APPLIED STATISTICS FOR DATA SCIENCE

SW01 - Einführung
SW02 - Deskriptive Statistik (eindimensional)
SW03 - Histogramm, zweidimensionale deskriptive Statistik
SW04 - Korrelation, Wahrscheinlichkeitstheorie
SW05 - Zufallsvariable, Wahrscheinlichkeitsverteilung
SW06 - Bedingte Wahrscheinlichkeit
SW07 - Normalverteilung
SW08 - Gesetz der grossen Zahlen, zentraler Grenzwertsatz
SW09 - Hypothesentest, z-Test, t-Test
SW10 - Vertrauensintervall, Zweistichprobentest, Wilcoxon-Test
SW11 - Lineare Regression
SW12 - Multiple lineare Regression
SW13 - Variablenselektion (NUR Qualitative Variablen)

SW01 – Einführung

- **Angewandte Statistik:** Anwendung der Statistik auf reale Alltagsprobleme.
- Relative Fallzahlen sind nicht immer besser oder aussagekräftiger als absolute.

Problemlösung in der angewandten Statistik

- Beispiel Tweet von Trump*
- Enthält aber viele Aspekte, die für die Lösung von Problem in der angewandten Statistik relevant sind.

1. Unklar, was die Frage oder das Problem ist.
- Begonnen mit dem Trump Tweet.
- Und jetzt?
- CNN, im allgemeinen Anti-Trump, hielt es für erwähnenswert.

2. Es ist nicht klar, wie die Lösung aussehen soll.
- Wie reagiert man auf einen solchen Tweet?
- CNN: Argumentation auf der Basis von Fallzahlen

3. Es ist nicht klar, wie das Ergebnis zu interpretieren ist.
- CNN beschloss, absolute Fallzahlen anstelle der relativen Fallzahlen zu verwenden.

4. Es ist nicht klar, wie das Ergebnis zu interpretieren ist.
- Häufig der schwierigste Punkt zur Lösung von Problemen in der angewandten Statistik.
- CNN ging in die falsche Richtung: Vergleich Dinge, die nicht vergleichbar sind.

Alltagsprobleme

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

- Für all diese Probleme gibt es keine Lösungsanweisung.
- Kann auch nicht durch Formeln gelöst werden.
- Vier-Schritte Problemlösungsstrategie aus Mazur.
- Die meisten statistischen Probleme sind in Worten formuliert.

Vier-Schritte-Problemlösungsstrategie aus Mazur

1. Erste Schritte
2. Plan erstellen
3. Plan ausführen
4. Resultat interpretieren

Was ist Statistik, die nicht angewandt ist?

- **Angewandte Statistik:** Verfahren und Methoden werden verwendet und beschrieben.
- Kann oft auf einfache Weise erklärt werden.
- Aber was nicht getan wird, warum diese Verfahren und Methoden genau das tun, was sie tun sollten.
- Obwohl das Prinzip oft einfach ist, sind die Details schwierig.

1. Schritt: Erste Schritte (V.S.P.A. Mazur)

- Es ist nicht klar, welches der effizienteste Weg ist, um auf ein bestimmtes Problem zu reagieren.
- Der erste Schritt unserer Problemlösungsstrategie, der Anfang, ist oft der schwierigste
- Es ist daher sinnvoll, mit etwas zu beginnen, das Sie tun können.

1. Organisieren Sie die gegebenen Informationen und klären Sie, was genau das Problem erfordert.
2. Stellen Sie sicher, dass Sie alle enthaltenen Informationen verstehen.
3. Formulieren Sie das Problem in Ihren eigenen Worten.
4. Prüfen Sie, ob Sie alle notwendigen Informationen zur Lösung des Problems haben.

2. Schritt: Plan erstellen (V.S.P.A. Mazur)

- Der nächste Schritt besteht darin, einen Lösungsplan für Ihr Problem zu erstellen.
- Überlegen Sie, welche Schritte erforderlich sind, um das Problem zu lösen.
- Ein Guter Plan ist es, die Schritte aufzuzeigen, die Sie ergreifen müssen, um eine Lösung zu finden.

3. Schritt: Plan ausführen (V.S.P.A. Mazur)

- Sie führen Ihren Plan aus, indem Sie die von Ihnen umrissenen Schritte befolgen.

4. Schritt: Resultat interpretieren (V.S.P.A. Mazur)

- Letzter und entscheidender Schritt: **Interpretieren Sie Ihre Antwort!**
- Überprüfen Sie, ob Ihr Ergebnis überhaupt möglich ist.
  - Wenn beispielsweise eine Wahrscheinlichkeit negativ wird, ist etwas schief gelaufen.
- Interpretieren Sie das Ergebnis in Bezug auf das ursprüngliche Problem.

SW02 – Deskriptive Statistik (eindimensional)

Daten

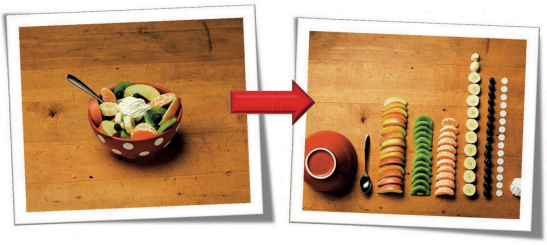
- Daten und Statistiken bestimmen immer mehr unser Leben
- Zeitung: Prognose zur nächsten Abstimmung oder Wahlen –> Befragung.
- Googeln: Wie „weiss" Google so genau, was man suchen will? –> Google wertet Suchanfragen aus.
- Passkontrolle am Flughafen: Wie erkennt die Software Gesichter? –> Gesichter werden charakterisiert.
- Wetterbericht: Wie kommt die Vorhersage zustande? –> Modell aufgrund früherer Wetterdaten (und Theorie)
- Börsenkurse: Wie lässt sich aus dem Börsenverlauf der letzten paar Tage, der Kurs für die nächsten paar Tage vorhersagen?

Deskriptive Statistik

- **Deskriptive Statistik:** Darstellung von Datensätzen.
- Datensätze
  - Durch gewisse Zahlen charakterisieren (z.B. Mittelwert).
  - und graphisch darstellen.
- D.S. beschäftigt sich damit, auf welche Weisen Daten organisiert und zusammengefasst werden können.

Ziele der deskriptiven Statistik

- Daten zusammenfassen durch numerische Kennwerte
- Graphische Darstellung der Daten



Datensätze (eindimensional)

- Liste: einfachste Variante eines Datensatzes.
- Bsp: Körpergrössen von 5 Personen.
  - 1.75, 1.80, 1.72, 1.65, 1.54
- Solche Listen heissen: **eindimensionale Datensätze oder Messreihen**

Datensätze (zweidimensional)

- Häufigste Form von Datensätzen: **Tabellen oder zweidimensionale Datensätze.**
- Grösse und Gewicht: Quantitative Daten, also (gemessene Zahlen)
- Können zumindest theoretisch, jeden beliebigen Zahlenwert in einem Bereich annehmen.
- Geschlecht und Nationalität: Qualitative Daten
- Nehmen nur bestimmte Anzahl Werte an (müssen keine Zahlen sein).

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

Beispieldatensatz

- Messungen Körpergewicht.
- Erfahrung: Steht am Morgen auf Waage und merkt sich Gewicht.
- Steht nochmals auf die Waage und erhält leicht anderes Resultat.
- Wir wollen es genauer wissen.
- Nehmen 80 Kg schweren Metallblock, der geeicht ist, d.h. er hat mit sehr grosser Genauigkeit 80 Kg.
- Gewicht dieses Metallblocks wird mehrere Male mit zwei Waagen A und B gemessen.
- Zwei Datensätze mit Gewichten (in Kg; auf 10 g genau messen)

Warum führen verschiedene Objektmessungen, zu unterschiedl. Resultaten?

