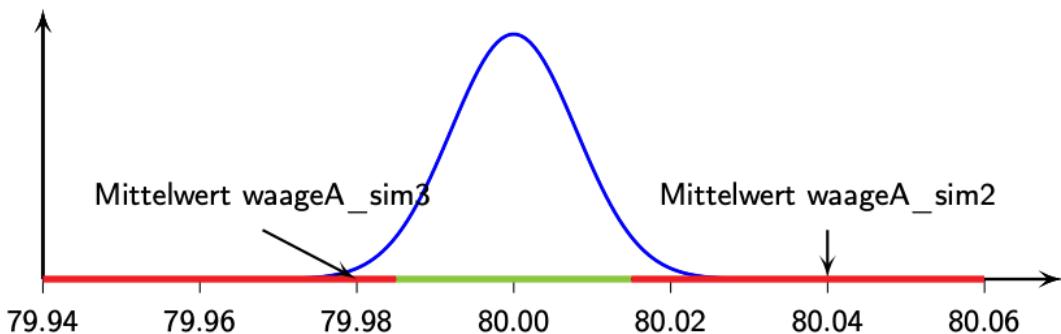
waageA.sim3
[1] 80.00420 79.95783 79.96086 79.95553 79.97645 79.99413
mean(waageA.sim3)
[1] 79.97483
sd(waageA.sim3)
[1] 0.02046691
```

- Mittelwert etwa 3 Standardabweichungen unter 80
- Dies ist zwar immer noch weit weg, aber nicht so stark wie im vorher

- Abbildung:



- Erwarten allerdings, dass der Durchschnitt der Messreihe in der Nähe vom wahren  $\mu = 80$  liegt
- Liegt der Durchschnitt weit weg  $\rightarrow$  beginnen zu zweifeln, ob der wahre Mittelwert tatsächlich 80 ist
- Idee: Legen Bereich fest, was „nahe bei“ oder „weit entfernt von“  $\mu$  ist
- Skizze:



Fragestellung:

- Ist eine Messreihe mit der Annahme  $\mu = 80$  noch kompatibel oder müssen an dieser Annahme zweifeln?
- Das heisst: Liegt der Mittelwert der Messreihe in der „Nähe“ des wahren Mittelwertes  $\mu = 80$  oder liegt er so „weit“ entfernt, dass wir an der Angabe des wahren  $\mu = 80$  zweifeln müssen?
- Hier stellt sich natürlich die Frage, was „nahe“ heisst (gleich)
- Der wahre Mittelwert ist grundsätzlich *nicht* bekannt

## Beispiel Vorgehen Hypothesentest

- Annahme: Daten normalverteilt sind mit  $\mu = 80.00$  und  $\sigma = 0.02$
- Wie kann man überprüfen, ob der Mittelwert  $\mu = 80$  auch realistisch ist?
- Grundidee: Mit Messreihe überprüfen, ob *unter dieser Annahme*  $\mu = 80$ , der Mittelwert dieser Messreihe w'lich ist oder nicht
- Wählen dazu eine Messreihe der Länge 6 aus und gehen von folgendem Modell aus:

### Modell

6 Messwerte sind Realisierungen der Zufallsvariablen  $X_1, X_2, \dots, X_6$ , wobei  $X_i$  eine kontinuierliche Messgrösse ist. Es soll gelten:

$$X_1, \dots, X_6 \text{ i.i.d. } \sim \mathcal{N}(80, 0.02^2)$$

- Wollen nun überprüfen, ob die *Annahme*  $\mu = 80$  auch gerechtfertigt ist
- Führen folgende Begriffe ein:

### Nullhypothese

$$H_0 : \mu = \mu_0 = 80$$

### Alternativhypothese

$$H_A : \mu \neq \mu_0 = 80 \quad \text{oder } „<“ \text{ oder } „>“$$

- Wählen Messreihe `waageA.sim3`:

```
[1] 80.00 79.96 79.96 79.96 79.98 79.99
Mittelwert: 79.975
Standardabweichung: 0.01760682
```

- Geschätzter Mittelwert:  $\hat{\mu} = 79.98$
- Konkretisieren, was es heisst, dass dieser Mittelwert (un)wahrscheinlich ist
- Folgende W'keit bringt uns hier nicht weiter, da diese 0 ist:

$$P(\bar{X}_6 = 79.98) = 0$$

- Da  $\hat{\mu} < 80$  ist, betrachten folgende Wahrscheinlichkeit:

$$P(\bar{X}_6 \leq 79.98)$$

- Verteilung von  $\bar{X}_6$  unter unseren Annahmen  $\mu = 80$  und  $\sigma = 0.02$ :

$$\bar{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

- Testen mit dieser Verteilung, ob die Annahme  $\mu = 80$  gerechtfertig ist

### Teststatistik

Verteilung der Teststatistik  $T$  unter der Nullhypothese  $H_0$ :

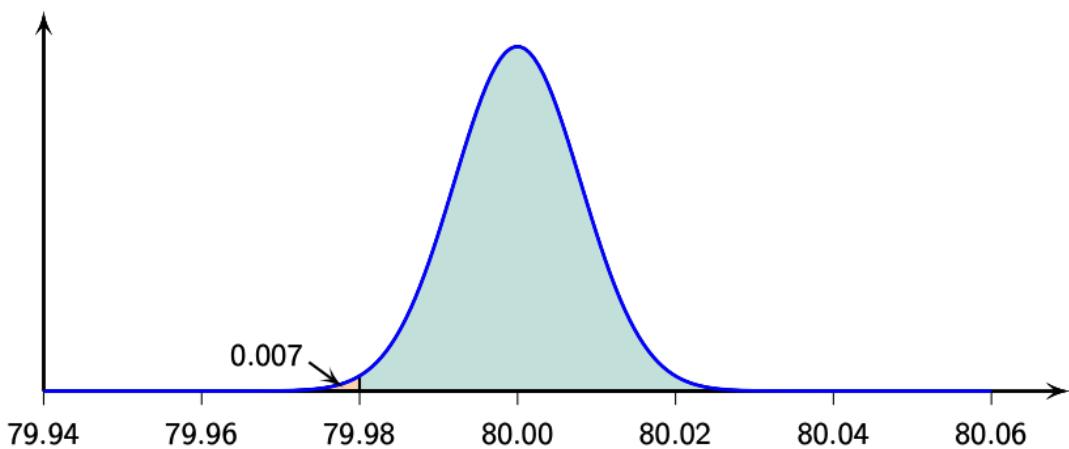
$$T = \bar{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

- Erhalten für die W'keit

$$P(\bar{X}_6 \leq 79.98) = 0.007$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))
[1] 0.007152939
```

- Skizze:



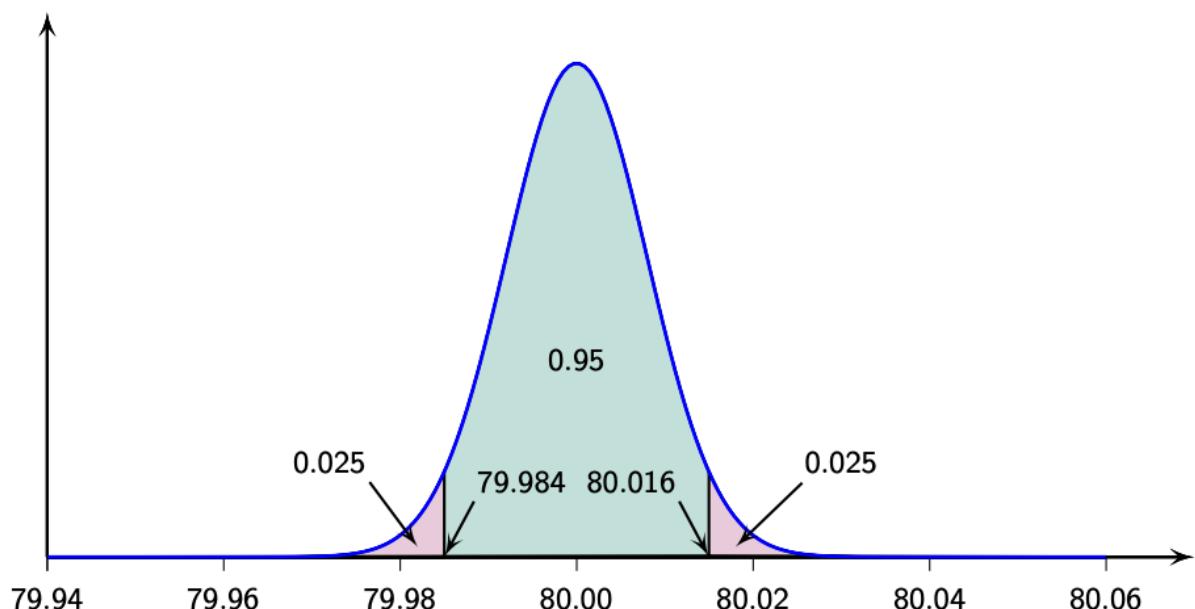
- Diese W'keit ist klein: 0.7 %
- Ist sie aber *zu* klein?
- Nun kommt eine *Abmachung*: Es hat sich als praktisch erwiesen, diese Grenze, was zu klein ist und was nicht bei 2.5 % festzulegen
- Warum dies 2.5 % sind, kommt gleich
- Gemäss dieser Abmachung ist:

$$P(\bar{X}_6 \leq 79.98) < 0.025$$

- Geschätzter Mittelwert  $\hat{\mu} = 79.98$  zu *zu unwahrscheinlich*, als dieser zum Wert  $\mu = 80$  passen könnte
- *Gehen also davon aus, dass der angegebene Mittelwert von  $\mu = 80$  nicht stimmen kann!*

## Graphische Darstellung

- Normalverteilungskurve in drei Teile auf:



- Symmetrischer Teil um Mittelwert  $\mu = 80$  soll 0.95 (95 %) betragen
- Beide Teile links und rechts müssen zusammen 0.05 ergeben
- Also ergibt sich für jeden Teil 0.025
- Grenzen entsprechen den 0.025- und 0.975-Quantilen

```
qnorm(p = c(0.025, 0.975), mean = 80, sd = 0.02/sqrt(6))
[1] 79.984 80.016
```

- Fläche 0.05 des gesamten roten Bereiches heisst *Signifikanzniveau*

### Signifikanzniveau $\alpha$

Signifikanzniveau  $\alpha$ , gibt an, wie hoch das Risiko ist, das man bereit ist einzugehen, eine falsche Entscheidung zu treffen  
Für die meisten Tests wird ein  $\alpha$ -Wert von 0.05 bzw. 0.01 verwendet. Hier

$$\alpha = 0.05$$

- Liegt der gemessene Mittelwert im roten Bereich in Abbildung, so zweifelt man an der Nullhypothese

$$H_0 : \mu = 80$$

- Wir sagen, wir *verwerfen* die Nullhypothese  $\mu = 80$
- Bereich, wo die Nullhypothese verworfen wird, heisst deshalb

### Verwerfungsbereich

$$K = (-\infty, 79.984] \cup [80.016, \infty)$$

- Liegt der gemessene Mittelwert im roten Bereich in Abbildung, so zweifelt man an der Nullhypothese

$$H_0 : \mu = 80$$

- Wir sagen, wir *verwerfen* die Nullhypothese  $\mu = 80$
- Bereich, wo die Nullhypothese verworfen wird, heisst deshalb

### Verwerfungsbereich

$$K = (-\infty, 79.984] \cup [80.016, \infty)$$

- Gehen davon aus, dass ein Mittelwert einer Messreihe im Verwerfungsbereich so unwahrscheinlich ist, dass an der Richtigkeit von  $\mu = 80$  gezweifelt wird
- Müssen annehmen, dass das wahre  $\mu$  nicht 80 ist
- Mit Messreihe überprüfen, ob deren Mittelwert im Verwerfungsbereich liegt oder nicht
- Machen den sogenannten

### Testentscheid

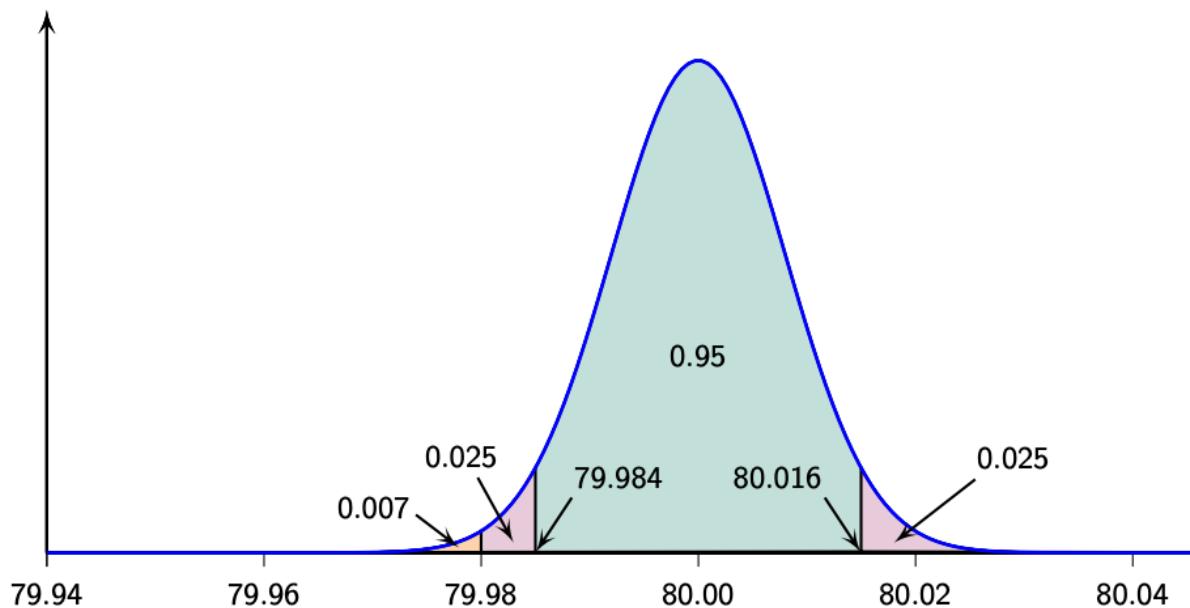
- ▶ In Beispiel oben

$$\bar{X}_6 = 79.98 \in K$$

- ▶ Dieser Wert liegt im Verwerfungsbereich
- ▶ Gehen nicht vom wahren  $\mu = 80$  aus, da der Mittelwert der Messreihe nicht zu diesem Parameter passt
- ▶ D.h.: Dieser Wert ist zu unwahrscheinlich, als dass  $\mu = 80$  plausibel ist
- ▶ Nullhypothese wird verworfen und Alternativhypothese angenommen:

$$\mu \neq 80$$

- Abbildung:



- Wählen *andere* Messreihe: Beispiel früher

```
[1] 80.05403 80.03896 80.03671 80.06336 80.01052 80.04372
Mittelwert: 80.04122
```

- Modell, Nullhypothese, Alternativhypothese, Teststatistik, Signifikanzniveau und Verwerfungsbereich gleich vorher
- Nur noch den Testentscheid durchführen
- Geschätzter Mittelwert ist im Verwerfungsbereich
- Somit wird auch hier die Nullhypothese verworfen

- Es gilt für die W'keit

$$P(\bar{X}_6 > 80.04) \approx 5 \cdot 10^{-7}$$

```
1 - pnorm(q = 80.04, mean = 80, sd = 0.02/sqrt(6))
[1] 4.816785e-07
```

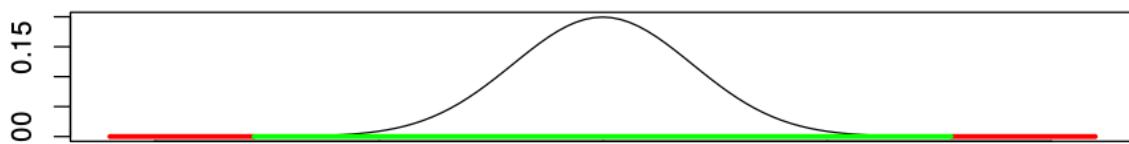
- Bei weitem kleiner als 0.025
- Damit so unwahrscheinlich, dass man auch auf diese Weise  $\mu = 80$  als nicht richtig annehmen (muss)
- Nullhypothese wird *verworfen*
- Verwerfungsbereich: Nur Entscheidung fällen, ob der geschätzte Mittelwert im Verwerfungsbereich liegt oder nicht
- Wert von  $P(\bar{X}_6 > 80.04)$  noch eine Aussage über die Sicherheit des Verwerfen
- In diesem Fall ist  $5 \cdot 10^{-7}$  sehr viel kleiner als 0.025 und damit können wir mit grosser Sicherheit davon ausgehen, dass  $\mu = 80$  nicht gilt
- Siehe *p*-Wert
- Aber nochmals: Messreihe stammt von der wirklichen Verteilung  $\mathcal{N}(80.00, 0.02^2)$
- Allerdings ist sie so unwahrscheinlich, dass an der Annahme  $\mu = 80$  gezweifelt werden muss

## Bemerkungen

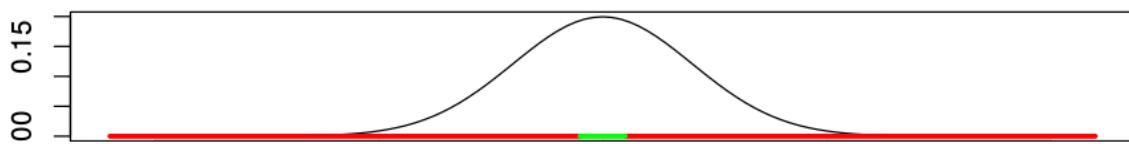
- Warum wurde Verwerfungsbereich nach oben und nach unten aufgeteilt, wenn man schon weiß, dass der gemessene Mittelwert kleiner als  $\mu = 80$  ist?
- Vor der Messung war dies nicht bekannt
- Der gemessene Mittelwert hätte also durchaus auch größer als  $\mu = 80$  sein können
- Man spricht in diesem Fall von einem *zweiseitigen Test*
- Es gibt auch *einseitige Tests* (Beispiel gleich)
- Annahme hier: Gesamter Verwerfungsbereich 5 %
- Annahme hat sich als praktisch erwiesen, aber auch 1 % oft gewählt

## Signifikanzniveau

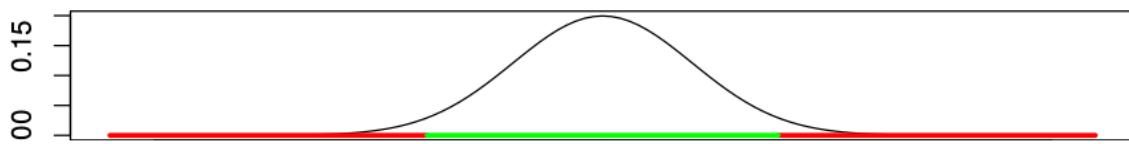
- Graphik:  $\alpha = 0.0001$  (nahe bei 0)



- Graphik:  $\alpha = 0.8$  (gross)



- Graphik:  $\alpha = 0.05$



- Ist  $\alpha$  sehr nahe bei null, so Bereich wo *nicht* verworfen wird (grüner Bereich) sehr gross
- D.h.: Es braucht ein sehr Ereignis bis verworfen wird
- Es wird viel zu wenig verworfen
- Im Extremfall  $\alpha = 0$ : Es wird gar nicht verworfen
- Für  $\alpha$  gross: Grüner Bereich sehr klein
- D.h.: Es braucht ein sehr Ereignis bis verworfen wird
- Es wird viel zu wenig verworfen
- Im Extremfall  $\alpha = 1$ : Es wird immer verworfen
- $\alpha = 0.05$ : Kompromiss zwischen den beiden Extremen

### Beispiel Abfüllanlage

- Testen, ob die Angabe in Beispiel Abfüllanlage mit der Testreihe konform ist
- Herstellerfirma behauptet, dass die Maschine die Büchsen normalverteilt mit  $\mu = 500 \text{ ml}$  und  $\sigma = 1 \text{ ml}$  abfüllt
- Brauerei macht 100 Stichproben
- Mittelwert dieser Stichproben ist  $499.84 \text{ ml}$
- Annahme: Messungen sind normalverteilt mit bekanntem  $\sigma = 1$

- *Modell*

$X_i$ : Inhalt der  $i$ -ten Büchse

$$X_1, \dots, X_{100} \text{ i.i.d. } \sim \mathcal{N}(\mu, 1^2)$$

- *Nullhypothese*

$$H_0 : \mu = 500$$

- *Alternativhypothese*

$$H_A : \mu \neq \mu_0 = 500$$

- *Teststatistik mit Signifikanzniveau  $\alpha = 0.05$*

$$\bar{X}_{100} \sim \mathcal{N}\left(500, \frac{1^2}{100}\right)$$

Das Modell ist unsere Stichprobe. Jede Büchse sollte 500ml besitzen und eine Abweichung von 1ml.

Die Nullhypothese ist das, was wir erwarten, also 500ml.

Die Alternativhypothese könnte mit Erfahrungswerten gefüllt werden.