- Messungen finden nie unter exakt denselben Bedingungen statt.
- Scheinbar genaue Angaben sind nur <u>ungefähre</u> Angaben (Kalorienanzahl auf einer Packung, Inhalt 500ml Pet-Flasche: Keine zwei Pet-Flaschen sind absolut gleich usw.)

Bezeichnungen

- Standardbezeichnung von Daten mit: $x_1, x_2, \ldots, x_n$
- **n**: Umfang der Messreihe (Daten, Datensatz).
- Beispiel: Messreihe der Waage A hat Umfang n = 13:

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \ldots, \quad x_{13} = 80.02$$

Überblick über die Kennzahlen

- Bekannt: **n** beobachtete Datenpunkte (Messungen).

  $x_1, x_2, \ldots, x_n$

  (z.B. Verkehrsaufkommen an n verschiedenen Tagen).
- Unterscheidung zwischen Lage- und Streuungsparametern.
- **Lageparameter** („Wo liegen die Beobachtungen auf der Mess-Skala?")
  - Arithmetisches Mittel (Durchschnitt)
  - Median
  - Quantile
- **Streuungsparameter** („Wie streuen die Daten um ihre mittlere Lage?")
  - Empirische Varianz / Standardabweichung
  - Quartilsdifferenz

Zurück zum Beispiel der Waagen

- Messungen wurden mit grösstmöglichen Sorgfalt durchgeführt.
- Trotzdem variieren die Messwerte innerhalb beider Waagen
- Es stellen sich hier nun die folgenden Fragen:
  - Gibt es einen Unterschied zwischen der Waage A und B?
  - Falls ja, wie können wir diesen Unterschied ermitteln?
- Es fällt auf: Werte der Waage A sind also <u>eher</u> grösser als die der Waage B
- Was heisst hier aber eher?
- Wie kann man die beiden Messreihen miteinander vergleichen?
- Ziele:
  - Messreihen irgendwie zusammenzufassen, um die beiden miteinander vergleichen zu können.
  - Interpretation und darauffolgende statistische Analyse dieser Daten vereinfachen.
- Kennzahlen sollen Daten numerisch zusammenfassen und grob charakterisieren
- Bei statistischen Analysen ist es sehr wichtig, nicht einfach blind ein Modell anzupassen oder ein statistisches Verfahren anzuwenden.
- Daten sollten immer mit Hilfe von geeigneten graphischen Mitteln und den Kennzahlen dargestellt werden.
- Nur auf diese Weise kann man (teils unerwartete) Strukturen und Besonderheiten entdecken.
- Aber WARNUNG!!!: Wann immer ein Datensatz „reduziert" wird (durch Kennzahlen oder Graphiken), geht Informationen verloren!
- - Beispiel: Notendurchschnitt einer Klasse (4.51). Dieser Wert sagt über Klasse als Ganzes etwas aus, aber nichts mehr über die einzelnen Lernenden. Kennen nur Zahl 4.51: Keine Information mehr, wie die einzelnen Lernenden abgeschnitten haben. Wissen nicht einmal, wieviele Lernende in der Klasse sind.

Arithmetisches Mittel (Durchschnitt)

- Addiert alle Daten und teilt Summe durch Anzahl Daten (Umfang)
- Sprechweise: „x quer“
- Definition:

Arithmetisches Mittel

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

- Beispiel Waage A: Arithmetische Mittel der $n = 13$ Messungen

$$\overline{x} = \frac{79.98 + 80.04 + \ldots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.020\,77$$

- R-Befehl für arithmetisches Mittel `mean(...)`:

```
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
```

- Arithmetisches Mittel: Anschaulich

Vergleich der Waagen

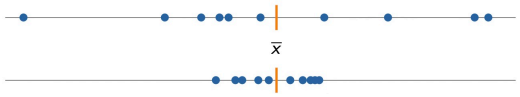
- Waagen mit den arith. Mittel miteinander vergleichen:

```
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)

mean(waageB)
## [1] 79.97875
```

- Waage B hat durchsch. die tieferen Werte als die Waage A
- Waagen vergleichbar, obwohl verschiedener Umfang.

Streuung graphisch

- Arithm. Mittel: „Wo ist die „Mitte“ der Daten?“
- Aber: Arith. Mittel sagt nicht alles über Messreihe aus.
- Graphisches Beispiel:



- Beide Datenreihen haben dasselbe arith. Mittel.
- Punkte 2. Datenreihe: Durchschnittlich viel näher beim Mittelwert x quer als Punkte der 1. Datenreihe.
- Verschiedene Streuung der Daten um den Durchschnitt

Streuung numerisch

- Beispiel von Schulnoten: 2, 6, 3, 5 und 4, 4, 4, 4
- Beide Mittelwert 4, aber Verteilung der Daten um Mittelwert ziemlich unterschiedlich
  –> 1. Fall: Zwei gute und zwei schlechte Lernende
  –> 2. Fall: Alle Lernende gleich gut
- Datensätze haben eine verschiedene Streuung um den Mittelwert
- 2. Klasse: Keine Streuung (Streuung 0)
- Wie kann man diese Streuung mathematisch beschreiben?
- Suchen also Mass für den Unterschied zum Mittelwert.
- 1. Idee: Durchschnitt der Unterschiede zum Mittelwert
  - 1. Fall:

$$\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4} = \frac{-2+2-1+1}{4} = \frac{0}{4} = 0$$

  - Zweiter Fall auch 0:

$$\frac{(4-4)+(4-4)+(4-4)+(4-4)}{4} = \frac{0+0+0+0}{4} = \frac{0}{4} = 0$$

- Dieser Ausdruck macht keine Aussage über unterschiedliche Streuung
- Problem: Unterschiede können negativ werden, können sich aufheben.
- Nächste Idee: Unterschiede durch die Absolutwerte ersetzen
  - 1. Fall:

$$\frac{|(2-4)|+|(6-4)|+|(3-4)|+|(5-4)|}{4} = \frac{2+2+1+1}{4} = 1.5$$

  - D.h.: Noten weichen im Schnitt 1.5 vom Mittelwert ab
  - 2. Fall: Dieser Wert natürlich auch 0

$$\frac{|(4-4)|+|(4-4)|+|(4-4)|+|(4-4)|}{4} = \frac{0+0+0+0}{4} = 0$$

- Je grösser dieser Wert (immer >= 0), desto mehr unterscheiden sich die Daten bei gleichem Mittelwert voneinander.
- Dieser Wert für die Streuung: Mittlere absolute Abweichung
- Aber: Theoretische Nachteile

Besser: Empirische Varianz und Standardabweichung

- Mass für Variabilität oder Streuung der Messwerte

- Definition:

Empirische Varianz $var(x)$ und Standardabweichung s_X

$$\mathrm{Var}(x) = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

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

Beispiel Waage A (von vorher)

- Arith. Mittel der $n = 13$ Messungen: $\overline{x} = 80.02$ (siehe Slide 16)
- Empirische Varianz:

$$\mathrm{Var}(x) = \frac{(79.98-80.02)^2+(80.04-80.02)^2+\cdots+(80.00-80.02)^2+(80.02-80.02)^2}{13-1}$$
$$= 0.000\,574$$

- Empirische Standardabweichung:

$$s_x = \sqrt{0.000574} = 0.024$$

- D.h.: „Mittlere" Abweichung vom Mittelwert 80.02 kg ist 0.024 kg

- Von Hand sehr mühsam
- sd: Standard deviation

- Mit R:

```
var(waageA)
## [1] 0.000574359
sd(waageA)
## [1] 0.02396579
```

Eigenschaften der Varianz

- Bei Varianz: Abweichungen $x_i - \overline{x}$ quadrieren , damit sich Abweichungen nicht gegenseitig aufheben können

- Nenner n - 1, anstelle von n: Mathematisch begründet.