- *Verwerfungsbereich*

Grenze des Verwerfungsbereichs:

```
qnorm(p = c(0.025, 0.975), mean = 500, sd = 1/sqrt(100))
[1] 499.804 500.196
```

- *Also*

$$K = (-\infty, 499.804) \cup (500.196, \infty)$$

- *Testentscheid* Es gilt

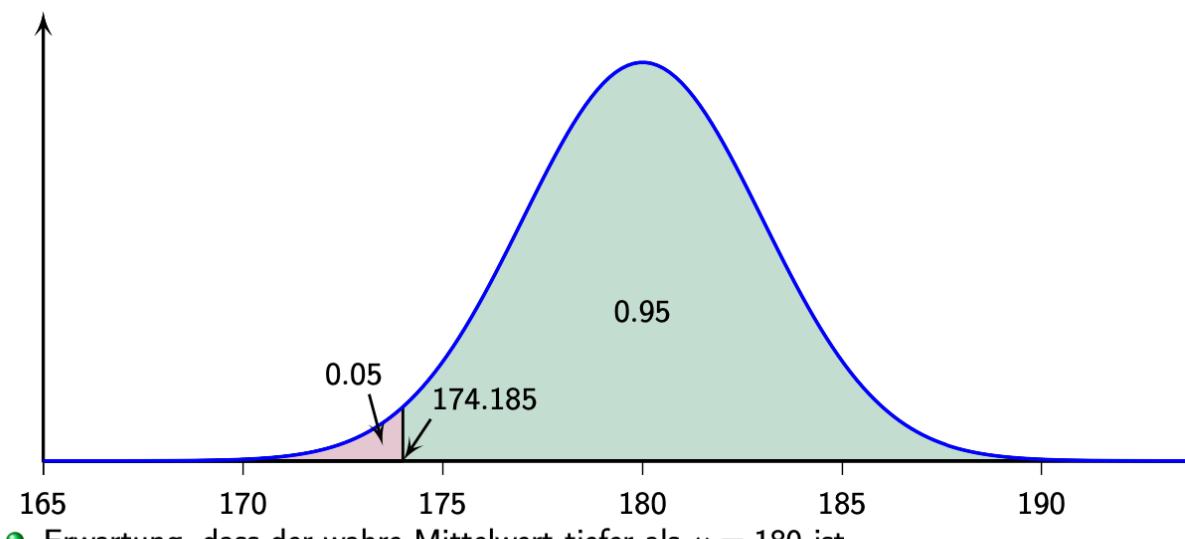
$$499.84 \notin K$$

- Nullhypothese wird nicht verworfen

- Vertrauen der Angabe des Hersteller der Abfüllanlage

### Beispiel: Körpergrösse Frauen

- Bundesamt für Statistik behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm liegt
- Vermutung: dieser Wert ist zu gross
- Zweiseitiger Test macht wenig Sinn, da man „weiss“, dass dieser Mittelwert zu gross ist
- D.h.: der wahre Wert liegt wohl eher tiefer
- Überlegung ist an sich dieselbe wie in Beispielen
- Verwerfungsbereich nicht auf beide Seiten verteilen, sondern nur nach unten



- Wählen zufällig 8 erwachsene Frauen aus, deren durchschnittliche Körpergrösse 171.54 cm beträgt (was immer noch sehr gross ist)
- Annahme: Körpergrösse normalverteilt mit  $\mathcal{N}(\mu, 10^2)$
- Annahme: Standardabweichung dieselbe, wie vom Bundesamt angegeben
- *Modell:*  
 $X_i$ : Körpergrösse der  $i$ -ten Frau. Es gilt

$$X_1, \dots, X_8 \text{ i.i.d. } \sim \mathcal{N}(\mu, 10^2)$$

- Gehen davon aus, dass der wahre Mittelwert wirklich 180 cm ist
- *Nullhypothese*

$$H_0 : \quad \mu_0 = 180$$

- *Alternativhypothese*

$$H_A : \quad \mu < \mu_0 = 180$$

- Testen ob jetzt der Wert

$$P(\bar{X}_8 < \bar{x}_8) < 0.05$$

ist oder nicht

- Verwerfungsbereich ist hier also einseitig nach unten
- Abbildung oben: Verwerfungsbereich für  $n = 8$  pink eingezeichnet

- Teststatistik mit Signifikanzniveau  $\alpha = 0.05$

$$\bar{X}_8 \sim \mathcal{N} \left( 180, \frac{10^2}{8} \right)$$

- Grenze des Verwerfungsbereichs:

```
qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8))
[1] 174.1846
```

- *Verwerfungsbereich*

Der Verwerfungsbereich ist also

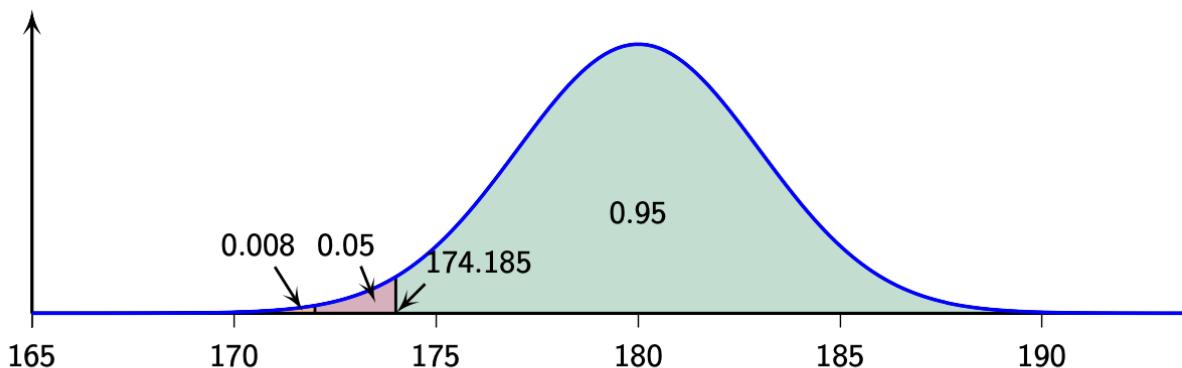
$$K = (-\infty, 174.185)$$

- Verwerfungsbereich ist natürlich viel zu gross, da wohl kaum Körpergrößen von erwachsenen Frauen unter 50 cm zu erwarten sind
- Arbeiten hier mit einem *Modell*, das eben nur in einem bestimmten Bereich Sinn macht
- *Testentscheid*  
Wert im Verwerfungsbereich und somit wird Nullhypothese *verworfen*, dass das wahre  $\mu = 180$  gilt
- Mittelwert der zufällig ausgewählten acht Frauen erscheint immer noch relativ hoch, aber er reicht schon, damit an der Annahme  $\mu = 180$  gezweifelt wird
- Wert für  $P(\bar{X}_6 < 171.54)$ :

$$P(\bar{X}_6 < 171.54) = 0.008$$

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))
[1] 0.008359052
```

- Abbildung:



- Dieser Wert heisst  $p$ -Wert und gibt die Sicherheit mit der man Testentscheid trifft
- Wird die Nullhypothese verworfen, so deutet ein sehr kleiner  $p$ -Wert darauf hin, dass die Nullhypothese sicherer verworfen wird, als wenn er in der Nähe des Signifikanzniveaus (hier  $\alpha = 0.05$ ) liegt.

### Einfluss der Anzahl Messungen auf Verwerfungsbereich

- Beispiel Waage von früher
- Messreihen verschiedener Länge  $n$ , alle mit geschätztem Mittelwert  $\hat{\mu} = 79.78$
- Bestimmen für alle Messreihen den Wert

$$P(\bar{X}_n \leq 79.98) \quad \text{mit} \quad \bar{X}_n \sim \mathcal{N}\left(80, \frac{0.02^2}{n}\right)$$

- Ist dieser Wert grösser als 0.025, dann wird die Nullhypothese nicht verworfen, ansonsten schon
- $n = 2$ :

$$P(\bar{X}_2 \leq 79.98) = 0.079 > 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(2))
[1] 0.0786496
```

- Die Nullhypothese wird also nicht verworfen
- Bei 2 Messwerten Abweichung vom wahren Mittelwert als zufällig möglich erachtet

- $n = 4$ :

$$P(\bar{X}_4 \leq 79.98) = 0.022 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(4))
[1] 0.02275013
```

- Hier wird die Nullhypothese (knapp) verworfen

- $n = 6$ :

$$P(\bar{X}_6 \leq 79.98) = 0.007 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))
[1] 0.007152939
```

- Nullhypothese wird klarer verworfen als für  $n = 4$
- Und schlussendlich noch für  $n = 8$ :

$$P(\bar{X}_8 \leq 79.98) = 0.002 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(8))
[1] 0.002338867
```

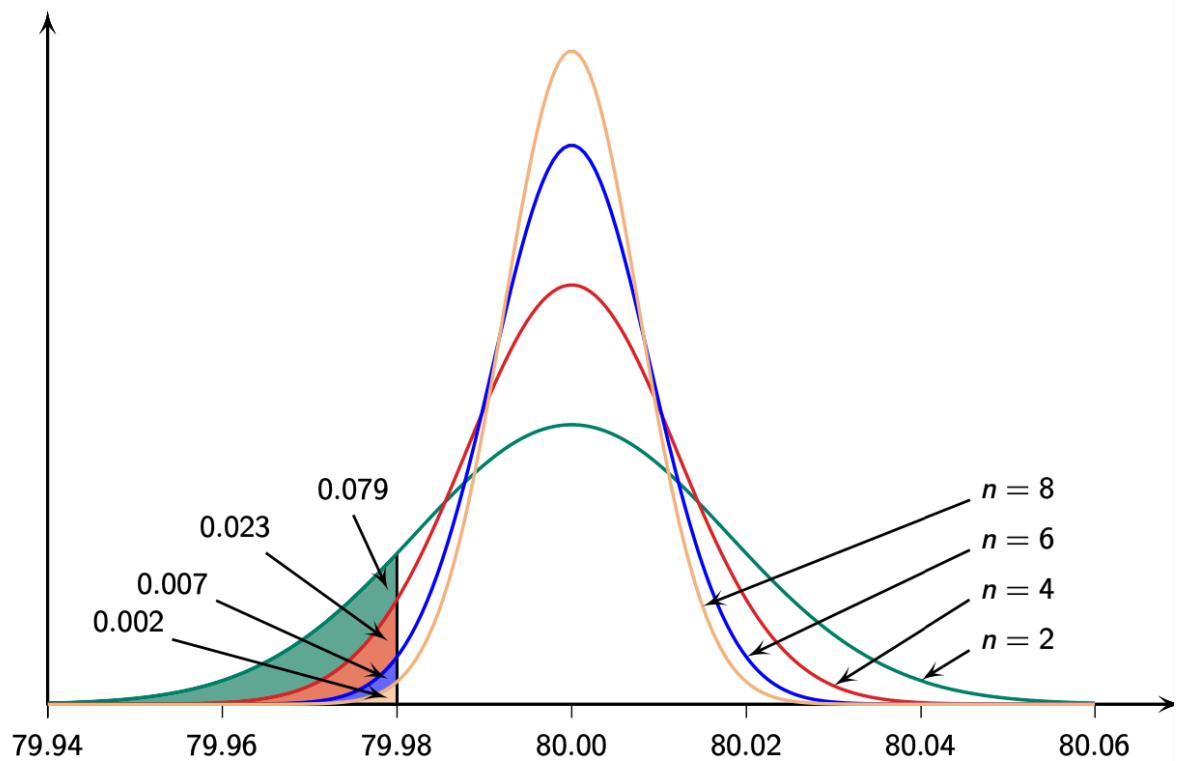
- Die Nullhypothese wird noch klarer verworfen, als bei  $n = 6$
- Mit zunehmendem  $n$  wird der Wert

$$P(\bar{X}_n \leq 79.98)$$

immer kleiner

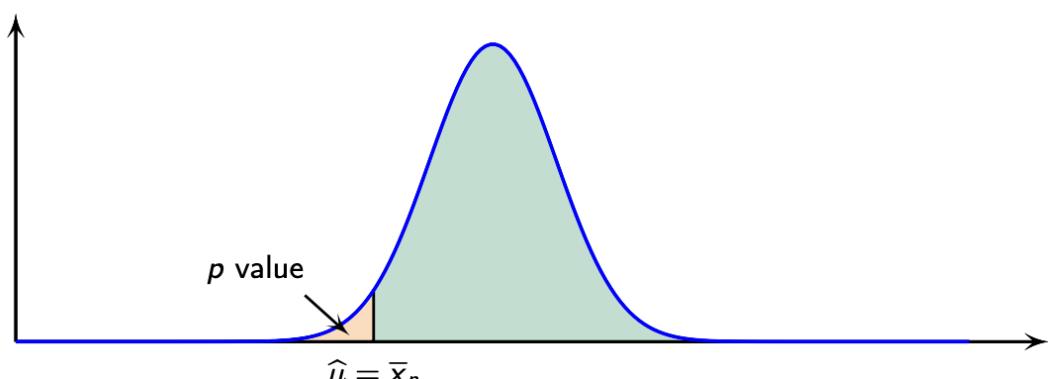
- Grund: Standardabweichung mit grösser werdendem  $n$  kleiner wird
- Normalverteilungskurven werden schmäler
- D.h.: je mehr Messungen wir haben, umso wichtiger ist eine Abweichung von wahren Mittelwert

- Abbildung:



### P-Wert

- $p$ -Wert ist ein Wert zwischen 0 und 1, der angibt, wie gut Nullhypothese und Daten zusammenpassen
  - ▶ 0: passt gar nicht
  - ▶ 1: passt sehr gut
- $p$ -Wert ist die W'keit, unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein *extremeres* zu erhalten



- Mit dem  $p$ -Wert wird also angedeutet, wie extrem das Ergebnis ist
- Je kleiner der  $p$ -Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese
- Werte kleiner als eine im voraus festgesetzte Grenze, wie 5 %, 1 % oder 0.1 % sind Anlass, die Nullhypothese abzulehnen

### **$p$ -Wert**

Der  $P$ -Wert ist die Wahrscheinlichkeit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten wie das aktuell beobachtete.

- Testentscheid auch mit Hilfe des  $p$ -Wertes durchführen

### **$p$ -Wert und Statistischer Test**

Bei einem vorgegebenen Signifikanzniveau  $\alpha$  (z.B.  $\alpha = 0.05$ ) gilt aufgrund der Definition des  $p$ -Werts für einen einseitigen Test:

- ▶ Verwerfe  $H_0$  falls  $p$ -Wert  $\leq \alpha$
- ▶ Belasse  $H_0$  falls  $p$ -Wert  $> \alpha$

- Viele Computer-Pakete liefern den Testentscheid nur mit  $p$ -Wert
- Wie signifikant?

$p$ -Wert  $\approx 0.05$  : schwach signifikant, “.“

$p$ -Wert  $\approx 0.01$  : signifikant, “\*“

$p$ -Wert  $\approx 0.001$  : stark signifikant, “\*\*“

$p$ -Wert  $\leq 10^{-4}$  : äusserst signifikant, “\*\*\*“

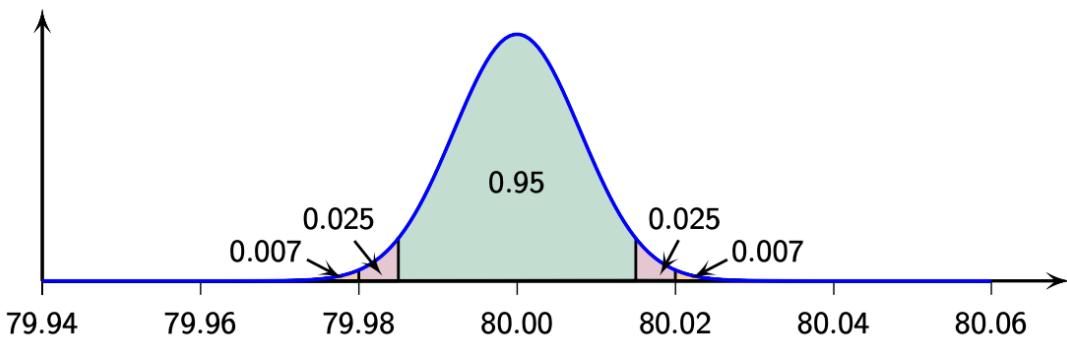
## P-Wert für zweiseitigen Test

- Haben den  $p$ -Wert für einseitige Tests definiert
- Wie sieht nun aber der  $p$ -Wert für zweiseitige Tests aus?
- Beispiel von früher:

$$P(\bar{X}_6 \leq 79.98) = 0.007$$

der kleiner ist als 0.025

- Könnten dies als  $p$ -Wert betrachten, tun es aber nicht
- Skizze:



- Da aber das Signifikanzniveau auf  $\alpha = 0.05$  liegt, wird die W'keit oben auf 5 % umgerechnet, also verdoppelt:

$$p\text{-Wert} = 2 \cdot P(\bar{X}_6 \leq 79.98) = 0.014$$

- Dieser  $p$ -Wert dann mit dem Signifikanzniveau vergleichen
- Computersoftware gibt den  $p$ -Wert *immer* auf Signifikanzniveau an

## t-Test

- Bisher: Verfahren heisst z-Test
- Stillschweigend vorausgesetzt: Standardabweichung *bekannt*
- Praxis: Praktisch nie der Fall
- Folgender *t-Test*: Setzt keine Standardabweichung voraus
- Darum: *t-Test* viel wichtiger als z-Test
- Vorgehen sehr ähnlich z-Test → Nur andere Verteilung
- Wie vorher Annahme: Daten Realisierungen von

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2)$$

- Praxis: Annahme, dass  $\sigma_X$  bekannt ist, meist unrealistisch
- Können aber  $\sigma_X$  aus Daten schätzen →  $\hat{\sigma}_X^2$

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

- Zusätzliche Unsicherheit → Verteilung der Teststatistik ändern

### t-Verteilung

Die Verteilung der Teststatistik beim *t-Test* unter der Nullhypothese

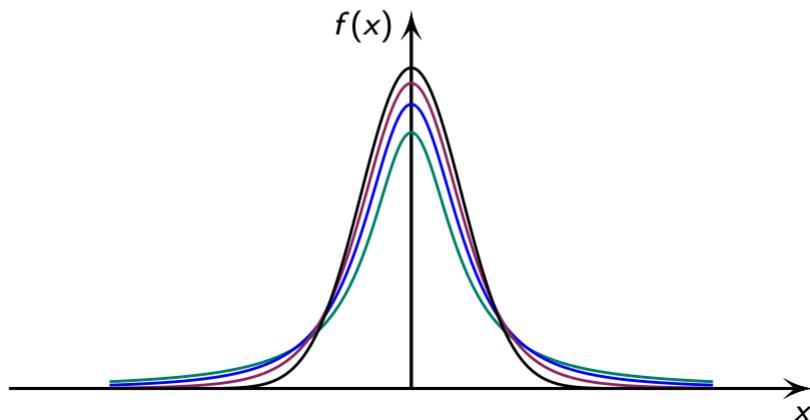
$$H_0 : \mu = \mu_0$$

ist gegeben durch

$$T = \bar{X}_n \sim t_{n-1} \left( \mu, \frac{\hat{\sigma}_X}{\sqrt{n}} \right)$$

wobei  $t_{n-1}$  eine *t-Verteilung* mit  $n - 1$  Freiheitsgraden ist

- Normalverteilung wird also durch eine  $t$ -Verteilung ersetzt
- Was aber ist eine  $t$ -Verteilung?
- Ähnlich Normalverteilung, aber flacher, wegen grösserer Unsicherheit
- Hängt von der Anzahl Beobachtungen
- Skizze für  $\mu = 0$  und  $\sigma \approx 1$  (hängt von  $n$  ab):



- Grün:  $n = 1$ , blau:  $n = 2$ , violet:  $n = 5$ , schwarz:  $\mathcal{N}(0, 1)$
- $t_n$ -Verteilung symmetrische Verteilung um 0, aber langschwänziger ist Standardnormalverteilung  $\mathcal{N}(0, 1)$
- Für grosse  $n$  ist  $t_n$  ähnlich zu  $\mathcal{N}(0, 1)$
- $t_n$  strebt für  $n \rightarrow \infty$  gegen Standardnormalverteilung  $\mathcal{N}(0, 1)$
- Wichtig: Für  $t$ -Test  $t_{n-1}$  verwenden
- $t$ -Verteilung wurde von William Gosset (Chefbrauer Guiness Brauerei) 1908 gefunden

- Alle Begriffe vom z-Test können für den t-Test übernehmen
- Verwerfungsbereich mit t `qt(...)` anstatt `qnorm(...)`
- $p$ -Wert mit `pt(...)` anstatt `pnorm(...)`
- t-Test kommt sehr oft vor: Ganzes Verfahren in R implementiert
- Daten in Befehl `t.test(...)` eingeben und R übernimmt Arbeit
- Verwerfungsbereich nicht *nicht* ausgegeben
- Aber  $p$ -Wert wird ausgegeben, reicht für Testentscheid
- Datensatz aus normalverteilten Datenpunkten  $x_1, \dots, x_{20}$

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 5.9 | 3.4 | 6.6 | 6.3 | 4.2 | 2.0 | 6.0 | 4.8 | 4.2 | 2.1 |
| 8.7 | 4.4 | 5.1 | 2.7 | 8.5 | 5.8 | 4.9 | 5.3 | 5.5 | 7.9 |

- Vermutung:  $x_1, x_2, \dots, x_{20}$  Realisierungen von

$$X_i \sim \mathcal{N}(5, \sigma_X^2)$$

- $\sigma_X$  unbekannt  $\rightarrow$   $\sigma_X$  also aus Daten schätzen

```
x <- c(5.9, 3.4, 6.6, 6.3, 4.2, 2, 6, 4.8, 4.2, 2.1, 8.7, 4.4,
 5.1, 2.7, 8.5, 5.8, 4.9, 5.3, 5.5, 7.9)

mean(x)
[1] 5.215
sd(x)
[1] 1.883802
```

- Nullhypothese lautet in diesem Fall:

$$H_0 : \mu_0 = 5$$

- Test, ob Mittelwert 5.215 zum vermuteten Wert  $\mu_0$  passt oder nicht
- Befehl `t.test(...)`:

```
t.test(x, mu = 5)
##
One Sample t-test
##
data: x
t = 0.51041, df = 19, p-value = 0.6156
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
4.333353 6.096647
sample estimates:
mean of x
5.215
```

- **One Sample t-test**

Es wird ein Einstichprobentest gemacht (Zweistichproben nächste Woche)

- **data: x**  
Datensatz, der verwendet wurde
- **t = 0.51041**
  - ▶ *t*-Wert
  - ▶ Dieser ist an sich uninteressant
  - ▶ „Grosser“ *t*-Wert: Nullhypothese wird verworfen
  - ▶ *t*-Wert „nahe“ bei 0: Nullhypothese wird *nicht* verworfen
  - ▶ Entscheidender ist der *P*-Wert weiter unten
- **df = 19**  
Freiheitsgrad (degree of freedom): Auch uninteressant

- `p-value = 0.6156`
  - ▶  $p$ -Wert
  - ▶ Dies ist *der entscheidende Wert*
  - ▶ Entscheidet, ob die Nullhypothese verworfen wird oder nicht
  - ▶ Hier: Nullhypothese auf Signifikanzniveau 5 % nicht verwerfen, da  $p$ -Wert grösser als 0.05
- `alternative hypothesis: true mean is not equal to 5`  
Hier wird die Alternativhypothese aufgeführt
- `95 percent confidence interval: 4.33 6.09`  
Vertrauensintervall (wird gleich eingeführt)
- `mean of x 5.215`  
Mittelwert von `x`

## Beispiel: Waage A

- Schätzen Standardabweichung  $\sigma_X$  aus den Daten
- Behauptung: Wahres  $\mu = 80$
- $t$ -Test auf dem 5 % Signifikanzniveau
- $t$ -Test

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
 79.97, 80.05, 80.03, 80.02, 80, 80.02)

t.test(waageA, mu = 80)
##
One Sample t-test
##
data: waageA
t = 3.1246, df = 12, p-value = 0.008779
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
80.00629 80.03525
sample estimates:
mean of x
80.02077
```

- $p$ -Wert: 0.009
- Dieser Wert kleiner als Signifikanzniveau 0.05
- Nullhypothese  $H_0$  wird verworfen
- Müssen davon ausgehen, dass der wahre Mittelwert statistisch signifikant *nicht* 80 ist

## Beispiel: Körpergrösse Frauen

- Bundesamtes für Statistik: Durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz ist 180 cm
  - Vermutung: Wert zu gross
  - Auf einem Signifikanzniveau von 5 % untersuchen
  - Wählen zufällig 10 Frauen aus und messen deren Körpergrösse (in cm)
  - Gemessene Grössen:
- 165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9, 170.4, 163.4, 152.5

- Vermutung: Durchschnittliche Körpergrösse *kleiner* als 180 cm

- *t*-Test nach unten:

```

groesse <- c(165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9,
 170.4, 163.4, 152.5)

t.test(groesse, mu = 180, alternative = "less")
##
One Sample t-test
##
data: groesse
t = -5.4836, df = 9, p-value = 0.0001942
alternative hypothesis: true mean is less than 180
95 percent confidence interval:
-Inf 169.382
sample estimates:
mean of x
164.05

```

- *p*-Wert: 0.0002, also weit unter dem Signifikanzniveau von 0.05

- Nullhypothese

$$H_0 : \mu_0 = 180$$

verwerfen

- Alternativhypothese

$$H_A : \mu_0 < 180$$

annehmen

- Aussage des Bundesamtes für Statistik stimmt also statistisch signifikant (sehr wahrscheinlich) *nicht*

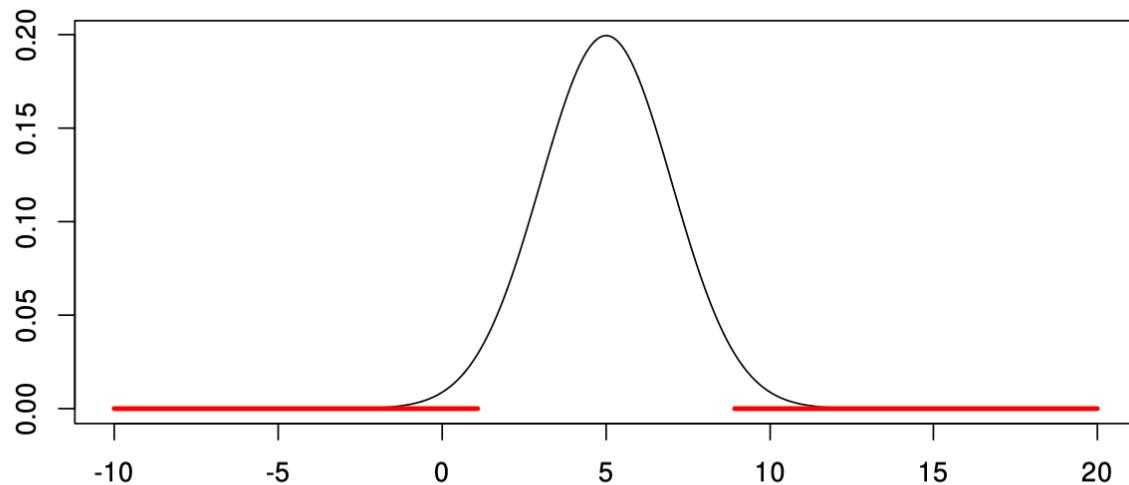
# Vertrauensintervall, Zweistichprobentest, Wilcoxon-Test

## Vertrauenintervall

- Bei Punktschätzung für Mittelwert  $\mu$  einer Messreihe → *Einriger Zahlwert*
- Wissen aber nicht, wie nahe dieser geschätzte Mittelwert beim wahren, aber meist unbekannten, Mittelwert der Verteilung der Messreihe liegt
- Vertrauensintervall: Intervall, das angibt, wo, grob gesagt, der wahre Mittelwert mit einer bestimmten Vorgegebenen W'keit liegt
- Wollen mit Beispiel Vertrauensintervall verschaulichen
- Bestimmung Verwerfungsbereiches beim  $z$ -Test: Gehen vom einem wahren (aber unbekannten) Wert  $\mu$  aus und einer bekannten Standardabweichung aus
- Bestimmen Quartile  $q_{0.025}$  und  $q_{0.975}$  (zweiseitiger Test,  $\alpha = 0.05$ )
- Normalverteilungskurve  $X \sim \mathcal{N}(5, 2^2)$
- $q_{0.025}$ - und  $q_{0.975}$ -Quantile

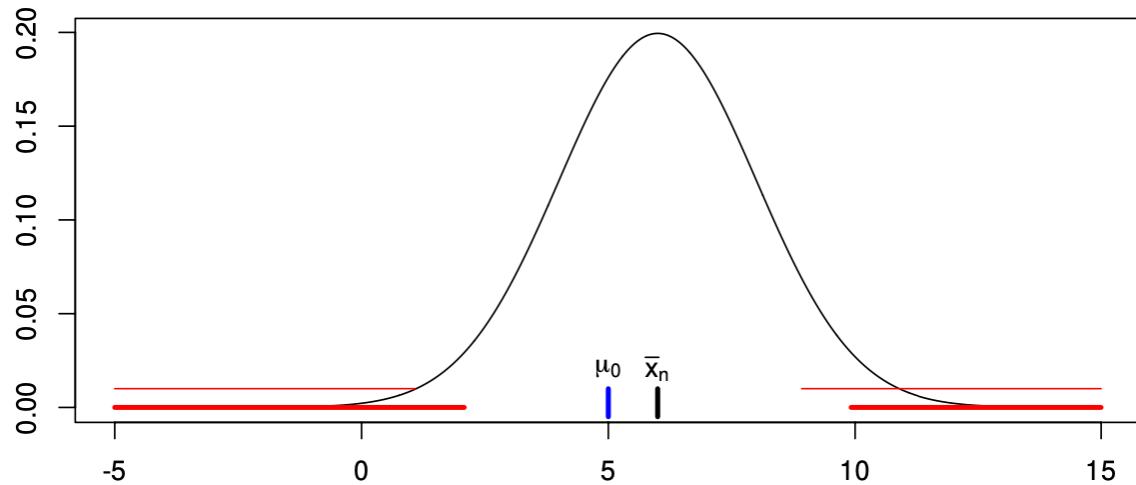
```
qnorm(p = c(0.025, 0.975), mean = 5, sd = 2)
[1] 1.080072 8.919928
```

- Abbildung: Normalverteilungskurve mit dem Verwerfungsbereich rot:



- Liegt nun  $\bar{x}_n$  im Verwerfungsbereich (roter Bereich), dann wird die Nullhypothese  $H_0$  verworfen
- Wahres  $\mu$  praktisch immer unbekannt: Wert einfach angenommen
- Frage umkehren: Kennen  $\bar{x}_n$  und fragen, für welche  $\mu$  wird  $H_0$  nicht verworfen
- Kann man rechnerisch herleiten, machen es aber nur graphisch
- Annahme  $\mu_0 = 5$
- Gegeben  $\bar{x}_n = 6$  und zeichnen Verwerfungsbereich für diesen Wert ein

- Abbildung:



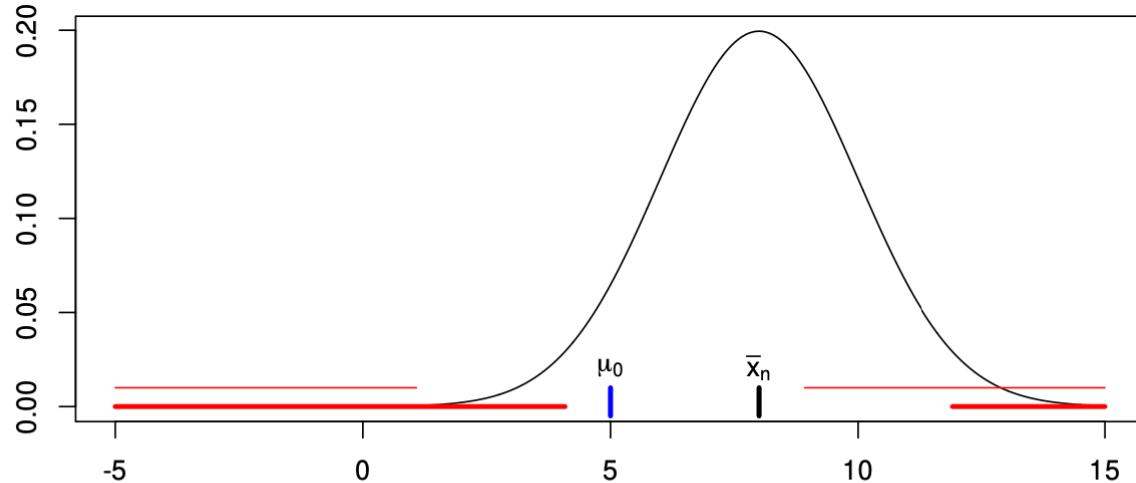
- Linien:

- ▶ Dicke rote Linien: Verwerfungsbereich für  $\bar{x}_n = 6$
- ▶ Dünne rote Linien: Verwerfungsbereich für  $\mu_0 = 5$
- ▶ Vertikaler schwarzer Strich:  $\bar{x}_n = 6$
- ▶ Vertikaler blauer Strich:  $\mu_0 = 5$

- Stellen fest:  $\bar{x}_n$  und  $\mu_0$  nicht innerhalb in einem der beiden Verwerfungsbereiche

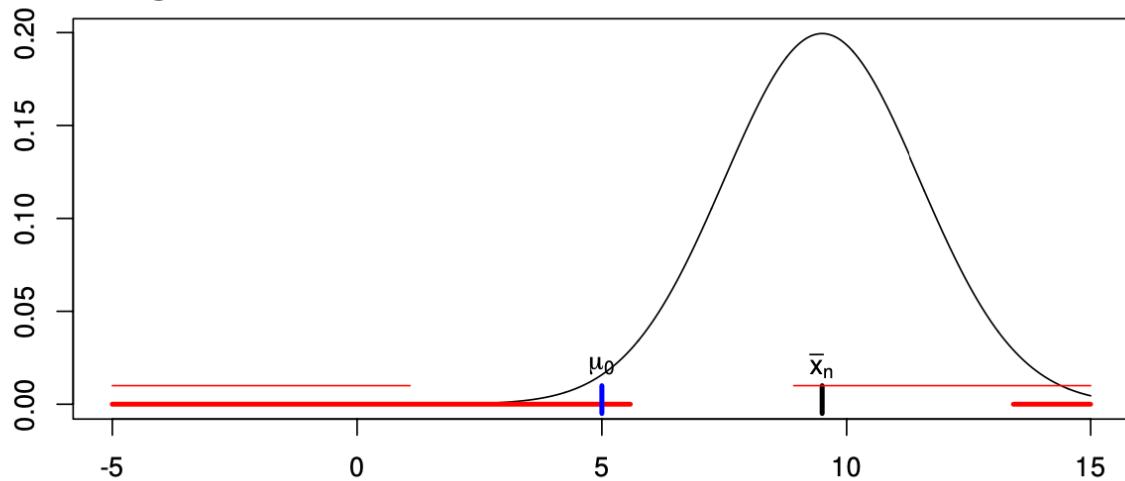
- Idee:  $\bar{x}_n$  vergrössern und  $\mu_0 = 5$  konstant lassen

- Abbildung:  $\bar{x}_n = 8$ :

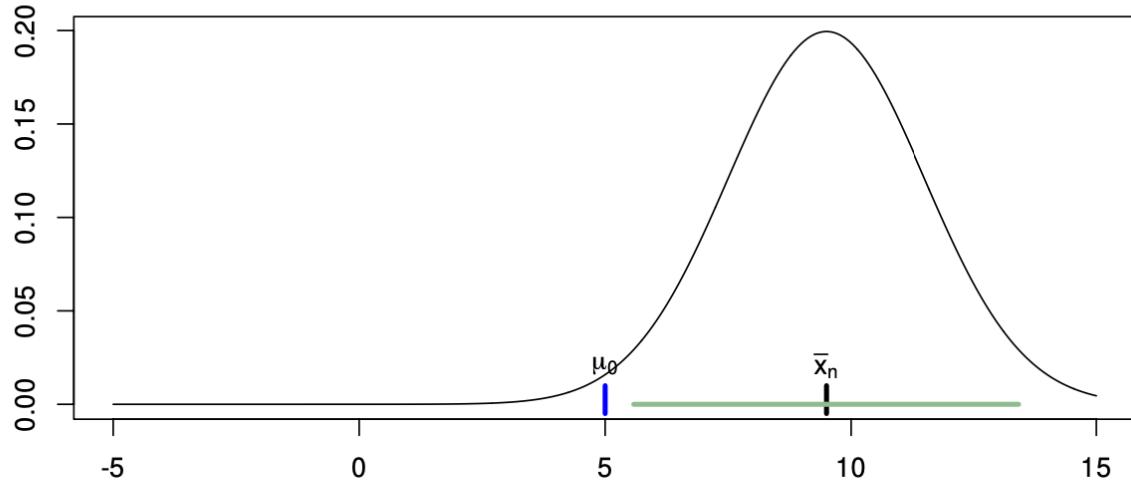


- Auch hier:  $\bar{x}_n$  und  $\mu_0$  nicht innerhalb der beiden Verwerfungsbereiche

- Abbildung:  $\bar{x}_n = 9.5$



- Andere Situation:  $\bar{x}_n$  (schwarze Linie) im Verwerfungsbereich von  $\mu_0 = 5$  (dünne blaue Linien)
- Nullhypothese  $H_0$  nun verworfen
- Aber:  $\mu_0 = 5$  im Verwerfungsbereich für  $\bar{x}_n = 9.5$ .
- Andere Darstellung: Bereich, der *nicht* zum Verwerfungsbereich gehört
- Dieses Intervall heisst *Vertrauensintervall*



- Wert 5 liegt nicht im Vertrauensintervall:

```
qnorm(p = c(0.025, 0.975), mean = 9.5, sd = 2)
[1] 5.580072 13.419928
```

Grün: Der Bereich, in welchem der Wert nicht verworfen wird

## Interpretation Vertrauensintervall

- Annahme: Kennen wahre Verteilung  $\mathcal{N}(5, 2^2)$
- Es gilt also  $\mu = \mu_0 = 5$
- Wählen Zufallszahl aus, die Normalverteilung  $\mathcal{N}(5, 2^2)$  folgt

```
set.seed(1)
m <- rnorm(n = 1, mean = 5, sd = 2)
m
[1] 3.747092
```

- Bestimmen Vertrauensintervall von 3.7470924:

```
qnorm(p = c(0.025, 0.975), mean = m, sd = 2)
[1] -0.1728356 7.6670203
```

- Feststellung  $\mu_0 = 5$  liegt im Vertrauensintervall
- Wählen andere Zufallszahl aus, die Normalverteilung  $\mathcal{N}(5, 2^2)$  folgt

```
set.seed(7)
m <- rnorm(n = 1, mean = 5, sd = 2)
m
[1] 9.574494
```

- Bestimmen Vertrauensintervall von 9.5744943:

```
qnorm(p = c(0.025, 0.975), mean = m, sd = 2)
[1] 5.654566 13.494422
```

- Feststellung  $\mu_0 = 5$  liegt nicht im Vertrauensintervall
- Frage: In wievielen Fällen liegt  $\mu_0 = 5$  im Vertrauensintervall einer zufällig ausgewählten Zahl, die  $\mathcal{N}(5, 2^2)$  folgt?

- Folgender Code: Bestimmt 1000 mal Zufallszahl und zählt wieviele Male  $\mu_0 = 5$  im Vertrauensintervall der Zufallszahl liegt:

```
n <- 1000

r <- rnorm(n = n, mean = 5, sd = 2)
q_u <- qnorm(p = c(0.025), mean = r, sd = 2)
q_o <- qnorm(p = c(0.975), mean = r, sd = 2)

k <- 0

for (i in 1:n) {
 if ((q_u[i] <= 5 & 5 <= q_o[i]) == FALSE) {
 k <- k + 1
 }
}

print(k)
[1] 47
```

- In 47 von 1000 Fällen liegt  $\mu_0 = 5$  nicht im Vertrauensintervall der Zufallszahlen

Es müsste =5 sein.

- Folgender Code: Macht das 20 mal

```
vi2 <- function(n) {
 r <- rnorm(n = n, mean = 5, sd = 2)
 q_u <- qnorm(p = c(0.025), mean = r, sd = 2)
 q_o <- qnorm(p = c(0.975), mean = r, sd = 2)

 k <- 0

 for (i in 1:n) {
 if ((q_u[i] <= 5 & 5 <= q_o[i]) == FALSE) {
 k <- k + 1
 }
 }
 cat(k, " ")
}

for (i in 1:20) {
 vi2(1000)
}

52 52 50 58 46 55 60 53 48 47 56 48 52 50 66 61 47 51 57 42
```

- In etwa 50 von 1000 Fällen liegt  $\mu_0 = 5$  nicht im Vertrauensintervall der Zufallszahlen
- Das sind etwa 5 %

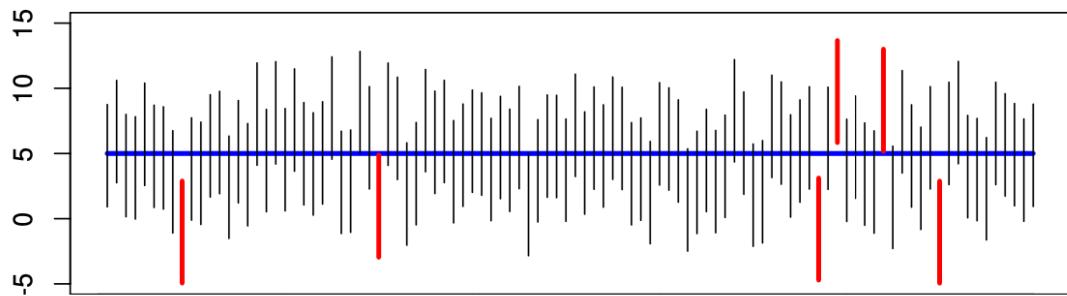
## Graphisch

- Rote Linien zeigen an, wo  $\mu_0 = 5$  nicht im Vertrauensintervall enthalten ist:

```
r <- rnorm(n = 100, mean = 5, sd = 2)
q_u <- qnorm(p = c(0.025), mean = r, sd = 2)
q_o <- qnorm(p = c(0.975), mean = r, sd = 2)

plot(NULL, xlim = c(1, 100), ylim = c(-5, 15), xlab = "", ylab = "")

lines(c(1, 100), c(5, 5), lwd = 3, col = "blue")
for (i in 1:100) {
 lines(c(i, i), c(q_u[i], q_o[i]))
 if ((q_u[i] <= 5 & 5 <= q_o[i]) == FALSE) {
 lines(c(i, i), c(q_u[i], q_o[i]), col = "red", lwd = 3)
 }
}
```



- Hier 100 Vertrauensintervalle
- In 6 Fällen ist  $\mu_0 = 5$  nicht im Vertrauensintervall
- Führen dies oft durch: In etwa 5 % der Fälle liegt  $\mu_0 = 5$  nicht im Vertrauensintervall
- Interpretation: Zu 95 % liegt der wahre Mittelwert im Vertrauensintervall
- Man spricht von einem 95 %-Vertrauensintervall

- Gesehen: Nullhypothese wird verworfen
- Fällt wahres  $\mu$  also aus dem Vertrauensintervall, dann wird Nullhypothese verworfen
- Weiteren Interpretation des Vertrauensintervalls: Enthält alle  $\mu_0$ 's für die Nullhypothese *nicht* verworfen wird
- Es sagt uns also in welchem Intervall sich das wahre  $\mu_0$  befindet
- Gilt nicht absolut: Mit einer bestimmten W'keit liegt wahres  $\mu_0$  in diesem Intervall
- Hier: Wahres  $\mu_0$  liegt zu 95 % im Vertrauensintervall
- Sprechen deswegen auch von einem 95 %-Vertrauensintervall
- Weitere Möglichkeit für Testentscheid:
  - ▶ Liegt  $\mu_0$  der Nullhypothese im Vertrauensintervall, so wird die Nullhypothese *nicht* verworfen
  - ▶ Liegt  $\mu_0$  der Nullhypothese *nicht* im Vertrauensintervall, so wird die Nullhypothese verworfen
- R-Output: Gibt Vertrauensintervall (`confidence interval`) an
- Dieses besagt, dass bei einem Signifikanzniveau von 5 % das wahre  $\mu$  zu 95 % in diesem Intervall liegt
- Mit Vertrauensintervall kann man ebenfalls Testentscheid durchführen

## Beispiel: Waage A

- Nullhypothese

$$H_0 : \mu_0 = 80$$

- R-Output: Vertrauensintervall:

$$[80.00629, 80.03525]$$

- Zu 95 % liegt das wahre  $\mu$  in diesem Intervall
- Aber  $\mu_0 = 80$  *nicht* in diesem Intervall
- Zu 95 % Sicherheit ist das wahre  $\mu$  *nicht* 80
- Nullhypothese wird verworfen und Alternativhypothese angenommen

## Beispiel: Körpergrösse Frauen

- Nullhypothese:

$$H_0 : \mu_0 = 180$$

- R-Output: Vertrauensintervall:

$$(-\infty, 169.382]$$

- Zu 95 % liegt das wahre  $\mu$  in diesem Intervall
- $\mu_0 = 180$  *nicht* in diesem Intervall
- Mit 95 % Sicherheit ist das wahre  $\mu$  *nicht* 80
- Nullhypothese verwerfen; Alternativhypothese annehmen

## Bemerkungen

- Je schmäler das Vertrauensintervall ist, umso sicherer weiss man, wo sich der wahre Mittelwert befindet
- Ist das Vertrauensintervall schmal, wie

$$[105.12, 105.23]$$

so wissen wir sehr genau, wo der wahre Mittelwert mit 95 % Wahrscheinlichkeit liegt

- Ist das Vertrauensintervall breit, wie

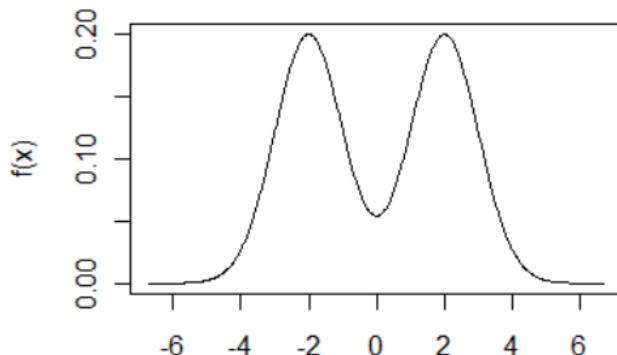
$$[10, 1000]$$

so besteht grosse Unsicherheit, wo das wahre  $\mu$  liegt

## Nicht-Normalverteilte Daten: Wilcoxon-Test

- Alternative zu  $t$ -Test
- Wilcoxon-Test ist der weniger voraussetzt als der  $t$ -Test
- Voraussetzung: Verteilung unter Nullhypothese *symmetrisch* bez.  $\mu_0$
- Annahme:

$$X_i \sim F \text{ iid, } F \text{ ist symmetrisch}$$



- Es wird ein *V*-Wert (*Rangsumme*) berechnet
- Grundidee diesselbe wie bei Hypothesentest bisher:
  - ▶ *V*-Wert „weit“ weg vom *Median*: Nullhypothese verwerfen
  - ▶ *V*-Wert „nahe“ beim *Median*: Nullhypothese nicht verwerfen
  - ▶ R berechnet *p*-Wert