- In einigen Büchern: n anstatt n - 1: Spielt für grosse n keine Rolle (Siehe Jupyter Notebook varianz_n.ipynb)

- Standardabweichung ist die Wurzel der Varianz

- Durch Wurzelziehen wieder dieselbe Einheit wie bei den Daten selbst.
  –> Daten in cm
  –> Varianz wegen der Quadrierung der Daten in cm^2
  –> Einheit der ursprünglichen Daten: Aus Varianz Wurzel ziehen.

- Ist empirische Varianz (und damit die Standardabweichung) **gross**, so ist die Streuung der Messwerte um das arithmetische Mittel **gross**

- Wert der empirischen Varianz hat keine physikalische Bedeutung.

- Man weiss nur, je grösser der Wert umso grösser die Streeung

- **Wichtig**: Nur die Standardabweichung s_x lässt sich konkret interpretieren

- Für normalverteilte Daten hat die Standardabweichung noch eine schöne geometrische Interpretation.

Median

- Ein weiteres Lagemass für die „Mitte": Median

- Sehr vereinfacht: Wert, bei dem die Hälfte der Messwerte unter oder gleich diesem Wert sind.

- Andere Hälfte ist gleich diesem Messwert oder darüber.

- Beispiel: Prüfung in der Schule ist Median 4.6
  –> Hälfte der Klasse hat diese Note oder ist schlechter
  –> Umgekehrt: Andere Hälfte der Klasse hat diese Note oder ist besser

- Obige Interpretation: Medians sehr einfach dargestellt, exakte Definition folgt nun.

Geordnete Strichprobe

- Datensatz in aufsteigender Reihenfolge ordnen.

- Bestimmung Median: Daten zuerst der Grösse nach ordnen.

- Bezeichnung der *geordneten Daten* mit $x_{(i)}$:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

- Runde Klammern im Index: Daten geordnet

Beispiel Waage A (Median)

- Daten Waage A geordnet:
- Median ist nun sehr einfach zu bestimmen.
- Unter diesen 13 Messung: Wert der mittleren Beobachtung.
- Wert der 7. Beobachtung:

79.97, 79.98, 80.00; 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Median des Datensatzes der Waage A ist 80.03.
- D.h.: Knapp die Hälfte der Messwerte, nämlich 6 Beobachtungen sind kleiner oder gleich 80.03.
- Ebenso sind 6 Messwerte grösser oder gleich dem Median.
- Ungerade Anzahl Messungen: GENAU eine mittlere Messung.

Beispiel Waage B (Median)

- Vorher: Anzahl der Daten ungerade und damit ist die mittlere Beobachtung eindeutig bestimmt.
- Anzahl Daten gerade: Keine mittlere Beobachtung
- Definition Median: Mittelwert der beiden mittleren Beobachtungen
- Beispiel Datensatz der Waage B hat 8 Beobachtungen
- Ordnen den Datensatz: Median Durchschnitt von der 4. und 5. Beobachtung.

79.94, 79.95, 79.97, 79.97, 79.97, 79.94, 80.02, 80.03

$$\frac{79.97 + 79.97}{2} = 79.97$$

Beispiele mit R–Befehl

- R–Befehl:

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

median(waageB)
## [1] 79.97
```

- Als Median kann Wert auftreten, der in Messreihe nicht ist.

- Beispiel: Median Schulklasse:

```
noten <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7,
           5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1)

median(noten)
## [1] 4.65
```

- ABER: Noten auf Zehntelnoten genau

Median vs. arithmetisches Mittel

- Zwei Lagemasse für die Mitte eines Datensatzes
- Welches ist nun „besser"? Dies kann man so nicht sagen:
  –> Kommt auf die jeweilige Problemstellung an.
  –> Am besten werden beide Masse gleichzeitig betrachtet.
- Eigenschaft des Medians: Robustheit
  –> Wird viel weniger stark durch extreme Beobachtungen beeinflusst als das arithmetische Mittel

- Beispiel: Bei der grössten Beobachtung ($x_9 = 80.05$) ist ein Tippfehler passiert und $x_9^* = 800.5$ eingegeben worden
- Arithmetische Mittel ist dann anstatt 80.02:

$$\overline{x}^* = 135.44$$

- Median ist aber nach wie vor

$$x^*_{(7)} = x_{(7)} = 80.03$$

- Arithmetisches Mittel: Durch Veränderung *einer* Beobachtung sehr stark beeinflusst
- Median bleibt hier gleich: Robust

Graphisch

Weiteres Beispiel zu Median

- Untersuchen typisches Haushaltseinkommen von Vororten von Seattle um Lake Washington.
- Durchschnittliches Einkommen von Medina und Windermere wird sehr unterschiedlich sein.
- Grund: Bill Gates lebt in Medina
- Grundsätzlich: Für das mittlere Einkommen wird praktisch immer der Median genommen und nicht der Durchschnitt, da dies gerichter ist.
- Zügelt Bill Gates von medina nach Windermere, so ändert sich das Durchschnitteinkommen, aber niemand hat etwas davon.

Quartile

- Median: Wert, wo die Hälfte der Beobachtungen kleiner oder gleich wie Median ist.
- Analoge Überlegung: Unteres und oberes Quartil.
- Unteres Quartil: Wert, wo 25% aller Beobachtungen kleiner oder gleich und 75% grösser oder gleich sind wie dieser Wert.
- Oberes Quartil: Wert, wo 75% aller Beobachtungen kleiner oder gleich und 25% grösser oder gleich wie dieser Wert sind.
- Achtung: Meist gibt es nicht exakt 25% der Beobachtungen.
- Man definiert Wert für das untere Quartil bzw. obere Quartil.

Bermerkungen Quartile

- Hier jeweils aufgerundet, falls 25% bzw. der 75% der Anzahl Beobachtungen nicht ganz ist.
- Hätten auch abrunden können: Andere Werte für die Quartile
- Für grosse Datensätze: Spielt praktisch keine Rolle, ob aufgerundet, abgerundet oder gerundet wird.
- Aber: Es gibt keine einheitliche Definition für die Quartile.

Median Bemerkung

- Median: Auch Zentralwert oder mittlerer Wert (nicht zu verwechseln mit Mittelwert) genannt.
- Exakte Interpretation des Medians ist noch erstaunlich schwierig.
- Für uns ausreichend: Hälfte der Werte kleiner oder gleich und die andere Hälfte grösser oder gleich dem Median.

Beispiel: Waage (Quartile)

- Waage A hat n = 13 Messpunkte: 25% davon ist 3.25
  - Man *wählt* nächstgrösseren Wert $x_{(4)}$ als unteres Quartil:

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Unteres Quartil ist 80.02
- Knapp ein Viertel der Messwerte ist gleich oder kleiner 80.02

  - Oberes Quartil: *Wählen* $x_{(10)}$, da für $0.75 \cdot 13 = 9.75$ die Zahl 10 der nächsthöhere Wert ist

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Knapp drei Viertel der Messwerte sind kleiner oder gleich 80.04

- Waage B: 25% der Werte 2: ganze Zahl.
- Man wählt Durchschnitt von $x_{(2)}$ und $x_{(3)}$ als unteres Quartil.
- Dann sind 2 Beobachtungen kleiner und 6 Beobachtungen grösser dieser Wert

79.94, 79.95, 79.97, 79.97, 79.97, 79.94, 80.02, 80.03

$$\frac{79.95 + 79.97}{2} = 79.96$$

- Das untere Quartil der Methode B ist also 79.96

Quartile in R

- Software R kennt keine eigenen Befehle für die Quartile
- Allgemeiner Befehl **quantile** (Quantile kommen gleich)
- R: Quartile nach unserer Definition: Option type=2

```
# 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 Beispiel

- Schulklasse mit 24 Lernenden: Noten an einer Prüfung

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

- Berechnen mit R die Quartile und die Quartilsdifferenz:

```
quantile(noten, p = c(0.25, 0.75), type = 2)
##  25%  75%
## 3.80 5.35
IQR(noten, type = 2)
## [1] 1.55
```

- Hälfte der Lernenden liegen innerhalb von 1.55 Noten, nämlich zwischen 3.8 und 5.35
- 25% der Klasse 3.8 oder weniger; rund 25% der Klasse 5.35 und mehr

Quantile Beispiel

- Noten an einer Prüfung in Schulklasse mit 24 Lernenden

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

- Verschiedene Quantile mit R:

```
quantile(noten, p = seq(from = 0.2, to = 1, by = 0.2), type = 2)
##  20%  40%  60%  80% 100%
##  3.6  4.2  5.0  5.6  6.0
```

- D.h.: Knapp 20% der Lernenden sind schlechter als 3.6
- Genau 20% der Lernenden nicht möglich: 4.8 Lernende
- 60%-Quantil: (Knapp) diese Anzahl Prozent der Lernenden waren schlechter oder gleich einer 5.