### Beispiel: Waage A

- R Output:

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
 80.05, 80.03, 80.02, 80, 80.02)

wilcox.test(x, mu = 80, alternative = "two.sided")
##
Wilcoxon signed rank test with continuity correction
##
data: x
V = 69, p-value = 0.0195
alternative hypothesis: true location is not equal to 80
```

- Auf Signifikanzniveau von 5 % wird Nullhypothese verworfen, da *p*-Wert kleiner als 0.05 ist

### Wilcoxon-Test vs. t-Test

#### Wilcoxon-Test versus t-Test

DWilcoxon-Test ist in den allermeisten Fällen dem *t*-Test vorzuziehen: Er hat in vielen Situationen oftmals wesentlich grössere Macht (Wahrscheinlichkeit Nullhypothese richtigerweise zu verwerfen)

Selbst in den ungünstigsten Fällen ist er nie viel schlechter

## Vergleich von zwei Stichproben: Mögliche Fragestellungen

- Vergleich von zwei Messverfahren (Messgerät A vs. Messgerät B): Gibt es einen signifikanten Unterschied?
- Vergleich von zwei Herstellungsverfahren (A vs. B): Welches hat die besseren Eigenschaften (z.B. bzgl. Brüchigkeit von Handy-Displays)?
- Werden männliche Dozenten von weiblichen Studierenden besser als von männlichen Studierenden bewertet?
- Sammeln also jeweils Daten von *zwei* Gruppen

## Gepaarte Stichproben

- Beispiel Messgeräte: Jeder Prüfkörper wird mit *beiden* Messgeräten gemessen
- Pro *Versuchseinheit* (hier: Prüfkörper) zwei Beobachtungen (einmal Gerät A und einmal Gerät B)
- Man spricht auch von *gepaarten Stichproben*
- Beide Beobachtungen sind *nicht* unabhängig, da an *gleicher* Versuchseinheit zwei Mal gemessen wird!

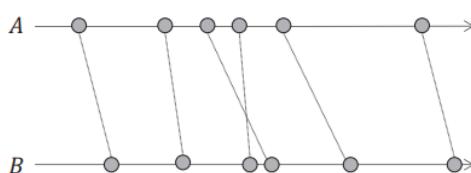
## Ungepaarte (unabhängige) Stichproben

- Beispiel Herstellungsverfahren: Stichprobe von Verfahren A und eine andere Stichprobe von Verfahren B und messen jedes Objekt aus
- Beobachtungen sind hier *unabhängig*: „Es gibt *nichts*, was sie verbindet“
- Man spricht auch von *ungepaarten (oder unabhängigen) Stichproben*

## Unterscheidung gepaarte vs. ungepaarte Stichproben

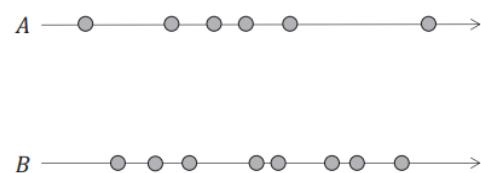
### Gepaarte Stichproben

- Jede Beobachtung einer Gruppe kann eindeutig einer Beobachtung der anderen Gruppe zugeordnet werden
- Stichprobengröße ist in beiden Gruppen zwangsläufig gleich



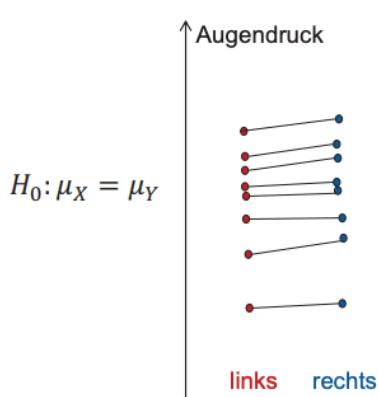
### Ungepaarte Stichproben

- Keine Zuordnung von Beobachtungen möglich
- Stichprobengrößen können verschieden sein (müssen aber nicht!)
- Man kann die eine Gruppe vergrößern, ohne dass man die andere vergrößert



## Gepaarte versus ungepaarte Stichproben

- Beispiel: Augeninnendruck; ein Auge behandelt, das andere nicht (gepaarter Test ist angebracht)
- Gemäss Voraussetzungen dürfte auch ein ungepaarter Test angewendet werden



Ungepaart:

$$\text{Intuition Teststatistik: } T = \frac{\bar{X} - \bar{Y}}{\widehat{\sigma}_{\bar{X}}}$$

Gepaart:

$$\text{Differenz } D_i = X_i - Y_i$$

$$\text{Teststatistik } T = \frac{\bar{D}}{\widehat{\sigma}_{\bar{D}}}$$

## Statistischer t-Test für gepaarte Stichproben mit

- *Gepaarte Stichproben:* Normalverteilte Daten

$$X_i \sim \mathcal{N}(\mu_X, \sigma_X^2) \quad \text{und} \quad Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

- Betrachten Differenzen

$$D_i = X_i - Y_i$$

- Führen einen *t*-Test durch

- Normalerweise für die Nullhypothese

$$\mathbb{E}(D) = \mu_D = 0$$

- *Kein Unterschied!*

- Falls Daten nicht normalverteilt → Wilcoxontest

- R-Output:

```
vorher <- c(25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28)
nachher <- c(27, 29, 37, 56, 46, 82, 57, 80, 61, 59, 43)

t.test(nachher, vorher, alternative = "two.sided", mu = 0, paired = TRUE,
 conf.level = 0.95)

##
Paired t-test
##
data: nachher and vorher
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
4.91431 15.63114
sample estimates:
mean of the differences
10.27273
```

- Nullhypothese wird auf Signifikanzniveau von 5 % verworfen, da *p*-Wert 0.001633 kleiner als 0.05

- Unterschied ist also auf dem 5 % Signifikanzniveau signifikant, weil der P-Wert kleiner als 5 % ist
- 95 %-Vertrauensintervall: Mittelwert der Unterschiede

$$[4.91431, 15.63114]$$

- Mit 95 % W'keit ist der Durchschnitt der Differenzen von **nachher** und **vorher** in diesem Bereich

### Statistischer t-Test für ungepaarte Stichproben

- *Ungepaarte Stichproben:* Daten  $X_i$  und  $Y_j$  normalverteilt, aber ungepaart
- Beispiel: Waage A und B
- Zwei-Stichproben  $t$ -Test für ungepaarte Stichproben mit Nullhypothese

$$\mu_X = \mu_Y$$

- **R-Output:**

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
 80.05, 80.03, 80.02, 80, 80.02)
y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

t.test(x, y, alternative = "two.sided", mu = 0, paired = FALSE,
 conf.level = 0.95)

##
Welch Two Sample t-test
##
data: x and y
t = 2.8399, df = 9.3725, p-value = 0.01866
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.008490037 0.073048425
sample estimates:
mean of x mean of y
80.02077 79.98000
```

- Auf Signifikanzniveau 5 % wird Nullhypothese verworfen, da  $p$ -Wert 0.01866 kleiner als 0.05 ist

- Unterschied ist also auf dem 5 % Signifikanzniveau signifikant, weil der P-Wert kleiner als 5 %
- 95 %-Vertrauensintervall: Unterschied in Gruppenmittelwerten:

$$[0.0167, 0.0673]$$

- Mit 95 % W'keit ist der Gruppenmittelwert von **x** um eine Zahl in diesem Bereich grösser als der Gruppenmittelwert von **y**

### Mann-Whitney U-Test (aka Wilcoxon Rank-sum Test)

- Falls Daten nicht normalverteilt
- R-Output:

```

x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
 80.05, 80.03, 80.02, 80, 80.02)

y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

wilcox.test(x, y, alternative = "two.sided", mu = 0, paired = FALSE,
 conf.level = 0.95)

##
Wilcoxon rank sum test with continuity correction
##
data: x and y
W = 76.5, p-value = 0.01454
alternative hypothesis: true location shift is not equal to 0

```

## Lineare Regression

- Fortsetzung von Block 3: Jetzt mit Hypothesentest
- Lineare Regression ist einer der Startpunkte in Machine Learning

- Auftrag als Statistiker einer Firma: Analyse, Strategie auszuarbeiten, wie Verkauf eines bestimmten Produktes gesteigert werden kann
- Firma stellt Daten von Werbebudget und Verkauf zur Verfügung
- Datensatz **Werbung** besteht aus:
  - ▶ Dem **Verkauf** dieses Produktes in 200 verschiedenen Märkten und den Werbebudgets für dieses Produkt in diesen Märkten
  - ▶ Werbebudget für die drei verschiedenen Medien **TV**, **Radio** und **Zeitung**

- Code:

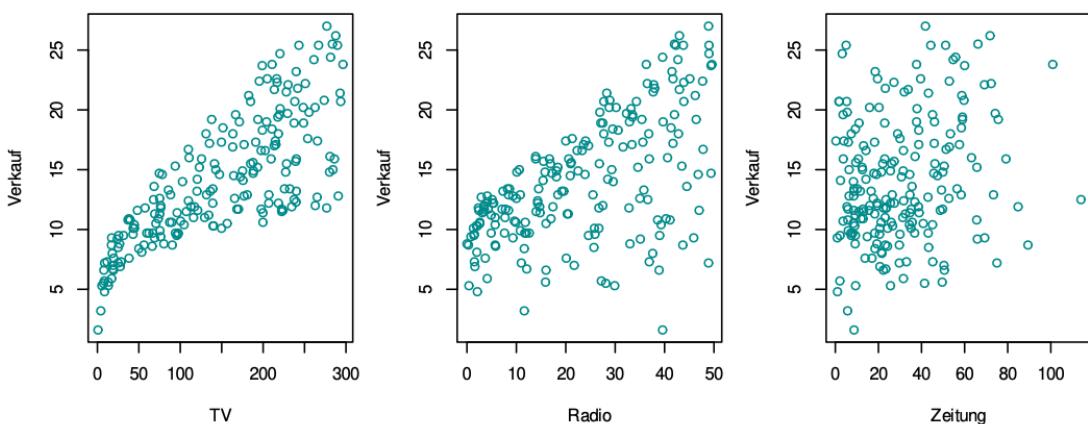
```
Werbung <- read.csv("../Data/Werbung.csv")[, -1]
head(Werbung, 3)

TV Radio Zeitung Verkauf
1 230.1 37.8 69.2 22.1
2 44.5 39.3 45.1 10.4
3 17.2 45.9 69.3 9.3
```

- Daten in Streudiagrammen in Abbildung dargestellt:

```
TV <- Werbung[, 1]
Radio <- Werbung[, 2]
Zeitung <- Werbung[, 3]
Verkauf <- Werbung[, 4]

plot(Verkauf ~ TV, col = "darkcyan", xlab = "TV", ylab = "Verkauf")
plot(Verkauf ~ Radio, col = "darkcyan", xlab = "Radio", ylab = "Verkauf")
plot(Verkauf ~ Zeitung, col = "darkcyan", xlab = "Zeitung", ylab = "Verkauf")
```



- Für Firma nicht möglich, Verkauf des Produktes direkt zu erhöhen
- Aber sie kann Werbeausgaben in den drei Medien kontrollieren
- Ziel: Zusammenhang zwischen Werbung und Verkauf herstellen, damit Firma ihre Werbebudgets anpassen kann, damit sie den Verkauf indirekt erhöhen kann
- Ziel: Möglichst genaues *Modell* zu entwickeln, damit auf Basis der drei Medienbudgets der Verkauf des Produkts *vorhersagt* werden kann
- Abbildung oben links: Deutlicher Zusammenhang zwischen dem Werbebudget und dem Verkauf des Produktes
- Je mehr in Werbung investiert wird, desto grösser Verkaufszahlen
- Frage: Welche *Form* dieser Zusammenhang?
  - Möglichkeit: Datenpunkte folgen einer Gerade siehe später
  - Abbildung oben rechts: Überhaupt keinen Zusammenhang
  - Folglich kann man die Zeitungswerbung hier sein lassen

- Mathematische Sichtweise: Gesucht Funktion  $f$ , die Werbebudgets  $X_1$  (**TV**),  $X_2$  (**Radio**) und  $X_3$  (**Zeitung**) den Verkauf  $Y$  ermittelt:

$$Y \approx f(X_1, X_2, X_3)$$

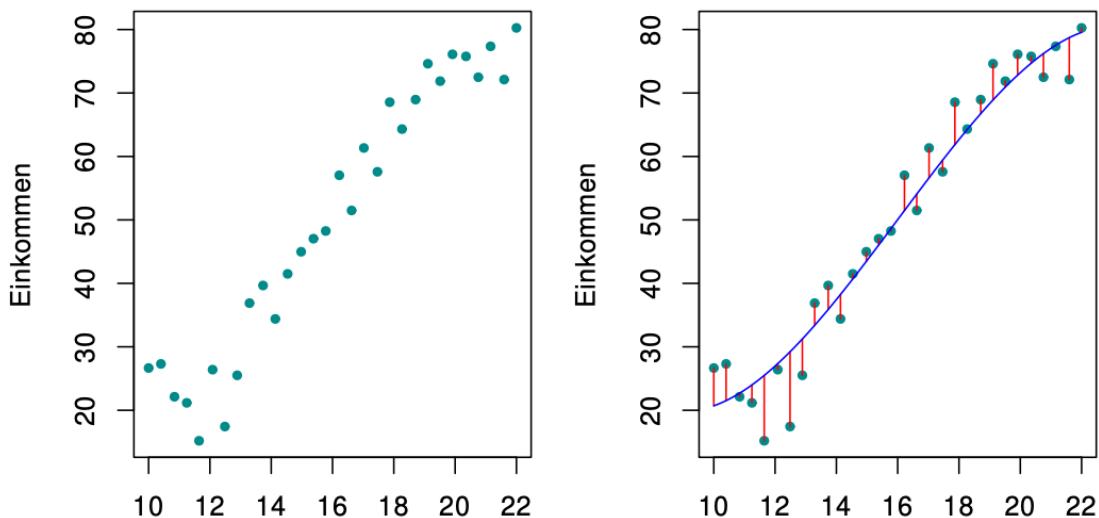
- Beziehung oben: Kein Gleichheitszeichen, da Streudiagramme keine Graphen einer Funktion darstellen
- Funktion  $f$  kann Zusammenhang zwischen  $X_1, X_2, X_3$  und  $Y$  nur *approximativ* darstellen
- Bezeichnung:
  - ▶ Variable  $Y$ : *Zielgröße, Outputvariable*,
  - ▶  $X_1, X_2$  und  $X_3$ : *Prädiktoren, erklärende Variable*
- Allgemein: Quantitative Zielgröße  $Y$  und  $p$  verschiedene Prädiktoren  $X_1, X_2, \dots, X_p$
- Annahme: Es besteht *irgendein* Zusammenhang zwischen  $Y$  und  $X_1, X_2, \dots, X_p$
- Allgemeine Form:

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$

- $f$  irgendeine feste, aber *unbekannte* Funktion von  $X_1, X_2, \dots, X_p$
- Größe  $\varepsilon$ : *Zufälliger Fehlerterm* unabhängig von  $X_1, X_2, \dots, X_p$  mit Mittelwert 0
- Bedeutung Fehlerterm  $\varepsilon$  → Folgendes Beispiel

### Beispiel: Einkommen

- Abbildung links: **Einkommen** von 30 Individuen in Abhängigkeit der **Ausbildungsdauer** (in Jahren)
- Graphik deutet an: **Einkommen** kann aus **Ausbildungsdauer** berechnet werden



- Aber: Funktion  $f$ , die Prädiktoren und Zielgröße miteinander in Verbindung bringt, in der Regel unbekannt
- In dieser Situation:  $f$  aus den Daten *schätzen*
- Datensatz simuliert: Funktion  $f$  bekannt (blaue Kurve) in Abb. rechts
- Einige Beobachtungen liegen überhalb, andere unterhalb der blauen Kurve
- Die roten vertikalen Linien repräsentieren den Fehlerterm  $\varepsilon$
- Insgesamt haben Fehler einen empirischen Mittelwert annähernd 0
- Ziel der Regression: Funktion  $f$  zu *schätzen*

- Schätzen in der Stochastik: Berechnung
- Schätzung ist Annäherung (Approximation) an wahre Grösse
- Geschätzte Grösse wird mit Hut  $\hat{\cdot}$  gekennzeichnet
- $\hat{Y}$ : Schätzung der unbekannten Grösse  $Y$
- $\hat{f}$ : Schätzung der unbekannten Funktion  $f$

### Warum soll f geschätzt werden?

- Hauptgründe, warum man unbekannte Funktion  $f$  schätzen will:
  - ▶ Datenpunkte vorherzusagen (Prognose)
  - ▶ Rückschlüsse auf Funktion selbst zu ziehen
- Prognose: Oft Prädiktoren  $X_1, X_2, \dots, X_p$  einfach verfügbar, aber die Zielgrösse nicht
- In so einem Fall:  $Y$  schätzen durch

$$\hat{Y} = \hat{f}(X_1, X_2, \dots, X_p)$$

- Fehlerterm im Mittel 0

### Beispiel

- Prädiktoren  $X_1, X_2, \dots, X_p$  seien die Werte von verschiedenen Charakteristiken einer Blutentnahme, die der Hausarzt des Patienten in seinem Labor bestimmen kann
- Zielgrösse  $Y$ : Mass für Risiko, dass der Patient starke Nebenwirkungen bei der Anwendung eines bestimmten Medikamentes erleidet
- Arzt möchte bei Verschreibung eines Medikamentes  $Y$  aufgrund von  $X_1, X_2, \dots, X_p$  vorhersagen können, damit er nicht ein Medikament Patienten verschreibt, die ein hohes Risiko für Nebenwirkungen bei diesem Medikament haben - d.h. bei denen  $Y$  gross ist

- Genauigkeit von  $\hat{Y}$  als Vorhersage von  $Y$  hängt von zwei Größen ab:
  - ▶ Reduzibler Fehler
  - ▶ Irreduzibler Fehler

• Allgemein:  $\hat{f}$  keine perfekte Schätzung von  $f$  und diese Ungenauigkeit führt zu einem Fehler

- *Reduzibler Fehler*: Schätzung mit statistischen Methoden verbessern
- Aber auch für perfekte Schätzung von  $f$ : Outputvariable hat Form

$$\hat{Y} = f(X_1, X_2, \dots, X_p)$$

- Vorhersage  $Y$  enthält immer noch Fehler
- Liegt am Fehlerterm  $\varepsilon$ : Hängt nicht von  $X_1, X_2, \dots, X_p$  ab
- Variabilität von  $\varepsilon$  beeinflusst die Genauigkeit der Vorhersage
- *Irreduzibler Fehler*: Fehler kann nicht beeinflusst werden, wie gut auch die Schätzung von  $f$  ist
- Woher kommt nun dieser Fehler  $\varepsilon$ , der grösser als Null ist?
- Größe kann Variablen enthalten, die nicht gemessen wurden, die aber für die Vorhersage von  $Y$  wichtig sind
- Da diese Variablen nicht gemessen wurden → Für die Vorhersage auch nicht verwendbar
- Größe  $\varepsilon$  kann aber auch nicht messbare Größen enthalten
- Bsp: Stärke der Nebenwirkungen eines Medikamentes abhängig sein von der Tageszeit der Einnahme des Medikamentes oder auch einfach vom allgemeinen Wohlbefinden des Patienten

## Rückschlüsse auf $f$ : Fragestellungen

- Welche Inputvariablen werden mit dem Output assoziiert?
  - ▶ Natürlich alle, denkt man zuerst
  - ▶ Aber oft sind es einige wenige Variablen, die auf  $Y$  einen substantiellen Einfluss haben
  - ▶ Sehr viele Inputvariablen: Wichtige Inputvariablen identifizieren
  - ▶ Beispiel **Werbung**:
    - ★ Ausgaben bei TV-Werbung grosser Einfluss auf die Verkaufszahlen
    - ★ Zeitungswerbung aber nicht
    - ★ Auf die TV-Werbung konzentrieren
- Wie sieht der Zusammenhang zwischen Outputvariable und jeder Inputvariable aus?
  - ▶ Einige Prädiktoren haben einen positiven Zusammenhang mit der Outputvariable
  - ▶ Vergrösserung der Inputvariable hat in diesem Fall Vergrösserung von  $Y$  zur Folge
  - ▶ Andere Inputvariablen haben einen negativen Zusammenhang mit  $Y$
  - ▶ In Abhängigkeit von der Komplexität von  $f$  kann der Zusammenhang zwischen der Zielvariablen und einer erklärenden auch von den Werten der anderen erklärenden Variablen abhängen (Interaktion)

- Kann der Zusammenhang zwischen der Outputvariable und jeder Inputvariable durch eine lineare Gleichung angemessen beschrieben werden oder ist der Zusammenhang komplizierter?
  - ▶ Historisch sind die meisten Schätzungen von  $f$  linear
  - ▶ Dies hat damit zu tun, dass solche Schätzungen sehr einfach sind
  - ▶ In vielen Situationen: Annahme Linearität ausreichend oder gar wünschenswert
  - ▶ Aber oft ist der wahre Zusammenhang komplizierter und das lineare Modell liefert keinen angemessenen Zusammenhang zwischen Input- und Outputvariablen

#### Fragen für Beispiel der Werbung

- Welche Medien tragen zum Verkauf des Produktes bei?
- Welche Medien haben den grössten Einfluss auf den Verkauf?
- Welchen Zuwachs im Verkauf hat eine bestimmte Vergrösserung der TV-Werbung zur Folge?