Quartilsdifferenz

- Quartilsdifferenz ist ein Streuungsmass für die Daten
  oberes Quartil – Unteres Quartil
- Es misst die Länge des Intervalls, das etwa die Hälfte der „mittleren“ Beobachtungen enthält.
- Je kleiner dieses mass, umso näher liegt die Hälfte aller Werte um den Median und umso kleiner ist die Streuung.
- Dieses Streuungsmass ist robust.
- Quartilsdifferenz der Methode A
  80.04 – 80.02 = 0.02

- R: Quartilsdifferenz (interquartile range) IQR

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

- Die (oder ungefähr die) Hälfte der Messwerte liegt also in einem Bereich der Länge 0.02
- Beim Boxplot: Anschauliche Interpretation der Quartilsdifferenz

Quantile

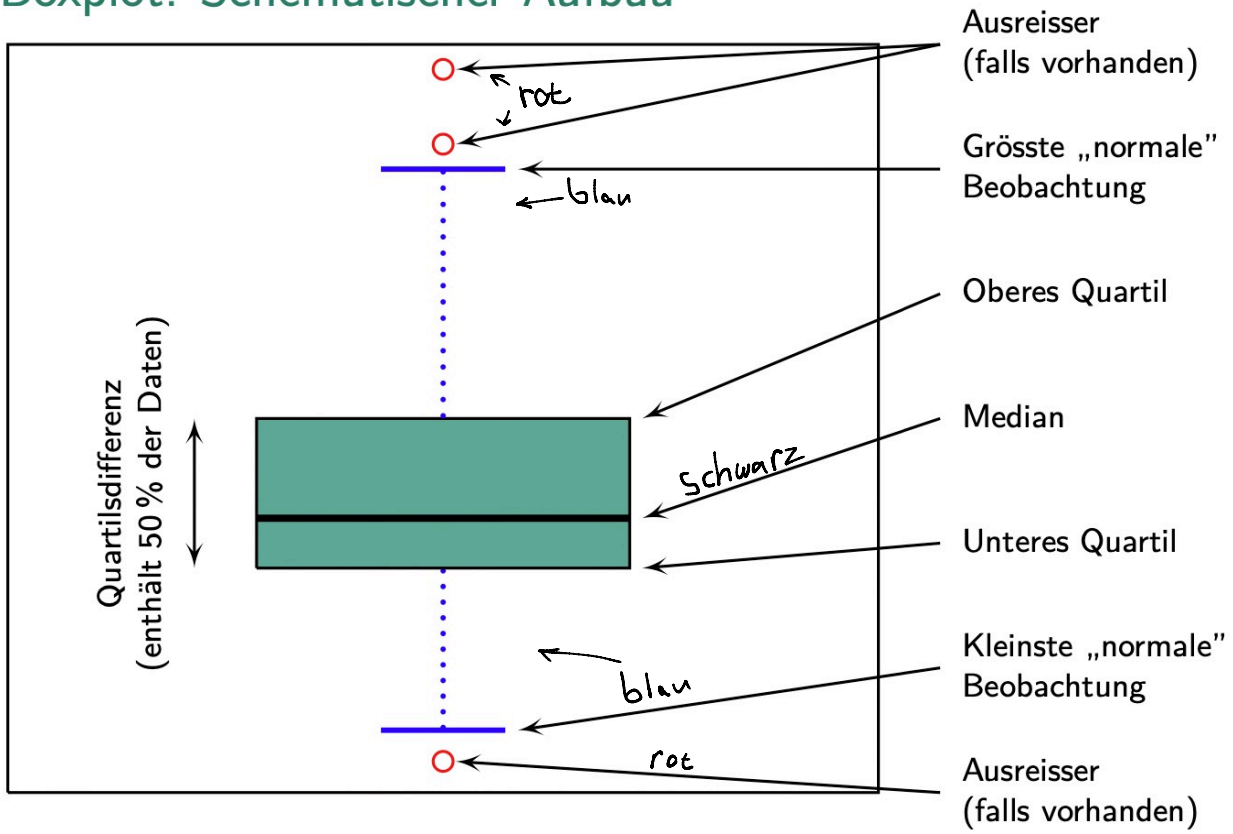
- Quartile auf jede andere Prozentzahl verallgemeinern: Quantile
- 10%-Quantil: Wert, wo 10% der Werte kleiner oder gleich und 90% der Werte grösser oder gleich diesem Wert sind.
- Definition analog wie bei Quartilen
- Median ist 50%-Quantil
- 25%-Quantil ist unteres Quartil
- 75%-Quantil ist oberes Quartil

- R: 10 %– und 70 %–Quantil der Waage A:

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

- Knapp 10% der Messwerte sind kleiner oder gleich 79.97
- Entsprechend: Knapp 70% der Messwerte kleiner oder gleich 80.04

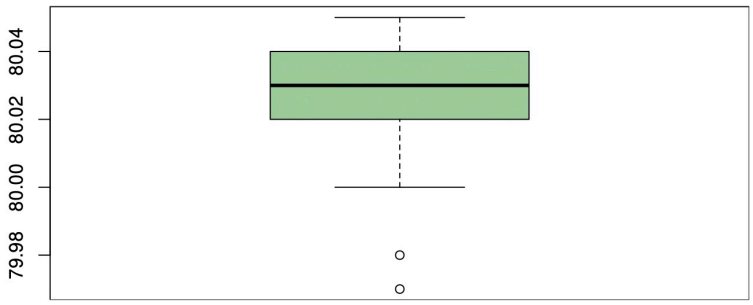
Boxplot: Schematischer Aufbau



- Rechteck, dessen Höhe vom empirischen 25%- und vom 75%-Quantil begrenzt wird: Box
- Horizontaler Strich für den Median in Box (Schwarz)
- Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten „normalen" Wert führen (blau eingezeichnet)
  –> Definition: „Normaler" Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt.
- Ausreisser: Kleine Kreise (rot)

R

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

)
```

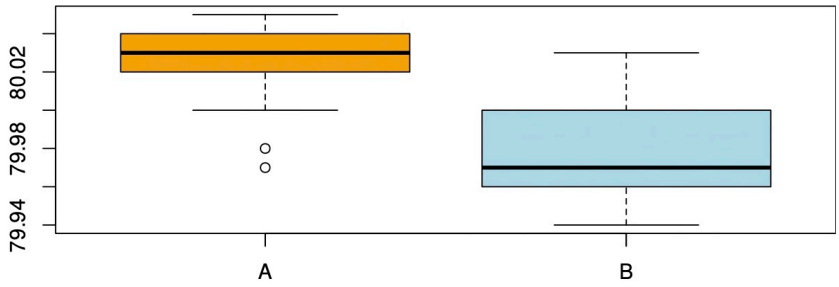


- Hälfte der Beobachtungen befindet sich zwischen dem oberen Quartil 80.04 und dem unteren Quartil 80.02, mit Quartilsdifferenz 0.02
- Median liegt bei 80.03
- Der „normale" Bereich der Werte liegt zwischen 80.00 und 80.05
- Zwei Ausreisser: 79.97 und 79.98
- Erste beiden Punkte: Früher berechnet
- Boxplot: Graphische Darstellung von Median und Quartilen

Vergleich von Datensätzen

- Boxplot: Darstellungen von verschiedenen Gruppen