## Schätzung von $f$ ?

- Mehrere Verfahren um zu  $f$  schätzen

- Hier nur *parametrische Methode*

- Vorgehen:
  - ▶ Annahme über die funktionale Form von  $f$
  - ▶ Einfachste Annahme:  $f$  linear in  $X_1, X_2, \dots, X_p$ :

$$f(X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

- ▶ Nach Wahl des Modells: Verfahren, das die Daten in das Modell *passt*
- ▶ Lineares Modell: Parameter  $\beta_0, \beta_1, \dots, \beta_p$  schätzen
- ▶ Parameter so bestimmen, dass

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

- ▶ Häufigste Methode zur Bestimmung von  $\beta_0, \beta_1, \dots, \beta_p$ : *Methode der kleinsten Quadrate*

## Beispiel

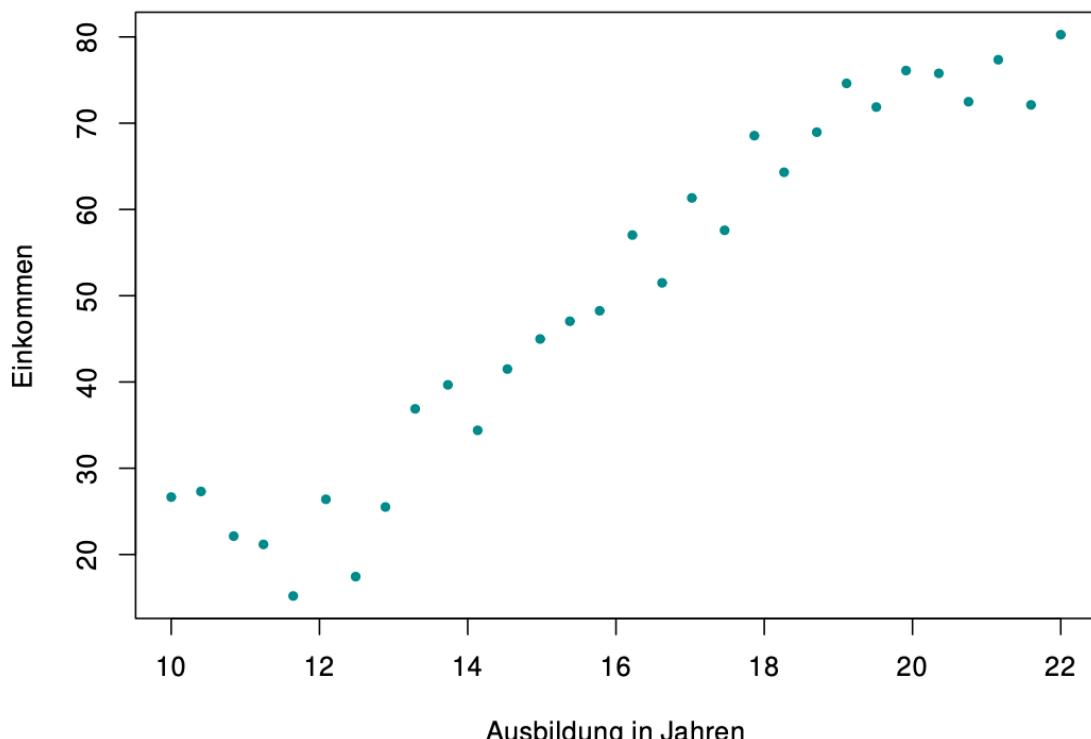
- Beispiel **Werbung**: Lineares Modell:

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung}$$

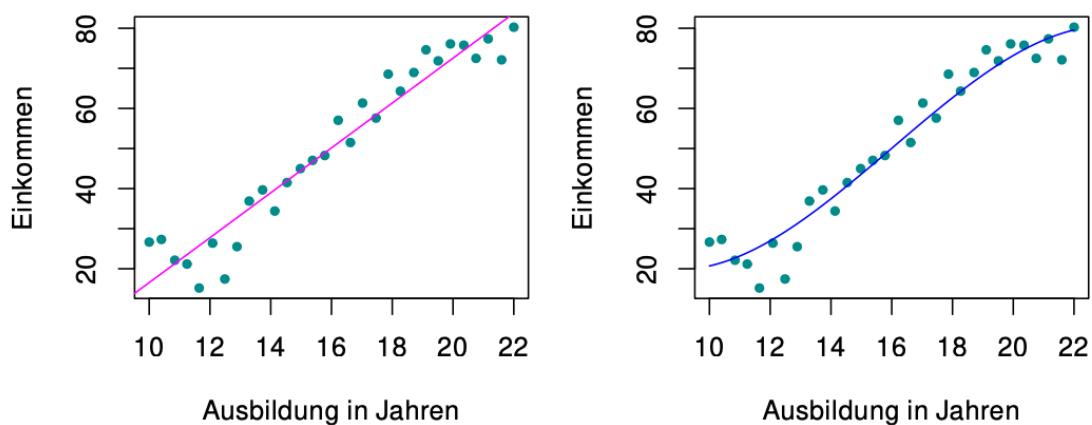
- Beispiel **Einkommen**: Lineares Modell:

$$\text{Einkommen} \approx \beta_0 + \beta_1 \cdot \text{Ausbildung}$$

- Datensatz **Einkommen**:



- Frage: Welches *Modell* wählen, oder welche Form soll  $f$  haben



- Aus Daten: Lineares Modell (oben links):

$$f(X) = \beta_0 + \beta_1 X$$

- Auch kubisches Modell (Polynom 3. Grades) möglich (oben rechts):

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

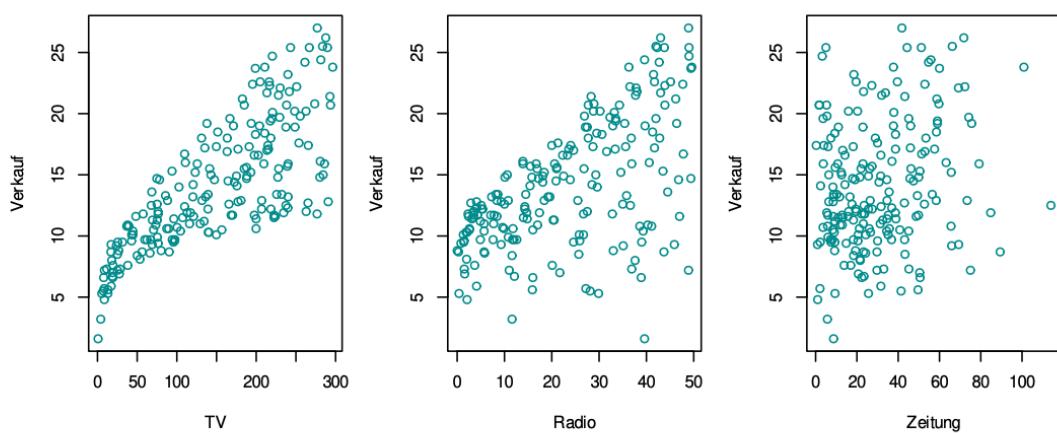
Viele weitere Modelle denkbar Aber welches ist nun das „richtige“? Dies lässt sich in dieser Absolutheit nicht entscheiden Funktion f i. A. unbekannt: Liegt an uns „bestes“ Modell zu wählen Statistik: Bei Entscheidungsfindung behilflich Welches Modell ist in unserem Beispiel das „bessere“? Kubisches Modell scheint besser zu passen, aber

auch komplizierter Lineares Modell einfacher (etwas weniger genau) hat Vorteil: Die Parameter  $\beta_0$  und  $\beta_1$  lassen sich geometrisch interpretieren:

- $\beta_0$  ist der y-Achsenabschnitt
- $\beta_1$  die Steigung der Geraden
- Komplizierteres Modell muss nicht das bessere Modell sein
- Phänomen: Overfitting
- Fehler oder Ausreißer werden zu stark berücksichtigt
- In sehr vielen Fällen: Lineares Modell ausreichend

## Lineare Regression

### • Datensatz Werbung:



- **Verkauf** für ein bestimmtes Produkt (in Einheiten von tausend verkauften Produkten) als Funktion von Werbebudgets (in Einheiten von tausend CHF) für **TV**, **Radio** und **Zeitung**

Aufgrund dieser Daten: Statistiker erstellen Marketingplan, der für nächstes Jahr zu höheren Verkäufen führen soll

Welche Informationen sind nützlich, um solche Empfehlungen auszuarbeiten?

### Einfaches Regressionsmodell

Einfache lineare Regression: Sehr einfaches Verfahren, um einen quantitativen Output Y auf der Basis einer einzigen Inputvariable X

Annahme: Annähernd lineare Beziehung zwischen X und Y

Mathematisch: Lineare Beziehung:

$$Y \approx \beta_0 + \beta_1 X$$

- Dabei steht „ $\approx$ “ für „ist annähernd modelliert durch“

## Beispiel

- Beispiel **Werbung**:  $X$  Grösse **TV** und  $Y$  Grösse **Verkauf**

- Nach dem linearen Regressionsmodell gilt dann

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV}$$

- Grössen  $\beta_0$  und  $\beta_1$  sind unbekannte Konstanten, die den  $y$ -Achsenabschnitt und die Steigung des linearen Modells darstellen
- $\beta_0$  und  $\beta_1$  die *Koeffizienten* oder *Parameter* des Modells
- Koeffizienten werden aus den gegebenen Daten geschätzt
- Schätzungen  $\hat{\beta}_0$  und  $\hat{\beta}_1$  für die Modellkoeffizienten
- Sind diese Koeffizienten bekannt, so können zukünftige Verkäufe auf der Basis eines bestimmten Werbebudgets für TV vorhersagen
- Berechnung mittels:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

wobei  $\hat{y}$  die Vorhersage von  $Y$  auf Basis des Inputs  $X = x$  bezeichnet.

## Schätzung der Parameter

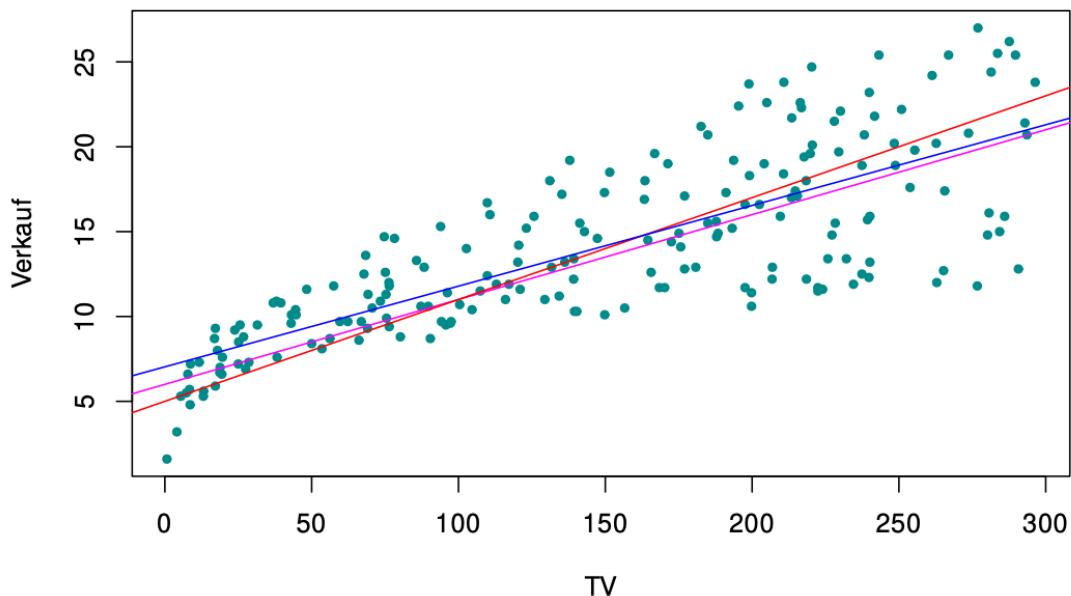
- Praxis:  $\beta_0$  und  $\beta_1$  unbekannt
- Bevor lineare Modell benützen → Koeffizienten schätzen
- Gehen von  $n$  Beobachtungspaaren aus:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

- Jedes Paar besteht aus je einer Messung von  $X$  und  $Y$
- Beispiel **Werbung**:  $n = 200$  verschiedene Beobachtungspaare (Märkte)
  - ▶  $x$ -Koordinate: TV-Budget
  - ▶  $y$ -Koordinate: entsprechenden Produktverkäufen
- Ziel:  $\hat{\beta}_0$  und  $\hat{\beta}_1$  so zu bestimmen, dass die Gerade  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  möglichst gut zu den Daten passt
- Das heisst, dass
$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$
für alle  $i = 1, \dots, n$
- Auf der linken Seite der obigen approximativen Beziehung steht der Messwert, auf der rechten der zugehörige  $y$ -Wert auf der Geraden
- Die Frage ist nun, was heisst „möglichst gut“?

## Beispiel

- Abbildung: Einige Geraden eingezeichnet, die gut zu Datenpunkten passen



- Welche passt am besten?

## Methode der kleinsten Quadrate

- Vorhergesagter Wert für  $Y$  abhängend vom  $i$ -ten Wert von  $X$ , also  $x_i$ :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

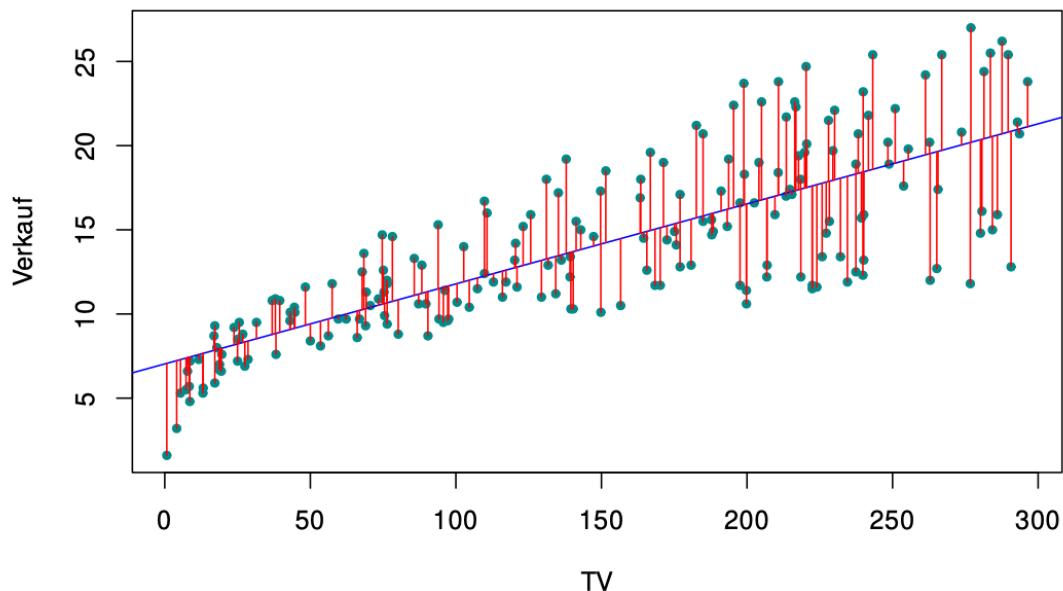
- $i$ -tes Residuum:

$$r_i = y_i - \hat{y}_i$$

- Differenz zwischen dem  $i$ -ten *beobachteten* Wert der Zielgrösse und dem  $i$ -ten von unserem linearen Modell *vorhergesagten* Wert der Zielgrösse

## Beispiel

- Abbildung: Residuen als Strecken rot eingezeichnet



- Residuen oberhalb der Geraden positiv, unterhalb der Geraden negativ
- Summe der *Quadrat*e der Residuen (RSS genannt)
- Es gilt dann

$$\text{RSS} = r_1^2 + r_2^2 + \dots + r_n^2$$

- Oder äquivalent:

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- Methode der kleinsten Quadrate:  $\hat{\beta}_0$  und  $\hat{\beta}_1$  so gewählt, dass RSS *minimal* wird

## Für die, die es interessiert

- Mit Differentialrechnung: Für  $\hat{\beta}_0$  und  $\hat{\beta}_1$  gilt:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

mit

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

- Mit Hilfe der Methode der kleinsten Quadrate geschätzte Koeffizienten für die einfache lineare Regression

## Beispiel

- Beispiel **Werbung**:  $\hat{\beta}_0$  und  $\hat{\beta}_1$  und die Regressionsgerade bestimmen:

```
lm(Verkauf ~ TV)
##
Call:
lm(formula = Verkauf ~ TV)
##
Coefficients:
(Intercept) TV
7.03259 0.04754
```

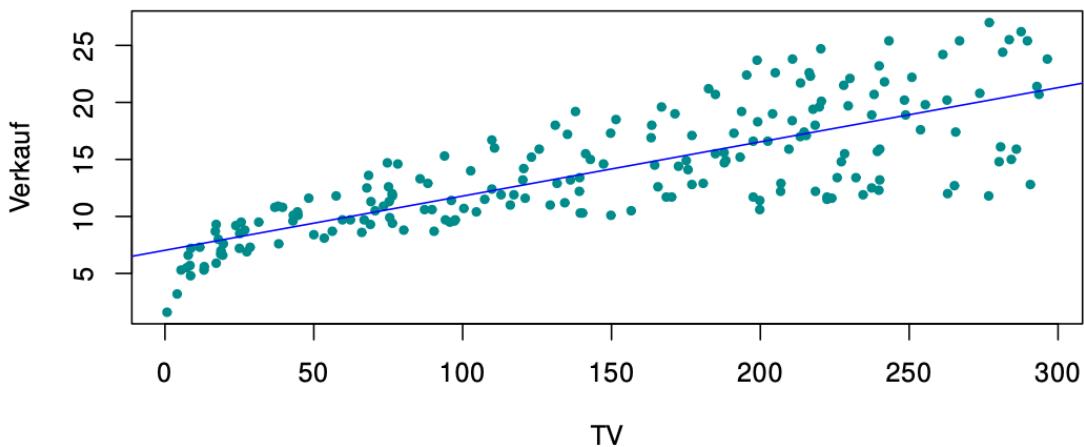
- Wert unter **Intercept**:  $\hat{\beta}_0 \rightarrow y$ -Achsenabschnitt
- Wert unter **TV**:  $\hat{\beta}_1 \rightarrow$  Steigung der Geraden
- Lineares Modell:

$$Y \approx 7.03 + 0.0475X$$

- Gemäss Näherung: Für zusätzliche CHF 1000 Werbeausgaben werden 47.5 zusätzliche Einheiten des Produktes verkauft
- Abbildung mit Regressionsgerade

```
plot(TV, Verkauf, col = "darkcyan", xlab = "TV", ylab = "Verkauf",
 pch = 20)

abline(lm(Verkauf ~ TV), col = "blue")
```



## Vertrauensintervall: Beispiel

- Vertrauensintervall Beispiel [Werbung](#) mit R:

```
confint(lm(Verkauf ~ TV), level = 0.95)
2.5 % 97.5 %
(Intercept) 6.12971927 7.93546783
TV 0.04223072 0.05284256
```

- 95 %-Vertrauensintervall von  $\beta_0$ :

$$[6.130, 7.935]$$

- Für  $\beta_1$ :

$$[0.042, 0.053]$$

- Ohne Werbung: Verkauf zwischen 6130 und 7935 Einheiten

- Für zusätzliche CHF 1000 für TV-Werbung durchschnittlich zwischen 42 und 53 Einheiten mehr verkaufen

## Hypothesentest: Statistische Signifikanz von $\beta_1$

- Standardfehler: Hypothesentest für die Regressionsparameter durchführen
- Häufigste Hypothesentest: Testen der *Nullhypothese*

$H_0$  : Es gibt *keinen* Zusammenhang zwischen  $X$  und  $Y$

- *Alternativhypothese*

$H_A$  : Es gibt *einen* Zusammenhang zwischen  $X$  und  $Y$

- Mathematisch:

$$H_0 : \beta_1 = 0$$

- Gegen:

$$H_A : \beta_1 \neq 0$$

- $\beta_1 = 0$ , dann:

$$Y = \beta_0 + \varepsilon$$

- $Y$  hängt *nicht* von  $X$  ab

- Nullhypothese testen:  $\widehat{\beta}_1$  genügend weit von 0 weg, damit  $\beta_1$  nicht 0

- Mit *t*-Statistik

## Beispiel

- $p$ -Wert von  $\beta_1$  im Beispiel **Werbung** berechnen:

```
summary(lm(Verkauf ~ TV))

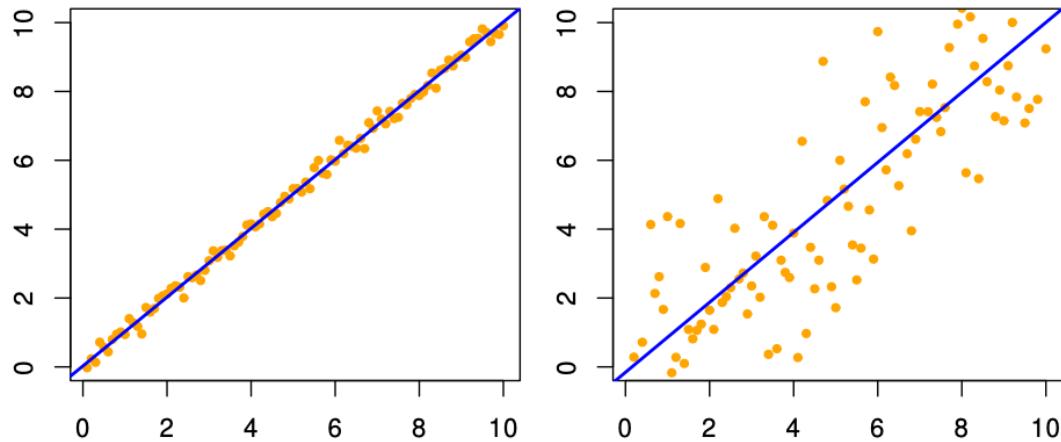
##
Call:
lm(formula = Verkauf ~ TV)
##
Residuals:
Min 1Q Median 3Q Max
-8.3860 -1.9545 -0.1913 2.0671 7.2124
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594 0.457843 15.36 <2e-16 ***
TV 0.047537 0.002691 17.67 <2e-16 ***

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099
F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16
```

- Eintrag **Coefficients** unter **Pr(>|t|)**:  $p$ -Wert  $2 \cdot 10^{-16}$
- Bei weitem kleiner als 0.05
- Nullhypotesen  $\beta_1 = 0$  verwerfen:  $\beta_1 \neq 0$
- Klarer Hinweise für Zusammenhang zwischen **TV** und **Verkauf**

## Abschätzung der Genauigkeit des Modells: $R^2$

- Nullhypothese verworfen: *In welchem Ausmass passt das Modell zu den Daten?*
- Abbildung:



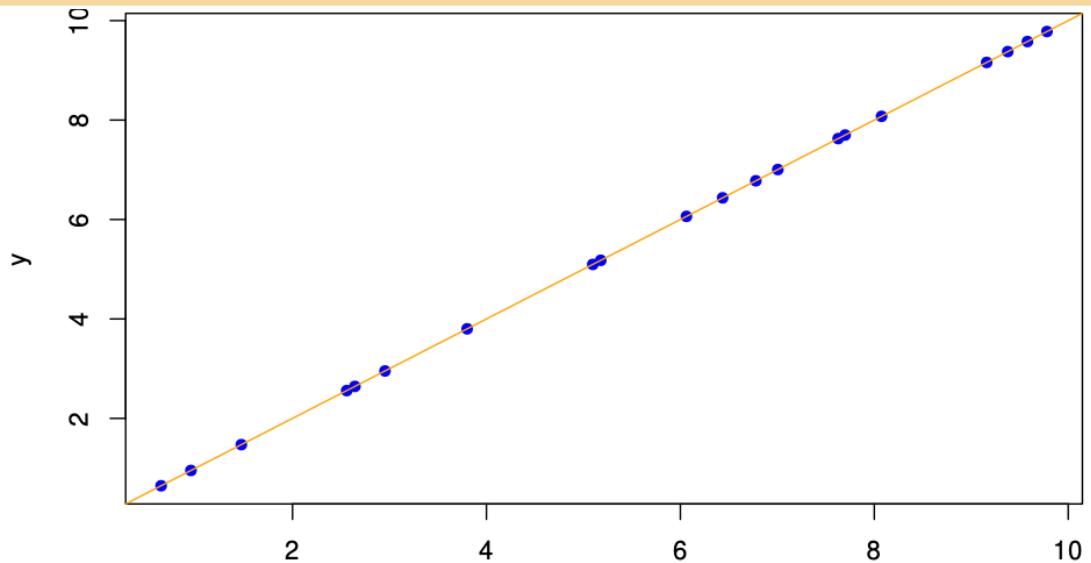
- ▶ Links: Steigende Gerade passt sehr gut zu Punkten
- ▶ Rechts: Steigende Gerade passt *nicht* gut zu Punkten

- Qualität einer linearen Regression abgeschätzt durch den *residual standard error* (RSE) und die  $R^2$ -Statistik
- $R^2$  wichtiger
- $R^2$ -Statistik: Wert zwischen 0 und 1
- Sie gibt an, welcher Anteil der Variabilität in  $Y$  mit Hilfe des Modells durch  $X$  erklärt werden
- Wert nahe bei 1: ein grosser Anteil der Variabilität wird durch die Regression erklärt. Das Modell beschreibt also die Daten sehr gut.
- Wert nahe bei 0: Regression erklärt die Variabilität der Zielvariablen nicht

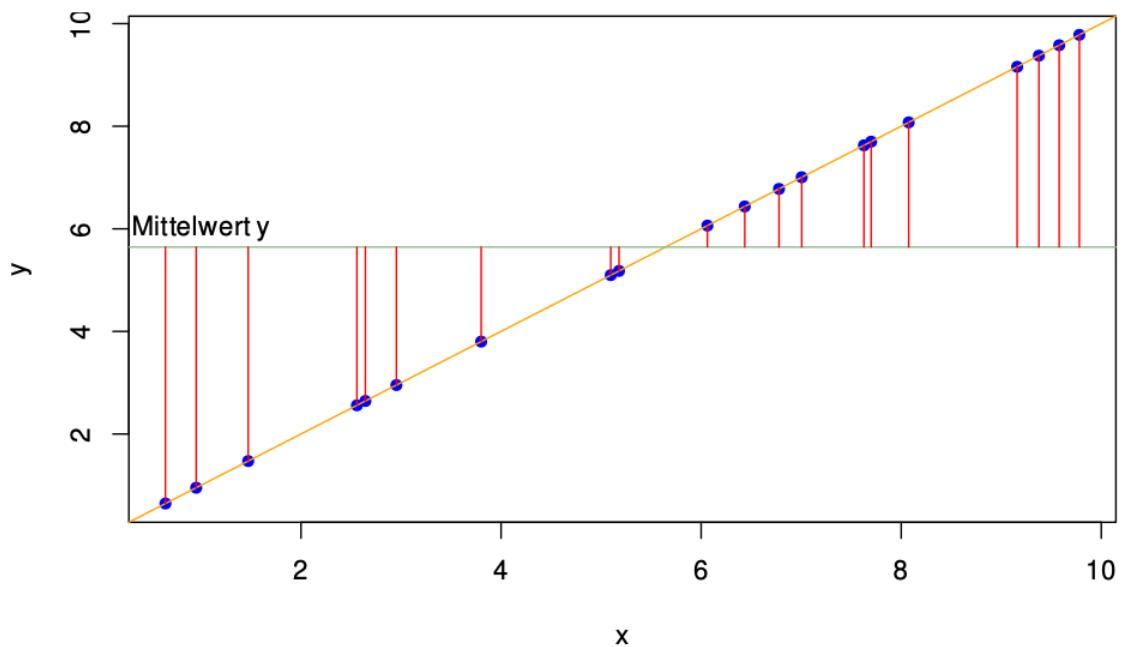
## Punkte folgen linearem Modell

- Abbildung:

```
x <- runif(min = 0, max = 10, n = 20)
y <- x
plot(x, y, col = "blue", pch = 16)
abline(lm(y ~ x), col = "orange")
```



- Abbildung Varianz:



- Varianz: „Mittelwert“ der quadrierten Unterschiede der y-Werte der Datenpunkte zu  $\bar{y}$

- Output:

- ▶ Korrelation:

```
cor(x, y)
[1] 1
```

- ▶  $R^2$ :

```
summary(lm(y ~ x))$r.squared
[1] 1
```

- ▶ Varianz:

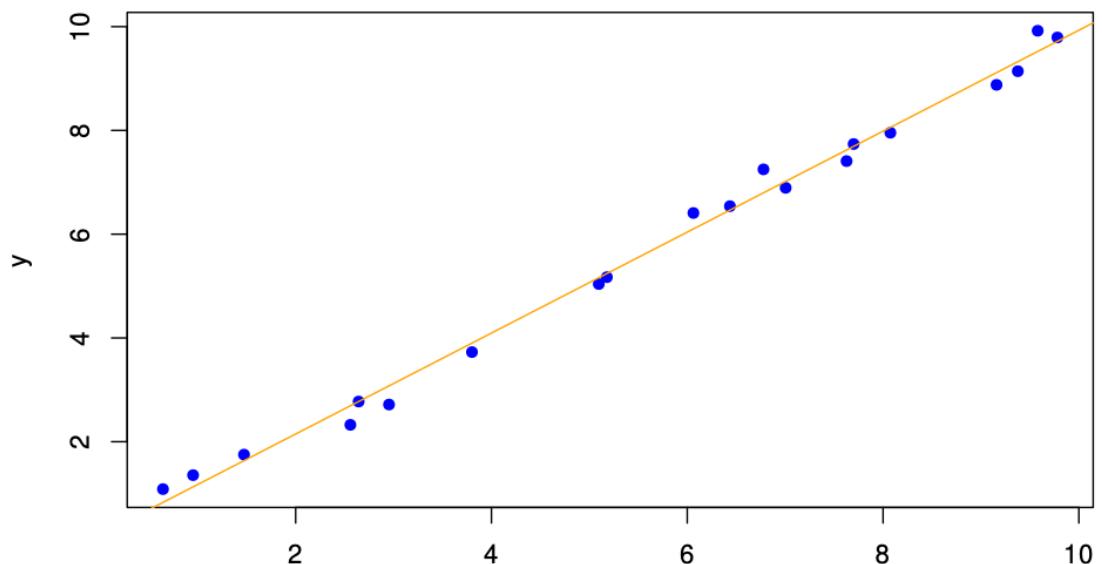
```
var(y)
[1] 8.998626
```

- ▶ 100% der Varianz von 9 wird durch das Modell erklärt

Punkte folgen mehr oder weniger linearem Modell

- Abbildung:

```
y <- x + rnorm(n = 20, mean = 0, sd = 0.2)
```



- Output:

- ▶ Korrelation:

```
cor(x, y)
[1] 0.9966885
```

- ▶  $R^2$ :

```
summary(lm(y ~ x))$r.squared
[1] 0.993388
```

- ▶ Varianz:

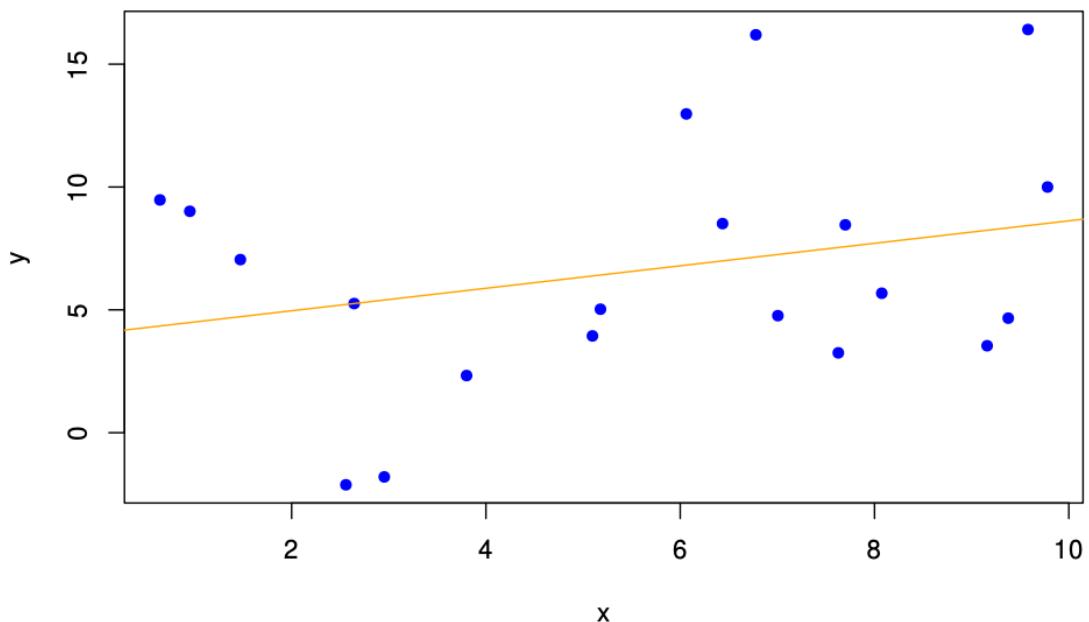
```
var(y)
[1] 8.573793
```

- ▶ 99.34% der Varianz von 8.57 wird durch das Modell erklärt

Punkte folgen dem linearen Modell nicht

- Abbildung:

```
y <- x + rnorm(n = 20, mean = 0, sd = 4)
```



- Output:

- ▶ Korrelation:

```
cor(x, y)
[1] 0.2769503
```

- ▶  $R^2$ :

```
summary(lm(y ~ x))$r.squared
[1] 0.07670148
```

- ▶ Varianz:

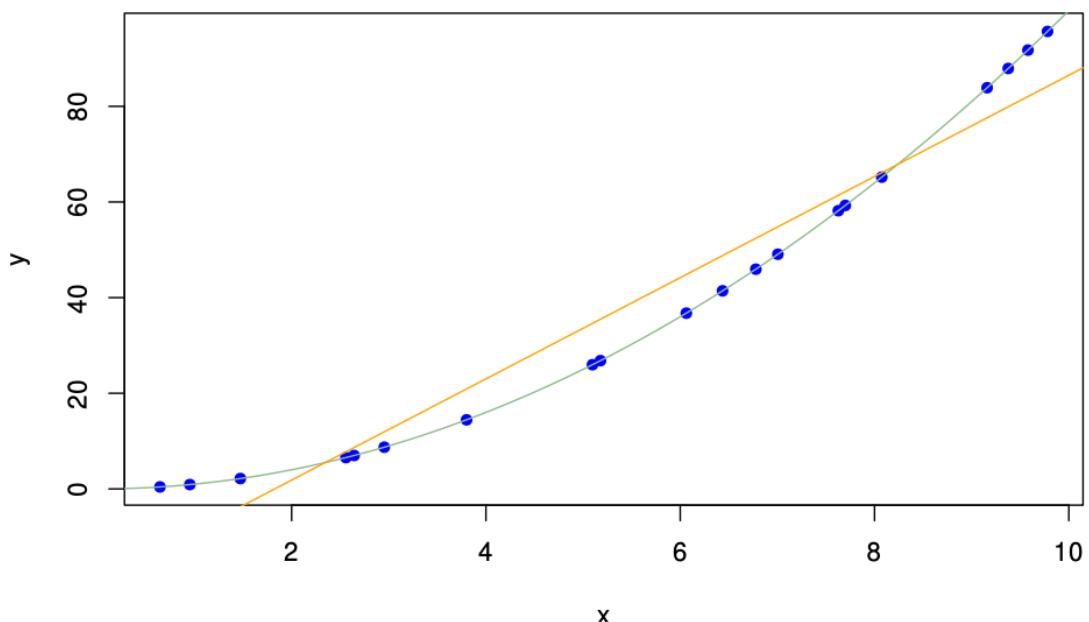
```
var(y)
[1] 24.55976
```

- ▶ 7.67% der Varianz von 24.56 wird durch das Modell erklärt

## Punkte folgen quadratischem Modell

- Abbildung:

```
y <- x^2
```



- Output:

- Korrelation:

```
cor(x, y)
[1] 0.9735588
```

- $R^2$ :

```
summary(lm(y ~ I(x^2)))$r.squared
[1] 1
```

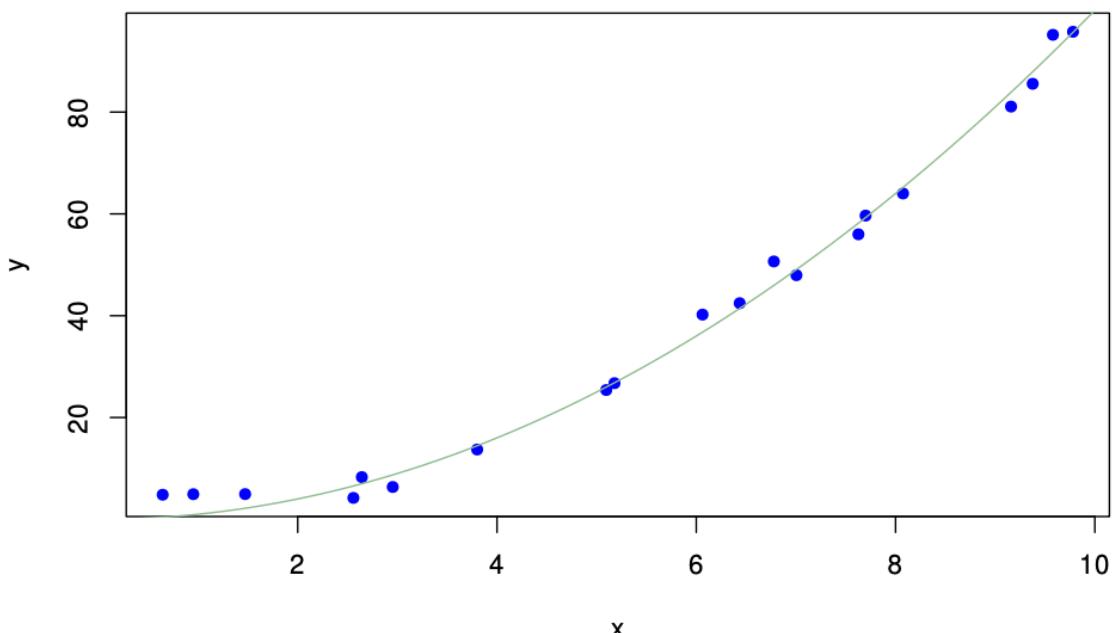
- Varianz:

```
var(y)
[1] 1063.22
```

- 100% der Varianz von 1063.22 wird durch das Modell erklärt

- Punkte folgen Modell:

```
y <- x^2 + rnorm(n = 20, mean = 0, sd = 2)
```



- Output:

- ▶ Korrelation:

```
cor(x, y)
[1] 0.9655864
```

- ▶  $R^2$ :

```
summary(lm(y ~ I(x^2)))$r.squared
[1] 0.9942619
```

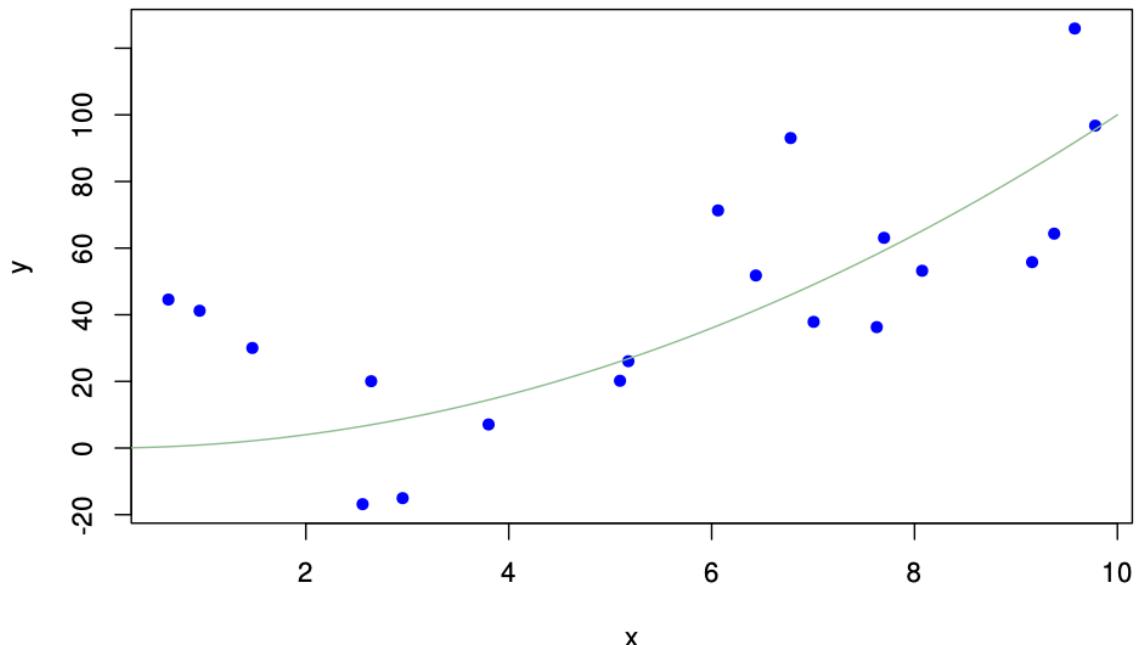
- ▶ Varianz:

```
var(y)
[1] 1026.155
```

- ▶ 99.43% der Varianz von 1026.15 wird durch das Modell erklärt

- Punkte folgen Modell:

```
y <- x^2 + rnorm(n = 20, mean = 0, sd = 20)
```



- Output:

- ▶ Korrelation:

```
cor(x, y)
[1] 0.6644753
```

- ▶  $R^2$ :

```
summary(lm(y ~ I(x^2)))$r.squared
[1] 0.5335559
```

- ▶ Varianz:

```
var(y)
[1] 1262.354
```

- ▶ 53.36% der Varianz von 1262.35 wird durch das Modell erklärt

## Multiple Lineare Regression

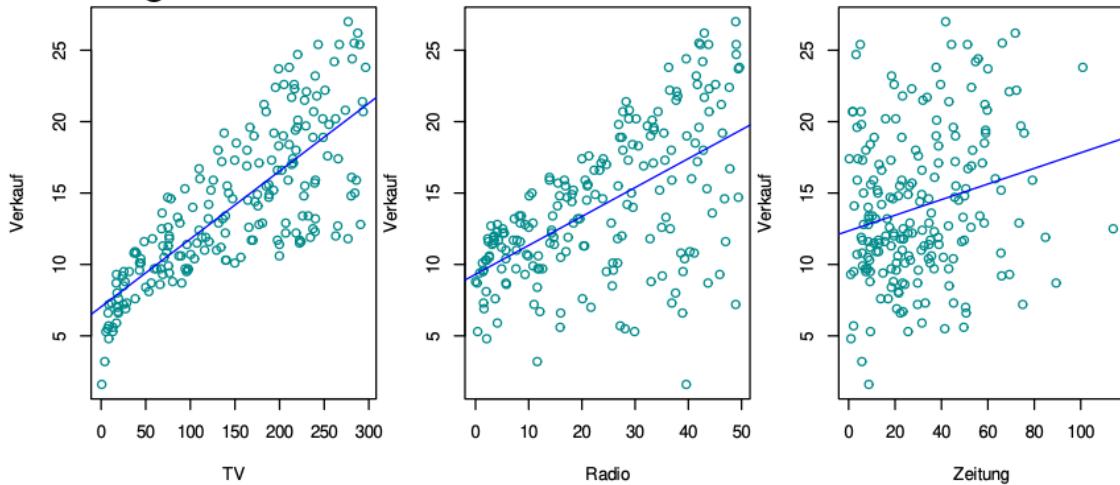
- Einfache lineare Regression: Nützliches Vorgehen, um Output aufgrund *einer* einzelnen erklärenden Variablen vorherzusagen
- Praxis: Output hängt oft von mehr als einer erklärenden Variablen ab

### Beispiel

- Datensatz **Werbung**: Zusammenhang zwischen **TV-Werbung** und **Verkauf** untersucht
- Auch Daten für Werbeausgaben für **Radio** und **Zeitung** vorhanden
- Frage: Wirken sich eine oder beide dieser Werbeausgaben auf Verkauf aus?
- Analyse der Verkaufszahlen erweitern: Beiden zusätzlichen Inputs mitberücksichtigen

- Möglichkeit: Für jedes separate Werbebudget eine einfache Regression durchführen

- Abbildung:



- Parameter und weitere wichtige Daten in Tabellen unten aufgeführt
- Einfache Regression von **Verkauf** auf **TV**:

|           | Koeffizient | Std.fehler | t-Statistik | P-Wert   |
|-----------|-------------|------------|-------------|----------|
| Intercept | 7.033       | 0.458      | 15.36       | < 0.0001 |
| TV        | 0.048       | 0.003      | 17.67       | < 0.0001 |

- Einfache Regression von **Verkauf** auf **Radio**:

|           | Koeffizient | Std.fehler | t-Statistik | P-Wert   |
|-----------|-------------|------------|-------------|----------|
| Intercept | 9.312       | 0.563      | 16.54       | < 0.0001 |
| Radio     | 0.203       | 0.020      | 9.92        | < 0.0001 |

- Einfache Regression von **Verkauf** auf **Zeitung**:

|           | Koeffizient | Std.fehler | t-Statistik | P-Wert   |
|-----------|-------------|------------|-------------|----------|
| Intercept | 12.351      | 0.621      | 19.88       | < 0.0001 |
| Zeitung   | 0.055       | 0.017      | 3.30        | < 0.0001 |

- Ansatz separate einfache lineare Regressionen: Nicht zufriedenstellend
- Erstens: Nicht klar, wie man für gegebene Werte der drei erklärenden Variablen eine Vorhersage für den Verkauf machen will:
  - ▶ Jeder Input durch *andere Regressionsgleichung* mit Verkauf verknüpft
- Zweitens: Jede der drei Regressionsgleichungen ignoriert die beiden anderen erklärenden Variablen für Bestimmung der Koeffizienten
- Kann zu sehr irreführenden Schätzungen der Wirkung der Werbeausgaben für jedes einzelne Medium auf den Verkauf haben kann, falls die drei erklärenden Variablen miteinander korrelieren
- Besser: Alle erklärenden Variablen direkt mitberücksichtigt
- Jeder erklärenden Variablen wird ein *eigener* Steigungskoeffizient in *einer* Gleichung zugeordnet
- Allgemein:  $p$  verschiedene erklärende Variablen
- *Multiples lineares Regressionsmodell:*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

- $X_j$ :  $j$ -ter Input
- $\beta_j$ : Zusammenhang zwischen *dieser* erklärenden Variablen und der Zielgröße  $Y$
- $\beta_j$ : Durchschnittliche Änderung der Zielgröße bei Änderung von  $X_j$  um eine Einheit, *wenn alle anderen erklärenden Variablen festgehalten werden*

## Beispiel

- Multiples lineares Regressionsmodell für den Datensatz **Werbung**:

$$\text{Verkauf} = \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung} + \varepsilon$$

- Also

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung}$$

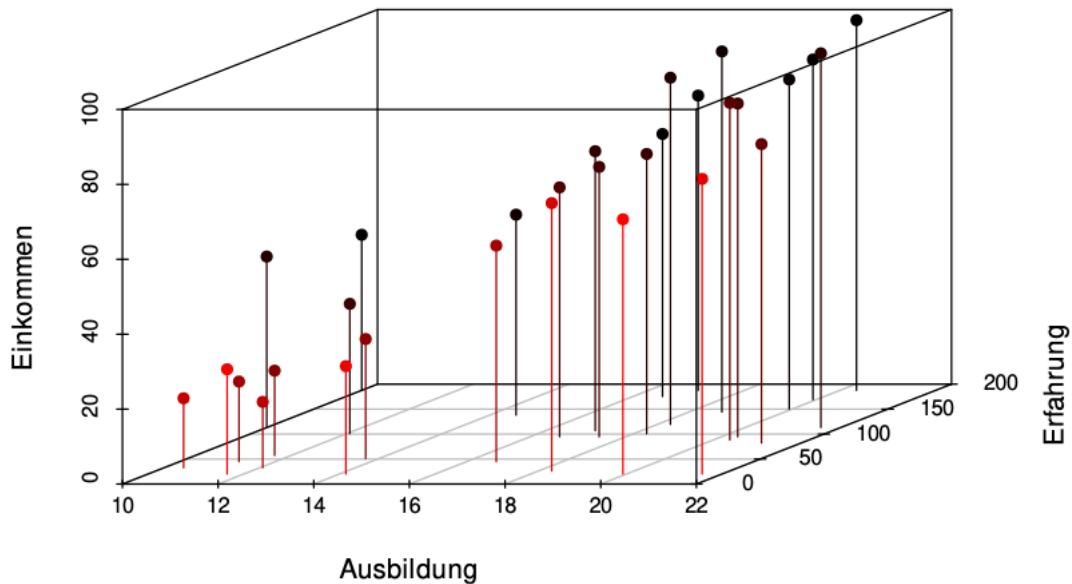
- Multiples lineares Modell verallgemeinert einfaches lineares Modell
- Berechnungen und Interpretationen für multiples Modell ähnlich, wenn auch meist komplizierter als beim linearen Modell
- Graphische Methoden: Entfallen für multiples lineare System praktisch vollends
- Datenpunkte für Beispiel vorher: Nicht darstellbar, da schon für erklärende Variablen drei Achsen gebraucht werden

## Beispiel: **Einkommen**

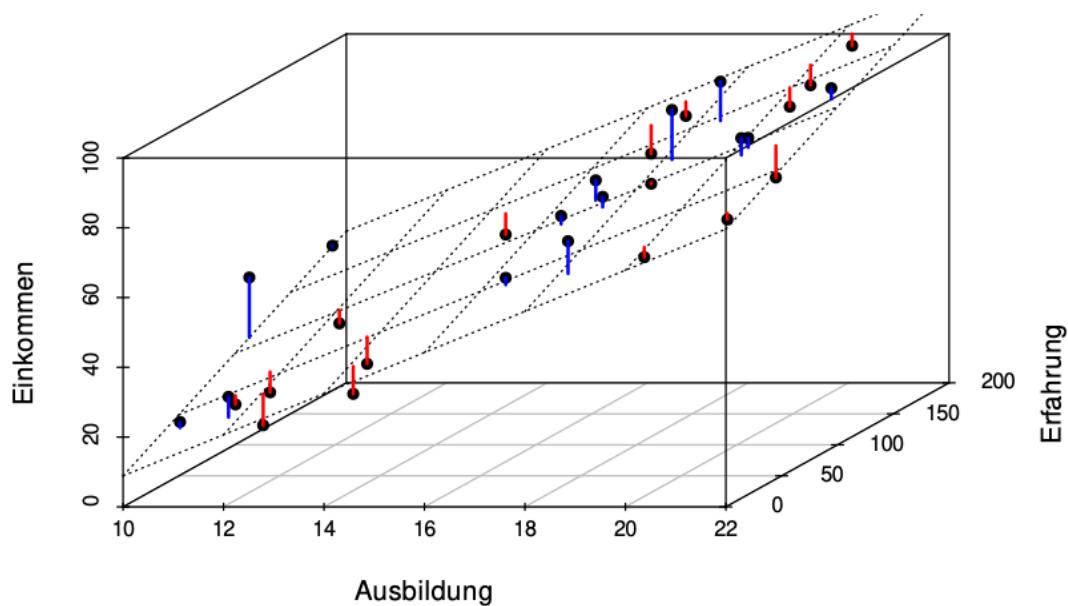
- Graphische Darstellung für zwei erklärende Variablen möglich
- Datensatz **Einkommen**
- Bis jetzt: **Ausbildung** einzige erklärende Variable
- Einkommen auch von **Erfahrung** (Anzahl Berufsmonate) abhängig
- Multiples lineares Modell:

$$\text{Einkommen} = \beta_0 + \beta_1 \cdot \text{Ausbildung} + \beta_2 \cdot \text{Erfahrung} + \varepsilon$$

- Datenpunkte im Raum:



- Analog einfaches lineares Regressionsmodell: Suchen *Ebene*, die am „besten“ zu den Datenpunkten passt



- Vorgehen analog zur einfachen linearen Regression
- Bestimmen Ebene so, dass Summe der Quadrate der Abstände der Datenpunkte zur Ebene minimal wird
- Strecken:
  - ▶ Blau: Punkte oberhalb der Ebene
  - ▶ Rot: Punkte unterhalb der Ebene
- Unterschiede von Punkten zu Ebene: *Residuen*
- Verwenden wieder *Methode der kleinsten Quadrate*
- Schätzung von  $\beta_0, \beta_1$  und  $\beta_2$  mit R:

$$\hat{\beta}_0 = -50.086; \quad \hat{\beta}_1 = 5.896; \quad \hat{\beta}_2 = 0.173$$

```
coef(lm(Einkommen ~ Ausbildung + Erfahrung))
(Intercept) Ausbildung Erfahrung
-50.0856388 5.8955560 0.1728555
```

- Multiples lineares Modell:

$$\text{Einkommen} \approx -50.086 + 5.896 \cdot \text{Ausbildung} + 0.173 \cdot \text{Erfahrung}$$

## Interpretation der Koeffizienten

- $\hat{\beta}_0 = -50.086$ :
  - ▶ Wenn Person keine Ausbildung und keine Erfahrung hat, so „erhält“ man CHF -50 086
  - ▶ Interpretation macht praktisch natürlich keinen Sinn
- $\hat{\beta}_1 = 5.896$ :
  - ▶ Bei konstanter Erfahrung verdient man pro zusätzliches Ausbildungsjahr Ausbildung CHF 5896 mehr
- $\hat{\beta}_2 = 0.173$ :
  - ▶ Bei konstanter Ausbildung verdient man pro zusätzlichen Monat Arbeitserfahrung CHF 173 mehr

## Allgemein: Schätzung der Regressionskoeffizienten

- Wie einfache linearer Regression: Regressionskoeffizienten  $\beta_0, \beta_1, \dots, \beta_p$  i. A. unbekannt
- Müssen sie aus Daten schätzen:

$$\hat{\beta}_0, \quad \hat{\beta}_1, \quad \dots, \quad \hat{\beta}_p$$

- Aufgrund der Schätzungen kann man Vorhersagen machen:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \dots + \hat{\beta}_p x_p$$

- Parameter wieder mit der Methode der kleinsten Quadrate schätzen

## Beispiel

- R: Multiples lineares Regressionsmodell für Werbung:

```
coef(lm(Verkauf ~ TV + Radio + Zeitung))
(Intercept) TV Radio Zeitung
2.938889369 0.045764645 0.188530017 -0.001037493
```

- Es gilt:

$$\text{Verkauf} \approx 2.94 + 0.046 \cdot \text{TV} + 0.189 \cdot \text{Radio} - 0.001 \cdot \text{Zeitung}$$

- Koeffizienten interpretieren:

- ▶ Für gegebene Werbeausgaben für Radio und Zeitung werden für zusätzliche CHF 1000 Werbeausgaben für das TV ungefähr 46 Einheiten mehr verkauft
- ▶ Für gegebene Werbeausgaben für TV und Zeitung werden für zusätzliche CHF 1000 Werbeausgaben für das Radio ungefähr 189 Einheiten mehr verkauft
- ▶ Interessant: Bei der Zeitung würde man *weniger* Produkte verkaufen, wenn man *mehr* investiert

- Tabelle: Weitere wichtige Werte:

|           | Koeffizient | Std.fehler | t-Statistik | P-Wert   |
|-----------|-------------|------------|-------------|----------|
| Intercept | 2.939       | 0.3119     | 9.42        | < 0.0001 |
| TV        | 0.046       | 0.0014     | 32.81       | < 0.0001 |
| Radio     | 0.189       | 0.0086     | 21.89       | < 0.0001 |
| Zeitung   | -0.001      | 0.0059     | -0.18       | 0.8599   |

- Code: `coef` durch `summary` ersetzen

```
fit <- lm(Verkauf ~ TV + Radio + Zeitung)

summary(fit)
##
Call:
lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
Residuals:
Min 1Q Median 3Q Max
-8.8277 -0.8908 0.2418 1.1893 2.8292
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.938889 0.311908 9.422 <2e-16 ***
TV 0.045765 0.001395 32.809 <2e-16 ***
Radio 0.188530 0.008611 21.893 <2e-16 ***
Zeitung -0.001037 0.005871 -0.177 0.86

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16
```

- Koeffizienten der separaten einfachen linearen Regressionen in Slide 5
- Steigungskoeffizienten der multiplen linearen Regression für `TV` und `Radio` sehr ähnlich:
  - ▶ `TV`: 0.46 (multiple), 0.48 (einfach)
  - ▶ `Radio`: 0.189 (multiple), 0.203 (einfach)
- Geschätzter Regressionskoeffizient  $\hat{\beta}_3$  für `TV` zeigt anderes Verhalten:
  - ▶ Einfach: 0.055 (ungleich 0)
  - ▶ Multiple: -0.001 (fast gleich 0)
- Entsprechende  $p$ -Werte:
  - ▶ Einfach: < 0.0001 (hochsignifikant)
  - ▶ Multiple: 0.86 (bei weitem nicht mehr signifikant)

- Einfache und multiple Regressionskoeffizienten können sehr verschieden sein
- Einfache Regression: Steigung gibt die Änderung der Zielgrösse **Verkauf** an, wenn man CHF 1000 mehr für die Zeitungswerbung ausgibt, wobei die beiden anderen erklärenden Variablen **TV** und **Radio** *ignoriert* werden
- Multiple lineare Regression: Steigung für **Zeitung** beschreibt die Änderung der Zielgrösse **Verkauf**, wenn man CHF 1000 mehr für Zeitungswerbung ausgibt, wobei die beiden anderen erklärenden Variablen **TV** und **Radio** *festgehalten* werden
- Macht es Sinn, dass die multiple Regression keinen Zusammenhang zwischen **Verkauf** und **Zeitung** andeutet, aber die einfache Regression das Gegenteil impliziert?
- Es macht in der Tat Sinn
- Tabelle mit Korrelationskoeffizienten:

|                | <b>TV</b> | <b>Radio</b> | <b>Zeitung</b> | <b>Verkauf</b> |
|----------------|-----------|--------------|----------------|----------------|
| <b>TV</b>      | 1.0000    | 0.0548       | 0.0567         | 0.7822         |
| <b>Radio</b>   |           | 1.0000       | 0.3541         | 0.5762         |
| <b>Zeitung</b> |           |              | 1.0000         | 0.2283         |
| <b>Verkauf</b> |           |              |                | 1.0000         |

- Code:

```
cor(data.frame(TV, Radio, Zeitung, Verkauf))
TV Radio Zeitung Verkauf
TV 1.00000000 0.05480866 0.05664787 0.7822244
Radio 0.05480866 1.00000000 0.35410375 0.5762226
Zeitung 0.05664787 0.35410375 1.00000000 0.2282990
Verkauf 0.78222442 0.57622257 0.22829903 1.0000000
```

- Korrelationskoeffizient **Radio** und **Zeitung**: 0.35
- Was bedeutet dies?
- Zeigt Tendenz bei höheren Werbeausgaben für **Radio** auch mehr in Werbung für **Zeitung** zu investieren
- Annahme: Multiples Regressionsmodell *korrekt*
- Ausgaben für **Zeitung**: Kein direkter Einfluss auf Zielgröße **Verkauf**
- Werbeausgaben für **Radio**: Höhere Verkäufe
  - In Märkten, wo mehr in die Werbung fürs Radio investiert wird, auch Ausgaben für **Zeitung** grösser, da Korrelationskoeffizienten von 0.35
  - Einfache lineare Regression: Nur Zusammenhang zwischen **Zeitung** und **Verkauf**, wobei für höhere Werte von **Zeitung** auch höhere Werte für **Verkauf** beobachtet werden
- Aber: Zeitungswerbung beeinflusst Verkäufe *nicht*
- Höhere Werte für **Zeitung** wegen Korrelation auch grössere Werte für **Radio** zur Folge: *Diese Grösse beeinflusst Verkauf*
- **Zeitung** schmückt sich hier mit fremden Lorbeeren, nämlich dem Erfolg von **Radio** auf **Verkauf**
- Dieses Resultat steht in Konflikt mit Intuition
- Tritt in realen Situationen aber häufig auf

## Absurdes Beispiel

- Einfache Regression: Zusammenhang zwischen Haiattacken und Glaceverkäufen an einem bestimmten Strand
- Je grösser Glaceverkäufe, desto häufiger ereignen sich Haiattacken
- Absurde Idee: Glaceverkäufe an diesem Strand verbieten, damit es keine Haiattacken auf Menschen mehr gibt
- Wo liegt aber der Zusammenhang?
- Real: Bei heissem Wetter kommen mehr Menschen an den Strand  
→ mehr Glaceverkäufe → mehr Haiattacken
- Confounder: Temperatur
- Multiples Regressionsmodell von Haiattacken mit Glaceverkäufen *und* Temperatur: Glaceverkauf keinen Einfluss mehr auf Haiattacken, Lufttemperatur allerdings schon

## Einige wichtige Fragestellungen

- *Ist mindestens eine der erklärenden Variablen  $X_1, \dots, X_p$  nützlich, um die Zielgröße vorherzusagen?*
- *Spielen alle erklärenden Variablen  $X_1, \dots, X_p$  für die Vorhersage von  $Y$  eine Rolle, oder nur eine Teilmenge der erklärenden Variablen?*
- *Wie gut passt das Modell zu den Daten?*
- *Welche Zielgröße kann man aufgrund konkreter Werte der erklärenden Variablen vorhersagen?*
- *Wie genau ist diese Vorhersage?*

# Gibt es einen Zusammenhang zwischen den erklärenden Variablen und der Zielgröße?

- Hypothesentest:
- Multiple lineare Regression mit  $p$  erklärenden Variablen: Alle Regressionskoeffizienten ausser  $\beta_0$  Null sind (keine Variable hat Einfluss):

$$\beta_1 = \beta_2 = \dots = \beta_p = 0$$

- Nullhypothese

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

- Alternativhypothese

$$H_A : \text{mindestens ein } \beta_i \text{ ist ungleich 0}$$

- Berechnung der  $F$ -Statistik mit  $p$ -Wert

## Beispiel

- $p$ -Wert für das multiple lineare Modell für den Datensatz [Werbung](#):

```
summary(lm(Verkauf ~ TV + Radio + Zeitung))

##
Call:
lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
Residuals:
Min 1Q Median 3Q Max
-8.8277 -0.8908 0.2418 1.1893 2.8292
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.938889 0.311908 9.422 <2e-16 ***
TV 0.045765 0.001395 32.809 <2e-16 ***
Radio 0.188530 0.008611 21.893 <2e-16 ***
Zeitung -0.001037 0.005871 -0.177 0.86

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16
```

- R-Ausgabe **p-value** in Zeile für *F*-Statistik: *p*-Wert für multiples lineares Modell praktisch null
- Sehr überzeugender Hinweis: Mindestens eine erklärende Variable ist für Zunahme von **Verkauf** bei vergrösserten Werbeausgaben verantwortlich

## Beispiel

- Warum betrachten wir nicht einfach die einzelnen *p*-Werte?
- Wenn einer unterhalb des Signifikanzniveaus liegt, dann weiss man, dass mindestens eine Variable Einfluss hat
- Aber: Wegen dem Prinzip des Hypothesentest, ist statistisch signifikanter *p*-Wert zu 5 % zufällig
- Folgendes Beispiel: Keine Variable ist signifikant
- Alle  $\beta_1$ -Werte in der Nähe von 0
- Aber: Gibt zufällige Abweichungen, wo die zugehörigen *p*-Werte signifikant sind
- Darum: Wenn sehr viele Variable vorhanden sind, ist praktisch immer eine signifikant, obwohl in Wahrheit keine ist
- Code:

```
set.seed(4)
v <- 20
d <- 500

df <- matrix(rnorm(v * d), nrow = d)
head(df)
df <- data.frame(df)

Y <- rnorm(d)
Y

df$Y <- Y

fit <- lm(Y ~ ., , data = df)
summary(fit)
```

### ● Output:

```

Call:
lm(formula = Y ~ ., data = df)

Residuals:
Min 1Q Median 3Q Max
-2.62976 -0.66857 0.00927 0.64462 2.81840

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.029669 0.047272 -0.628 0.5305
X1 -0.010970 0.048886 -0.224 0.8225
X2 -0.036943 0.049150 -0.752 0.4526
X3 -0.005961 0.047734 -0.125 0.9007
X4 -0.018073 0.047726 -0.379 0.7051
X5 0.005827 0.048524 0.120 0.9045
X6 -0.127798 0.049554 -2.579 0.0102 *
X7 -0.052386 0.049816 -1.052 0.2935
X8 0.020574 0.048557 0.424 0.6720
X9 -0.015178 0.047941 -0.317 0.7517
X10 -0.015107 0.046988 -0.322 0.7480
X11 0.005580 0.046517 0.120 0.9046
X12 -0.004676 0.046583 -0.100 0.9201
X13 -0.021652 0.049114 -0.441 0.6595
X14 -0.093800 0.046075 -2.036 0.0423 *
X15 0.019740 0.047451 0.416 0.6776
X16 0.042796 0.045267 0.945 0.3449
X17 -0.074511 0.049061 -1.519 0.1295
X18 0.041733 0.047568 0.877 0.3808
X19 -0.078238 0.047492 -1.647 0.1001
X20 -0.057475 0.048156 -1.194 0.2333

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.042 on 479 degrees of freedom
```

## Bestimmung der wichtigen erklärenden Variablen

- Zuerst entscheiden: Haben erklärende Variablen überhaupt Einfluss auf Zielgrösse
- Entscheid: Mit Hilfe  $F$ -Statistik und zugehörigem  $p$ -Wert
- Beeinflusst mindestens eine Variable die Zielgrösse: Welche erklärende Variablen sind dies?
- Können einzelne  $p$ -Werte wie in Tabelle betrachten

- Möglich: Alle erklärenden Variablen beeinflussen Zielgröße, aber meist sind es nur einige wenige
- Aufgabe: Variablen bestimmen und dann Modell aufstellen, welches nur diese Variablen enthält
- Interessiert an möglichst einfachen Modell, das zu den Daten passt
- Welche Variablen sind wichtig?
- Prozedere: *Variablenelektion* (nächstes Mal)

### Wie gut passt das Modell zu den Daten?

- Bestimmtheitsmaß  $R^2$
- Datensatz **Werbung** ist der  $R^2$ -Wert 0.8972
- $R^2$  erhöht sich, je mehr erklärende Variablen berücksichtigt werden

### Beispiel: Vorhersage

- *Vertrauensintervall*, um die Ungewissheit für den *durchschnittlichen Verkauf* für eine grosse Zahl von Städten zu quantifizieren
- Nur die erklärenden Variablen **TV** und **Radio** berücksichtigen, da **Zeitung** für **Verkauf** keinen Einfluss hat
- Wenden CHF 100 000 für **TV**-Werbung und CHF 20 000 für **Radio**-Werbung in jeder Stadt auf → 95 %-Vertrauensintervall

[10 985, 11 528]

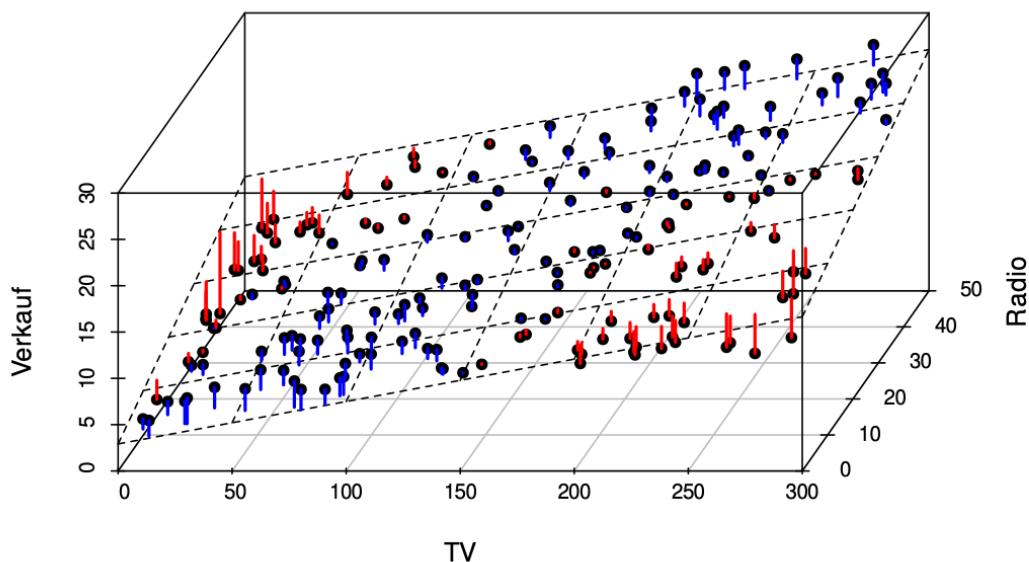
```
predict(lm(Verkauf ~ TV + Radio),
 interval = "confidence",
 data.frame(TV = 100, Radio = 20))

)
fit lwr upr
1 11.25647 10.98525 11.52768
```