```
boxplot(waageA, waageB,
        xlab = "Waage",
        col = c("orange", "lightblue")

)
axis(side = 1, at = c(1, 2), labels = c("A", "B"))
```



- Waage A grössere Werte als Waage B: Median von A grösser ist.
- Daten von Waage A haben weniger Streuung als die Daten von Waage B: Rechteck weniger hoch (Quartilsdifferenz!)
- R-Code axis(...)

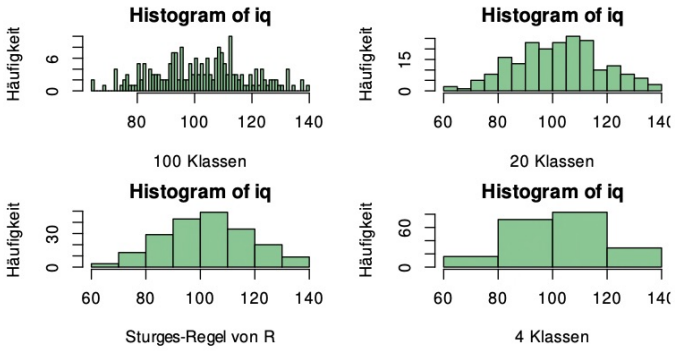
SW03 – 1. Histogramm, 2. zweidimensionale deskriptive Statistik

1. Histogramme

- Definition: Graphischer überblick über die auftretenden Werte.
- Aufteilung des Wertebereichs in k Klassen (Intervalle)
- Faustregel:
  –> Bei weniger als 50 Messungen ist die Klassenzahl 5-7
  –> Bei mehr als 250 Messungen wählt man 10-20 Klassen
- Zeichne für jede Klasse einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist.

Wahl der Klassen

- Wahl der Anzahl Klassen relevant für Aussagekraft eines Histogrammes
- Es gibt keine allgemeine Grundregel, wie man Anzahl Klassen wählt
- Abbildung: IQ Daten von Beispiel mit verschiedener Anzahl Klassen

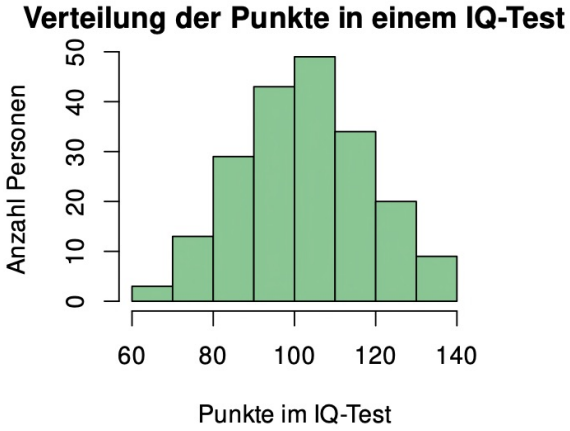


- Histogramm links oben: Viel zu detailliert, als dass man ein Muster erkennen könnte.
- Histogramm rechts unten zu ungenau.
- R: Anzahl Klassen nach der sogennanten Sturges-Regel.

Siehe nächste Seite für Code!

Beispiel: IQ-Test

- Abbildung: Histogramm vom Ergebnis eines IQ-Testes von 200 Personen



- Daten wurden hier allerdings simuliert.
- Breite der Klassen: 10 IQ-Punkte; für jede Klasse gleich.
- Höhe der Balken: Anzahl Personen, die in diese Klasse fallen.
- Bsp: ca. 20 Personen fallen in Klasse zwischen 120 und 130 Punkten.
- Form dieses Histogramm typisch für viele Histogramme –> Normalverteilungl

- Code: Der R-Code für das Histogramm oben lautet wie folgt:

```
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"
)
```

- Befehl `rnorm(n = 200, mean = 100, sd = 15)`: Wählt zufällig 200 normalverteilte Daten (siehe Kapitel Normalverteilung) mit Mittelwert 100 mit Standardabweichung 15 aus
- Befehl `hist(iq, ...)`: Histogramm für die Daten `iq`
- Die weiteren Optionen sollten klar sein:
  - `xlab` steht für x-Label, die Beschriftung der x-Achse
  - `ylab` steht für y-Label, die Beschriftung der y-Achse
  - `col` steht für color
  - `main` steht für Haupttitel

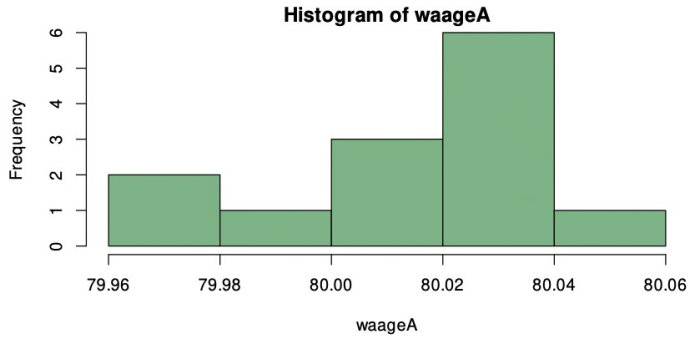
Code (Wahl der Klassen)

```
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"
)
```

- Befehl `par(mfrow = c(2, 2))`: Die vier Histogramme in 2 Zeilen (erste 2) und 2 Spalten gezeichnet (zweite 2)
- Option `breaks = ...`: Anzahl Klassen festlegen
- Beachte letzte Graphik: `breaks = 3`, aber vier Klassen gezeichnet
- `R` nimmt die Option `breaks = ...` nur als Vorschlag

R: hist()

```
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")
```



Bemerkungen

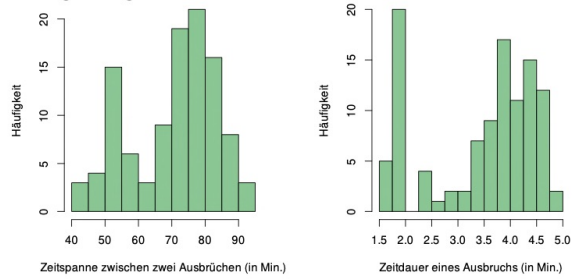
- Waage *A* 13 Messungen → 5 Balken
- `R`: Interne Regel (Sturges Regel)
- Bedeutung der Anzahlen (Frequency):
  - In 1. Klasse (79.96-79.98) sind die Beobachtungen mit den Werten 79.97 und 79.98 berücksichtigt
  - in der 2. Klasse 79.99 und 80.00; usw.
- Linke Grenze wird also *nicht* berücksichtigt, die rechte schon
- Umgekehrt auch möglich → Histogramm würde etwas anders aussehen
- Bei grossen Datensätzen spielt das kaum eine Rolle
- Mit Optionen lassen sich auch die Anzahl Klassen festlegen, Überschriften ändern, usw. (siehe Übungen)

Old Faithful Geysir (Yellowstone NP)

- Geysir Old Faithful (Yellowstone National Park): Bekannte heisse Quelle
- Für Zuschauer und Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse
- Von 1.8.1978 - 8.8.1978 insgesamt 107 Messungen von aufeinanderfolgenden Ausbrüchen gemacht
- Daten Datei `geysir.txt`:

```
geysir <- read.table("../Data/geysir.txt")
head(geysir)
##   X.Tag. Zeitspanne Eruptionsdauer
## 1      1         78            4.4
## 2      1         74            3.9
## 3      1         68            4.0
## 4      1         76            4.0
## 5      1         80            3.5
## 6      1         84            4.1
```

- Abbildung Histogramme:



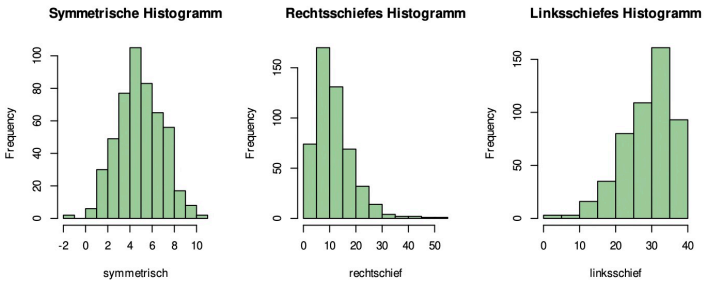
- Ausbruchsdauer eines Ausbruchs (rechts)
- Zeitspanne zwischen zwei Ausbrüchen (links)

- Bei beiden Histogrammen: *Bimodales* Verhalten sichtbar
- Es gibt zwei „Hügel" im Histogramm:
  - Zeitspanne zwischen zwei Ausbrüchen: Dauer relativ kurz (um die 50 Minuten) oder eher lang (um die 80 Minuten)
  - Zeitdauer zwischen zwei Ausbrüchen nicht „gleichmässig" verteilt
  - Dasselbe Verhalten bei der Zeitdauer eines Ausbruchs: Entweder ist der Ausbruch relativ kurz (um die 1.5-2 Minuten) oder lang (um die 4-4.5 Minuten)

Auf nächser Seite geht es weiter!

Schiefe von Histogrammen

- Abbildung:



- Histogramm links ist symmetrisch bezüglich ungefähr 5. Die Daten sind um 5 auf beiden Seiten ähnlich verteilt.
- Mittleres Histogramm: Die meisten Daten links im Histogramm –> Rechtsschiefes Histogramm.
- Rechtes Histogramm: Die meisten Daten rechts im Histogramm –> Linksschiefes Histogramm.
- Bezeichnung „rechts“ und „links“: Bezieht sich immer auf die Richtung, wo es weniger Daten hat.

Fortsetzung Old Faithful Geysir

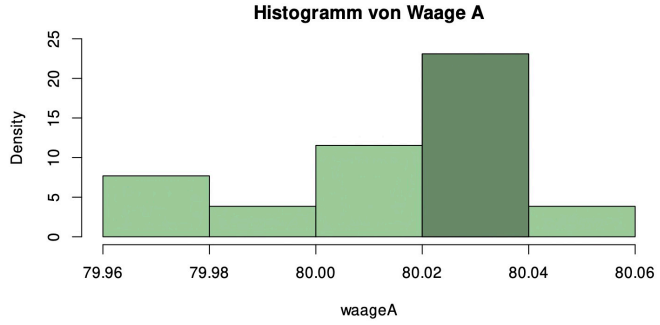
- Frage: Gibt es einen Zusammenhang zwischen Eruptionsdauer und Zeitspanne zwischen zwei Ausbrüchen gibt?
- Oder anders gefragt:
  – Geht es nach einem langen Ausbruch länger bis es wieder einen Ausbruch gibt?
  - Oder kommt ein Ausbruch schon sehr schnell wieder?
  - Oder gibt es gar keinen Zusammenhang?
- Fragen werden im 2. Teil dieses Blockes beantwortet.

Norminiertes Histogramm

- In Histogrammen bisher: Höhe der Blaken entspricht Anzahl Beobachtungen in einer Klasse.
- Oft besser und übersichtlicher: Balkenhöhe so wählen, dass die Balkenfläche dem prozentualen Anteil der jeweiligen Beobachtungen an der Gesamtanzahl Beobachtungen entspricht.
- Gesamtfläche aller Balken muss dann **gleich eins** sein.
- Auf der **vertikalen Achse** wird die **Dichte** angegeben.

Beispiel Waage *A*

- Normiertes Histogramm:



- Dichte der Klasse von $80.02 - 80.04$ ist etwa 23
- Fläche dieses Balkens (dunkelgrüne Fläche in Abbildung):

$$(80.04 - 80.02) \cdot 23 = 0.46$$

- Fläche mit 100 multipliziert: Prozentzahl der Daten, die in diesem Balken liegen
- Also etwa 46 % der Daten befinden sich zwischen 80.02 und 80.04

R-Code

- Code:

```
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")
```

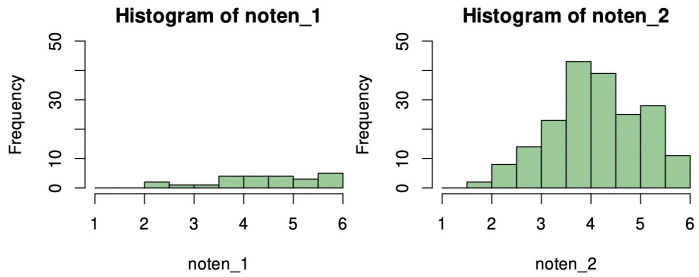
- Option `freq = F` (*frequency false*): Histogramm wird normiert gezeichnet
- Option `ylim = c(0, 25)`: siehe Skript
- `rect(80.02, 0, 80.04, 23.1, col = "darkseagreen4")`: siehe Skript

Warum normieren? Auf nächsten Seite

Warum normieren?

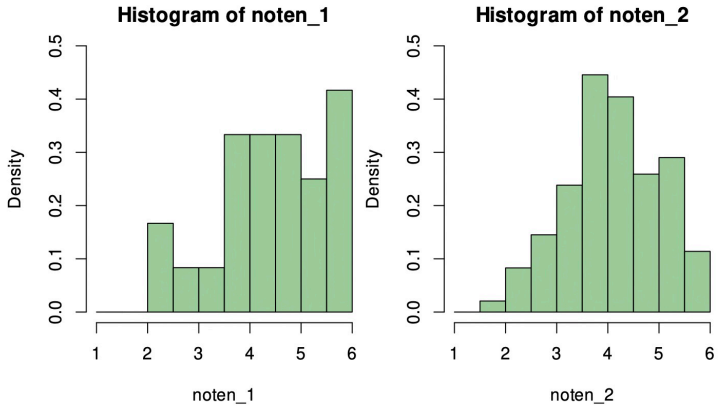
Beispiel

- Betrachten die Schulnoten einer Klasse von 24 Lernenden früher
- Vergleichen die Klasse mit einer (hypothetischen) anderen Klasse mit 194 Lernenden, die dieselbe Prüfung machten
- Histogramme mit der jeweiligen Häufigkeit



- Nicht vergleichbar

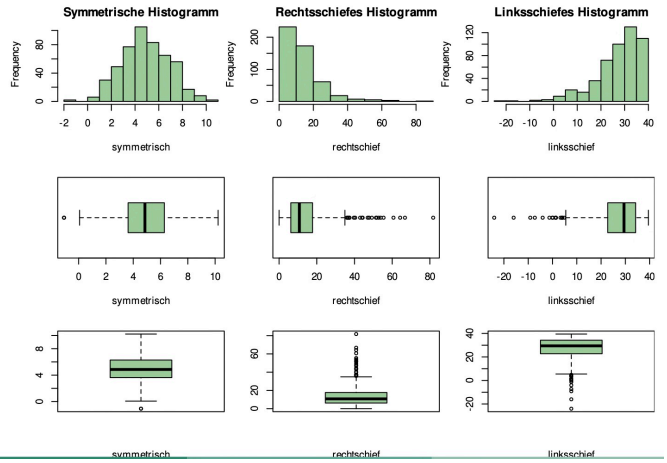
- Normierte Histogramme:



- Klasse 1 hat im oberen Bereich mehr Anteile als Klasse 2
- Vor allem der Balken von 5.5-6 von Klasse 1 ist sehr viel höher als der von Klasse 2
- Klasse 1 prozentual mehr starke Lernende als in Klasse 2
- Klasse 1 hat eher mehr schwächere Lernende als Klasse 2
- Im mittleren Bereich der Noten hat Klasse 2 prozentual mehr Lernende als Klasse 1

Schiefe im Boxplot

- Abbildung:



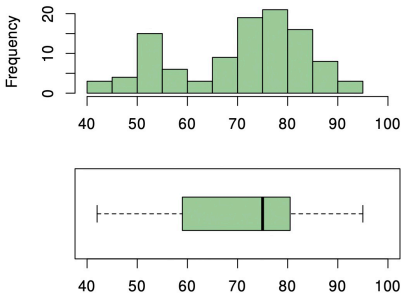
- Symmetrisches Diagramm links: Median in der Mitte der Box
- Rechtsschiefes Histogramm (Mitte): Median nicht mehr in der Mitte der Box, sondern nach links verschoben
- Der Abstand vom unteren Quartil zum Median ist kleiner als der Abstand vom Median zum oberen Quartil
- Vom unteren Quartil zum Median viele Daten in kleinem Bereich liegen
- Vom Median zum oberen Quartil braucht es ein viel grösserer Bereich bis 25 % der Daten in diesem Intervall liegen
- Beim linksschiefen Histogramm ist die Sachlage gerade umgekehrt

Boxplot: Bemerkungen

- Im *Boxplot* sind ersichtlich:
  - Lage
  - Streuung
  - Schiefe
- Man sieht aber z.B. *nicht*, ob eine Verteilung mehrere „Peaks" hat

Beispiel: Old Faithful

- Abbildung: Histogramm und der Boxplot von der Zeitspanne zwischen zwei Ausbrüchen von Old Faithful:



- Daten sind linksschief
- Boxplot: 50 % der mittleren Zeitspannen zwischen 60 und 80 Minuten
- Median liegt bei etwa 75 Minuten
- Daten zwischen dem Median und dem oberen Quartil liegen in einem Bereich von 5 Minuten (von 75-80 Minuten)
- D.h.: In diesem Bereich befinden sich relativ viele Zeitspannen verglichen zu Abstand 15 Minuten vom unteren Quartil zum Median

2. Deskriptive Statistik zweidimensionaler Daten

- Zweidimensionale Daten: An einem Versuchsobjekt werden jeweils zwei verschiedene Grössen gemessen.
- Beispiel: An einer Gruppe von Menschen wird jeweils die Körpergrösse und das Körpergewicht gemessen.
- Versuchsobjekt: Menschen, zudem je zwei Messungen gehören (Körpergrösse und Körpergewicht).
- Old Faithful: Versuchsobjekt ist ein Ausbruch, zudem je zwei Messungen gehören:
  – Die Eruptionsdauer
  – Die Zeit bis zum nächsten Ausbruch.

Daten: Weinkonsum - Mortalität

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

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

- Frage: Gibt es einen Zusammenhang zwischen der Sterblichkeitsrate aufgrund von Herzkreislauferkrankung und Weinkonsum?

Graphische Darstellung: Streudiagramm

- Wichtiger Schritt in der Untersuchung zweidimensionaler Daten: Graphische Darstellung.
- Meist über ein sogenanntes Streudiagramm (engl.: Scatterplot)
- Zwei Messungen als Koordinaten von Punkten in einem Koordinatensystem interpretiert und dargestellt.

Beispiel: Weinkonsum

- Grösse „Weinkonsum":

$$x_1,\ x_2,\ \ldots,\ x_{18}$$

- Zugehörige Grösse „Mortalität":

$$y_1,\ y_2,\ \ldots, y_{18}$$

- Koordinaten der Punkte:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_{18}, y_{18})$$

- Punkt mit den Koordinaten von Norwegen

$$(x_1, y_1) = (2.8,\ 6.2)$$

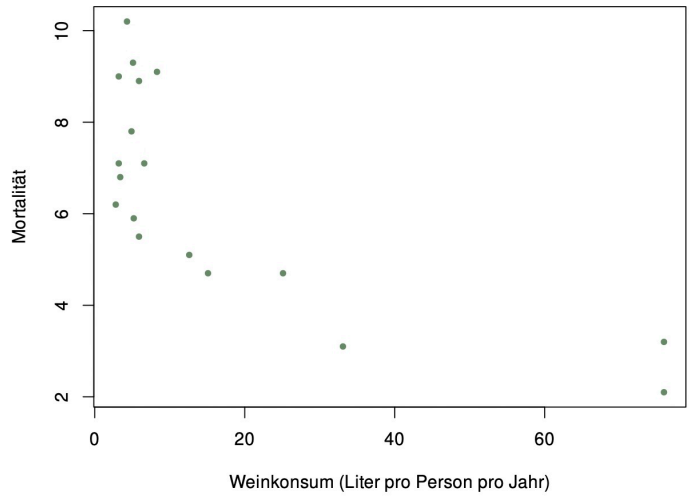
- Punkte in Koordinatensystem einzeichnen

Streudiagramm mit R

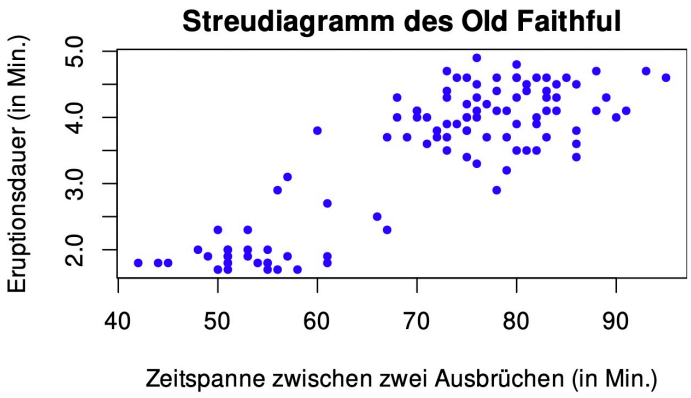
- Plot deutet an, dass hoher Weinkonsum weniger Sterblichkeit wegen Herz-Kreislauferkrankungen zur Folge hat
- Kann Zufall sein (keine Kausalität)
- Heisst *nicht*, dass Weinkonsum gesund ist (Leber!)
- R-Befehl