- Interpretation lautet wie folgt: 95 % aller Intervalle dieser Form enthalten den wahren Wert  $f(X_1, X_2)$
- Was heisst dies?
- Sammeln grosse Menge von Datensätzen wie den **Werbung**-Datensatz
- Für jeden Datensatz jeweils das Vertrauensintervall für den wahren durchschnittlichen **Verkauf** berechnen (bei CHF 100 000 für **TV**-Werbung und CHF 20 000 für **Radio**-Werbung)
- In 95 % dieser Intervalle liegt der wahre Wert vom mittleren **Verkauf**

## Keine lineare Regression

- Graphischer Überblick: Probleme mit dem Modell aufzeigen, die für die numerischen Werte unsichtbar sind:



- Dreidimensionales Streudiagramm: Nur **TV** und **Radio** berücksichtigt
- Gestrichelt: Regressionsebene
- Beobachtung: Werte der Ebene zu gross, wenn Werbeausgaben ausschliesslich entweder für **TV** oder **Radio** aufgewendet wurden
- Hinten links: Werbung nur für **Radio**
- Vorne rechts: nur für **TV**
- Werte der Ebene sind zu tief, wenn Werbeausgaben gleichmässig auf **TV** und **Radio** verteilt werden
- Nichtlineares Muster: Kann nicht genau durch eine lineare Regression beschrieben werden
- Plot deutet *Interaktion-* oder *Synergieeffekt* an: Grössere Verkäufen, wenn Werbeausgaben aufgeteilt werden

## Aufhebung der Annahme bezüglich Additivität

- Interaktionseffekte

- Beispiel Werbung:

```
fit <- lm(Verkauf ~ TV + Radio + TV * Radio)

summary(fit)
##
Call:
lm(formula = Verkauf ~ TV + Radio + TV * Radio)
##
Residuals:
Min 1Q Median 3Q Max
-6.3366 -0.4028 0.1831 0.5948 1.5246
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00 2.479e-01 27.233 <2e-16 ***
TV 1.910e-02 1.504e-03 12.699 <2e-16 ***
Radio 2.886e-02 8.905e-03 3.241 0.0014 **
TV:Radio 1.086e-03 5.242e-05 20.727 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared: 0.9678, Adjusted R-squared: 0.9673
F-statistic: 1963 on 3 and 196 DF, p-value: < 2.2e-16
```

- $p$ -Werte zu **TV**, **Radio** und dem Interaktionsterm **TV · Radio**: Statistisch signifikant
- Scheint klar: Alle diese Variablen sollten im Modell enthalten sein
- Möglich:  $p$ -Wert für den Interaktionsterm sehr klein ist, aber die  $p$ -Werte der Haupteffekte (hier **TV** und **Radio**) sind es nicht

## Qualitative Variablen

### Qualitative erklärende Variablen

Bisher angenommen: Alle Variablen quantitativ in linearem Regressionssystem. Aber: Oft sind einige erklärenden Variablen qualitativ.

### Beispiel

- Datensatz **Credit** wurde in den USA erhoben
- Enthält für eine grössere Anzahl Individuen:
  - ▶ **Balance** (monatliche Kreditkartenrechnung): Zielgrösse, quantitativ
  - ▶ **Age** (Alter): erklärend, quantitativ
  - ▶ **Cards** (Anzahl Kreditkarten): erklärend, quantitativ
  - ▶ **Education** (Anzahl Jahre Ausbildung): erklärend, quantitativ
  - ▶ **Income** (Einkommen in Tausenden Dollars): erklärend, quantitativ
  - ▶ **Limit** (Kreditkartenlimite): erklärend, quantitativ
  - ▶ **Rating** (Kreditwürdigkeit): erklärend, quantitativ

- Datensatz:

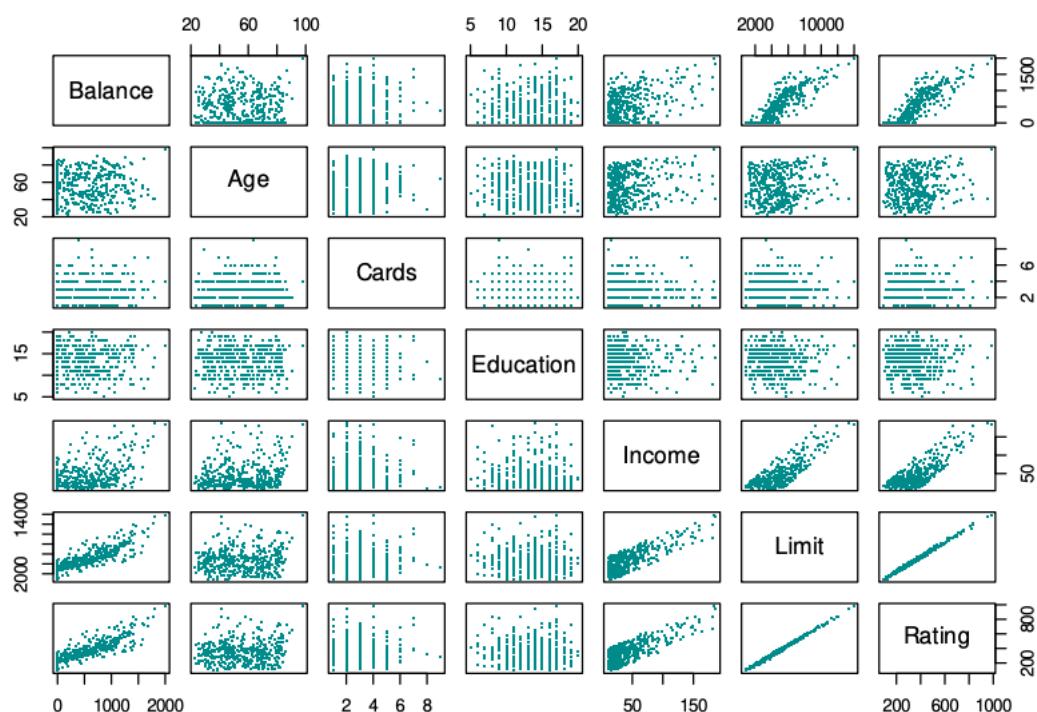
```
Credit <- read.csv("../Data/Credit.csv")[, -1]
head(Credit)

Income Limit Rating Cards Age Education Gender Student
1 14.891 3606 283 2 34 11 Male No
2 106.025 6645 483 3 82 15 Female Yes
3 104.593 7075 514 4 71 11 Male No
4 148.924 9504 681 3 36 11 Female No
5 55.882 4897 357 2 68 16 Male No
6 80.180 8047 569 4 77 10 Male No

Married Ethnicity Balance
1 Yes Caucasian 333
2 Yes Asian 903
3 No Asian 580
4 No Asian 964
5 Yes Caucasian 331
6 No Caucasian 1151

colnames(Credit)
[1] "Income" "Limit" "Rating" "Cards"
[5] "Age" "Education" "Gender" "Student"
[9] "Married" "Ethnicity" "Balance"
```

- Abbildung:



- Code:

```
Credit <- read.csv("../Data/Credit.csv")
pairs(~Balance + Age + Cards + Education + Income + Limit + Rating,
 data = Credit, pch = ".", col = "darkcyan")
```

- Streudiagramme von Paaren von Variablen: Identität gegeben durch entsprechenden Spalten- und Zeilenkennzeichnungen
- Plot direkt rechts des Wortes „Balance“: Streudiagramm der Variablen `age` und `balance`
- Streudiagramme:
  - ▶ `Age - Balance`: Kein Zusammenhang
  - ▶ `Education - Balance`: Kein Zusammenhang
  - ▶ `Income - Balance`: Schwacher Zusammenhang
  - ▶ `Limit - Balance`: Starker Zusammenhang
- Neben quantitativen noch vier erklärende qualitative Variablen:
  - ▶ `Gender` (Geschlecht)
  - ▶ `Student` (Studentenstatus)
  - ▶ `Ethnicity` (Ethnie)
- Qualitativ erklärende Variablen heißen auch *Faktoren*
- Faktoren nehmen *Stufen* oder *Levels* an:
  - ▶ `Gender`: male, female
  - ▶ `Student`: ja, nein
  - ▶ `Ethnicity`: Kaukasier, Afroamerikaner, Asiat

## Qualitative erklärende Variable mit nur zwei Levels

- Beispiel **Balance**: Unterschied zwischen Männern und Frauen
- Andere Variablen werden für den Moment ignoriert
- Qualitative erklärende Variable mit zwei *Levels* (mögliche Werte): Hinzunahme dieser Variable in Regressionsmodell sehr einfach
- Führen Indikatorvariable (oder *Dummy-Variable*) ein, die nur zwei mögliche numerische Werte annehmen kann

### Beispiel

- Für **Gender**:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ 0 & \text{falls } i\text{-te Person männlich} \end{cases}$$

- Verwenden diese Variable als erklärende Variable im Regressionsmodell
- Modell:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- $\beta_0$ : durchschn. Kreditkartenrechnungen der Männern
- $\beta_0 + \beta_1$ : durchschn. Kreditkartenrechnungen der Frauen
- $\beta_1$ : durchschn. *Unterschied* der Rechnungen Männern/Frauen

- Tabelle: Koeffizientenschätzungen für unser Modell:

|                 | Koeffizient | Std.fehler | t-Statistik | P-Wert   |
|-----------------|-------------|------------|-------------|----------|
| Intercept       | 509.80      | 33.13      | 15.389      | < 0.0001 |
| gender [female] | 19.73       | 46.05      | 0.429       | 0.6690   |

```
balance <- Credit[, "Balance"]
gender <- Credit[, "Gender"] == "Female"
round(summary(lm(balance ~ gender))$coef, digits = 5)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 509.80311 33.12808 15.38885 0.00000
genderTRUE 19.73312 46.05121 0.42850 0.66852
```

- Geschätzte durchschnittliche Rechnungen für Männer: \\$ 509.80
- Geschätzter Unterschied zu Frauen: \\$ 19.73
- Frauen: \\$ 509.80 + \\$ 19.73 = \\$ 529.53
- $p$ -Wert für Indikatorvariable  $\beta_1$  mit 0.6690 sehr hoch
- Kein statistisch signifikanter Unterschied der `balance` von Frauen und Männern
- Beispiel vorher: Frauen mit 1 und Männer mit 0 kodiert
- Völlig willkürlich
- Kodierung: *Kein* Einfluss auf Grad der Anpassung des Modells an Daten
- Unterschiedliche Kodierung: Unterschiedliche Interpretation der Koeffizienten
- Kodierung Männer mit 1 und Frauen mit 0
- Schätzung für die Parameter  $\beta_0$  und  $\beta_1$  \\$ 529.53, resp. \\$ -19.73
- Entspricht wiederum Rechnungen von:
  - ▶ Frauen: \\$ 529.53
  - ▶ Männer: \\$ 529.53 - \\$ 19.73 = \\$ 509.80
- Dasselbe Resultat wie vorher

## Beispiel

- Anstatt der 0/1-Kodierung:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ -1 & \text{falls } i\text{-te Person männlich} \end{cases}$$

- Regressionsmodell:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- $\beta_0$ : Durchschn. Rechnungen ohne Berücksichtigung des Geschlechts
- $\beta_1$ : Wert, mit welchem Frauen über dem Durchschnitt liegen und mit welchem Männer unter dem Durchschnitt liegen
- $\beta_0$  durch \\$ 519.665 geschätzt: Durchschn. Rechnungen von \\$ 509.80 für Männer und von \\$ 529.53 für Frauen
- Schätzung \\$ 9.865 für  $\beta_1$ : Hälfte vom Unterschied \\$ 19.73 zwischen Männern und Frauen
- Wichtig: Vorhersagen für die Zielgröße hängen *nicht* von Kodierung ab
- Einziger Unterschied: Interpretation der Koeffizienten

### Qualitative erklärende Variablen mit mehr als zwei Levels

- Qualitative erklärende Variable kann mehr als zwei Levels haben
- Eine Indikatorvariable für alle möglichen Werte reicht nicht
- In dieser Situation: Zusätzliche Indikatorvariable hinzufügen

## Beispiel

- Variable **Ethnicity**: *Drei mögliche Levels*

- Wählen *zwei* verschiedene Indikatorvariablen

- *Wahl* der 1. Indikatorvariablen:

$$x_{i1} = \begin{cases} 1 & \text{falls } i\text{-te Person asiatisch} \\ 0 & \text{falls } i\text{-te Person nicht asiatisch} \end{cases}$$

- 2. Indikatorvariable:

$$x_{i2} = \begin{cases} 1 & \text{falls } i\text{-te Person kaukasisch} \\ 0 & \text{falls } i\text{-te Person nicht kaukasisch} \end{cases}$$

- Beide Variablen in Regressionsgleichung aufnehmen:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch} \end{cases}$$

- $\beta_0$ : Durchschn. Kreditkartenrechnungen von Afroamerikanern
- $\beta_1$ : Differenz der durchschn. Rechnungen von Afroamerikanern und Asiaten
- $\beta_2$ : Differenz der durchschn. Rechnungen von Afroamerikanern und Kaukasiern

## Bemerkungen

- Es gibt immer eine Indikatorvariable weniger, als es Levels hat
- Level ohne Indikatorvariable (hier Afroamerikaner): *Baseline*
- Folgende Gleichung macht *keinen* Sinn:

$$y_i = \beta_0 + \beta_1 + \beta_2 + \varepsilon_i$$

► Person müsste asiatisch *und* kaukasisch sein

- Output: Geschätzte **balance** \$ 531.00 für Baseline (Afroamerikaner):

```
balance <- Credit[, "Balance"]
ethnicity <- Credit[, "Ethnicity"]
summary(lm(balance ~ ethnicity))

##
Call:
lm(formula = balance ~ ethnicity)
##
Residuals:
Min 1Q Median 3Q Max
-531.00 -457.08 -63.25 339.25 1480.50
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 531.00 46.32 11.464 <2e-16 ***
ethnicityAsian -18.69 65.02 -0.287 0.774
ethnicityCaucasian -12.50 56.68 -0.221 0.826

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared: 0.0002188, Adjusted R-squared: -0.004818
F-statistic: 0.04344 on 2 and 397 DF, p-value: 0.9575
```

- Schätzung für Kategorie Asiaten: \$ -18.69
- Durchschn. Rechnungen um diesen Betrag kleiner als die von Afroamerikanern
- Kaukasier haben um durchschn. \$ 12.50 kleinere Rechnungen als die Afroamerikaner
- $p$ -Werte gross → Zufällige Abweichungen
- Kein signifikanter Unterschied bei den Kreditkartenrechnungen zwischen den Ethnien
- Level, für Baseline willkürlich
- Vorhersage der Zielvariable hängt nicht von der Kodierung ab
  - $p$ -Werte hängen von der Kodierung ab
  - $F$ -Statistik betrachten
  - $F$ -Test und testen
$$H_0 : \beta_1 = \beta_2 = 0$$
  - $p$ -Wert dieser Statistik hängt *nicht* von der Kodierung ab
  - $p$ -Wert 0.96 → Relativ hoch
  - Vermutung bestätigt: Nullhypothese *nicht* verwerfen
  - Es gibt keinen Zusammenhang zwischen `balance` und `ethnicity`

- Indikatorvariablen: Qualitative *und* quantitative erklärende Variablen in Regressionsmodell integrieren
- Regression von **Balance** mit quantitativer erklärenden Variable **Income** und qualitativer erklärenden Variable **student** durchführen
- **Student** mit Indikatorvariablen
- Multiple lineare Regression

## Beispiel: Datensatz Credit

- Zielgröße **Balance** durch die erklärenden Variablen **Income** (quantitativ) und **Student** (qualitativ) vorhersagen
- Ohne Interaktionsterm:

$$\text{balance}_i \approx \beta_0 + \beta_1 \cdot \text{income}_i + \begin{cases} \beta_2 & \text{falls } i\text{-te Person Student} \\ 0 & \text{falls } i\text{-te Person kein Student} \end{cases}$$

$$= \beta_1 \cdot \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{falls } i\text{-te Person Student} \\ \beta_0 & \text{falls } i\text{-te Person kein Student} \end{cases}$$

- Output:

```

student <- Credit[, "Student"]
income <- Credit[, "Income"]
summary(lm(balance ~ income + student))

##
Call:
lm(formula = balance ~ income + student)
##
Residuals:
Min 1Q Median 3Q Max
-762.37 -331.38 -45.04 323.60 818.28
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.1430 32.4572 6.505 2.34e-10 ***
income 5.9843 0.5566 10.751 < 2e-16 ***
studentYes 382.6705 65.3108 5.859 9.78e-09 ***

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 391.8 on 397 degrees of freedom
Multiple R-squared: 0.2775, Adjusted R-squared: 0.2738
F-statistic: 76.22 on 2 and 397 DF, p-value: < 2.2e-16

```

- $\hat{\beta}_0$ :

Ohne Einkommen und als Nichtstudent zahlt man \$211 monatliche Kreditkartenrechnung

- $\hat{\beta}_1$ :

Pro \$1000 Einkommen mehr, zahlt man \$6 mehr Kreditkartenrechnung (unabhängig vom Studentenstatus)

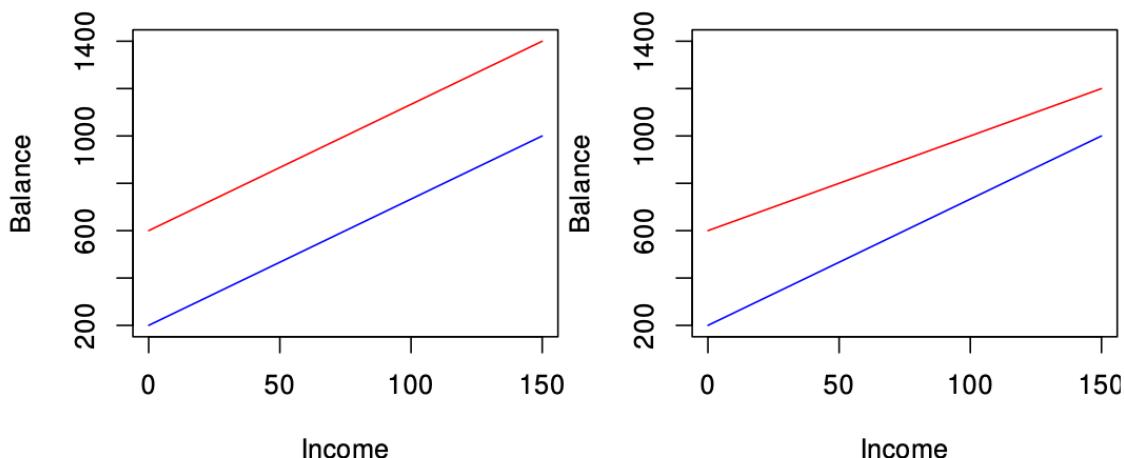
- $\hat{\beta}_2$ :

Studierende zahlen \$383 mehr Kreditkartenrechnung als Nichtstudierende (unabhängig vom Einkommen)

- Modell beschreibt zwei parallele Geraden: eine für Studierende und eine für Nichtstudierende

- ▶ Steigung  $\beta_1$  ist bei beiden gleich
- ▶  $y$ -Achsenabschnitte sind verschieden ( $\beta_0 + \beta_2$  und  $\beta_0$ )

- Abbildung links:



- Durchschn. Zunahme von **Balance** für Vergrösserung von **Income** um eine Einheit hängt nicht davon ab, ob entsprechendes Individuum studiert oder nicht
- Mögliche Einschränkung des Modells: Änderung in **Income** kann eine unterschiedliche Wirkung auf Rechnungen haben kann, ob jemand studiert oder nicht
- Lockerung dieser Einschränkung: Einführung einer Interaktionsvariablen
- **Income** wird mit der Indikatorvariablen für **Student** „multipliziert“

- Modell:

$$\begin{aligned}\text{balance}_i &\approx \beta_0 + \beta_1 \cdot \text{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \text{income}_i; & \text{falls studierend} \\ 0 & \text{falls nicht studierend} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{income}_i; & \text{falls studierend} \\ \beta_0 + \beta_1 \cdot \text{income}_i & \text{falls nicht studierend} \end{cases}\end{aligned}$$

- Zwei unterschiedliche Regressionsgeraden für Studierende und Nichtstudierende (Abbildung oben rechts):
  - ▶ Verschiedene Steigungen  $\beta_1 + \beta_3$  und  $\beta_1$
  - ▶ Unterschiedliche  $y$ -Achsenabschnitte  $\beta_0 + \beta_2$  und  $\beta_0$
- Möglichkeit, Änderung der Zielgröße (Kreditkartenrechnungen) aufgrund der Änderungen im Einkommen für Studenten und Nichtstudenten getrennt zu betrachten
- Rechte Seite von Abbildung oben: Geschätzter Zusammenhang zwischen **income** und **balance** für Studierende (rot) und Nichtstudierende (blau)
- Steigung für Studierende ist grösser als für Nichtstudierende
- Deutet an: Zunahme im Einkommen eines Studierenden eine grösse Zunahme der Kreditkartenrechnungen zur Folge hat als für Nichtstudierenden

- Output:

```
summary(lm(balance ~ income * student))

##
Call:
lm(formula = balance ~ income * student)
##
Residuals:
Min 1Q Median 3Q Max
-773.39 -325.70 -41.13 321.65 814.04
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 200.6232 33.6984 5.953 5.79e-09 ***
income 6.2182 0.5921 10.502 < 2e-16 ***
studentYes 476.6758 104.3512 4.568 6.59e-06 ***
income:studentYes -1.9992 1.7313 -1.155 0.249

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared: 0.2799, Adjusted R-squared: 0.2744
F-statistic: 51.3 on 3 and 396 DF, p-value: < 2.2e-16
```

- $p$ -Wert der Interaktion ist statistisch nicht signifikant
- Somit gibt es keine Interaktion
- Steigungen der beiden Geraden sind nicht signifikant unterschiedlich