```
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

)
```

Zweidimensionales Streudiagramm



Beispiel Old Faithful



- Zunächst ist die Punktwolke steigend:
  - Je länger die Zeitspanne zwischen den Ausbrüchen, umso länger dauert der Ausbruch
- Im Streudiagramm hat es zwei Gruppen:
  - Eine links unten und eine rechts oben
  - Zeitspanne zwischen zwei Ausbrüchen kurz und die nächste Eruptionsdauer kurz
  - Zeitspanne ist lang und der Eruptionsdauer ist lang
  - Eine mittlere Zeitspanne (um die 70 Minuten) mit einer mittleren Ausbruchsdauer (um die 3 Minuten) gibt es nicht

Abhängigkeiten und Kausalität

- Bei Streudiagrammen aufpassen: Abhängigkeit nicht mit Kausalität verwechseln.
- Gesetzmässigkeit vorhanden, heisst dies noch lange nicht, dass diese Gesetzmässigkeit auch kausal erklärt werden kann.
- Beispiel Old Faithful: „Je länger desto länger"-Sachverhalt festgestellt.
- Das dieser aber erklärt werden kann, reicht das Streudiagramm nicht.
- Dazu müssen andere Methoden verwendet werden.

(Fiktives) Beispiel für Lineare Regression

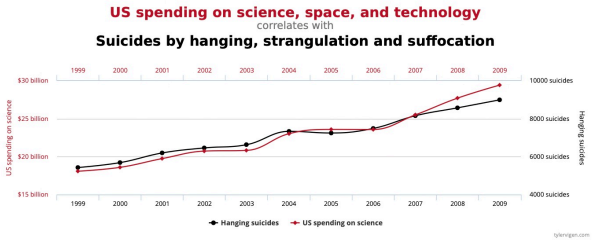
- Kunde kauft in Buchhandlung 10 Bücher

	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:
  - Wieviel kostet eine Seite?
  - Wie teuer ein Buch mit „null" Seiten wäre? $\rightarrow$ 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
  - Zusammenhang zwischen Seitenzahl x und Buchpreis y
- Vorhersagen über Buchpreis möglich für Bücher mit Seitenzahlen, die in Liste nicht auftauchen

Beispiel

- Abbildung:



- Kurven haben gleiche Form
- Es gibt aber keinen kausalen Zusammenhang zwischen den Kurven

- Abbildung:

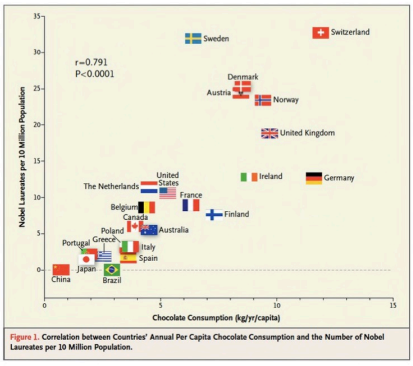
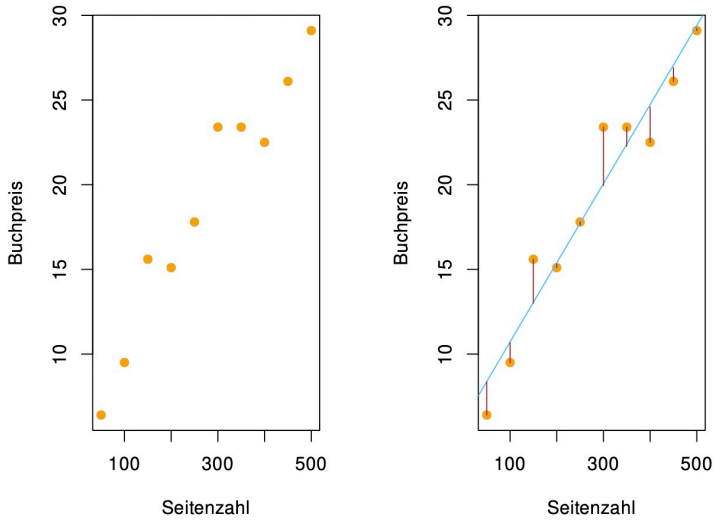


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- Auch hier keine kausale Gesetzmässigkeit vorhanden

1

Streudiagramm und Regressionsgerade



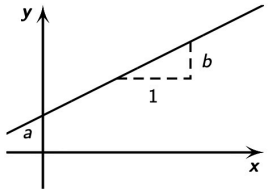
Repetition: Geradengleichung

- Gerade:

$$y = a + bx$$

  - a: y–Achsenabschnitt
  - b: Steigung

- Skizze:



- Interpretation der Steigung: Nimmt x um eine Einheit zu, so ändert sich y um b

2.

Regressionsgerade und Residuum

- Vermutung: Eine Gerade scheint recht gut zu den Daten zu passen
- Diese Gerade hätte die Form:

$$y = a + bx$$

mit

  - y: Buchpreis
  - x: Seitenzahl
  - Parameter a: Grundkosten des Verlags für ein Buch
  - Parameter b: Kosten pro Seite

- Problem: Gerade finden, die möglichst gut zu allen Punkten passt?
- Was heisst möglichst gut?

3.

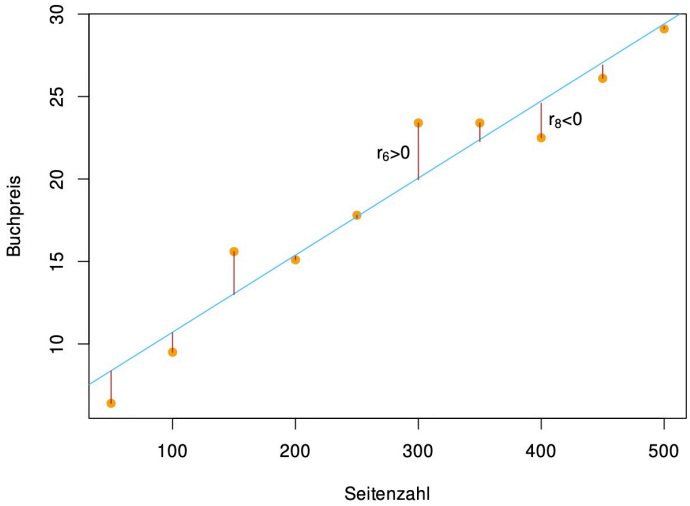
- Möglichkeit: Vertikale Abstände zwischen Beobachtung und Gerade zusammenzählen
- Dabei sollte eine kleine Summe der Abstände eine gute Anpassung bedeuten
- Abstände von Messpunkten zu Geraden $\rightarrow$ neuer Begriff:

Residuum

Ein *Residuum* r_i ist die vertikale Differenz zwischen einem Datenpunkt (x_i, y_i) und dem Punkt $(x_i, a + bx_i)$ auf der gesuchten Geraden:

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

4.



5.

- Beispiel: Residuen r_6 und r_8 für *diese* Gerade in Abbildung
- Residuum r_6 positiv, da Punkt überhalb der Gerade
- Entsprechend ist $r_8 < 0$
- Gerade $y = a + bx$ so bestimmen, dass die Summe

$$r_1 + r_2 + \ldots + r_n = \sum_i r_i$$

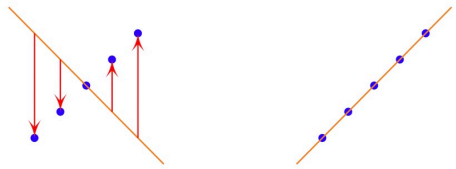
minimal wird

- Minimierung von $\sum_i r_i$ hat aber eine *gravierende Schwäche*: Falls Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegen: Summe der Abstände etwa null
- Dabei passt die Gerade gar nicht gut zu den Datenpunkten!

6

Beispiel

- Abbildung:



7.

- 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 Punktwolke passt?
- Verfahren gesucht, das diese Gerade eindeutig festlegt

Methode der kleinsten Quadrate

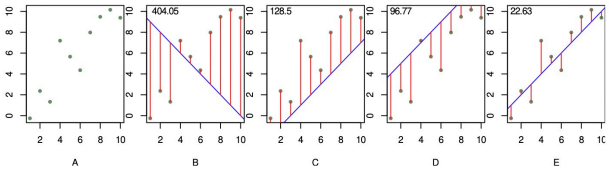
- Eine andere Möglichkeit besteht darin, die Quadrate der Abweichungen aufzusummieren, also

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

- Parameter a und b so wählen, dass diese Summe minimal wird

Graphisch: Methode der kleinsten Quadrate

- Skizze:



- Gesucht: Gerade, die gemäss der kleinsten Methode am besten zur Punktwolke links passt
- Punktwolke steigt: Steigung der Geraden positiv
- Abb. B: fallende Gerade:
  - Gerade passt nicht gut zur Geraden
  - Residuen (rot) sind sehr lang
  - Oben links: Summe der Quadrate der (roten) Residuen: 404.05
- Abb. C: steigende Gerade:
  - Gerade zwar steigend aber zu tief
  - Residuen (rot) immer noch lang, aber besser als bei Abb. A
  - Oben links: Summe der Quadrate der Residuen: 128.25
- Abb. D: steigende Gerade:
  - Gerade zwar steigend aber zu hoch
  - Residuen (rot) immer noch lang, aber besser als bei Abb. A
  - Oben links: Summe der Quadrate der Länge der Residuen: 96.77
- Abb. E: steigende Gerade:
  - Gerade passt gut zur Punktwolke
  - Residuen (rot) klein, verglichen zu den anderen Abb.
  - Oben links: Summe der Quadrate der Länge der Residuen: 22.63

Buchbeispiel

- R berechnet für Beispiel die Werte $a = 6.04$ und $b = 0.047$
  - Grundkosten des Verlags sind also rund 6 SFr. (Preis des Buches für 0 Seiten)
  - Pro Seite verlangt der Verlag rund 5 Rappen

Beispiel: Buchpreis

- Mit diesem Modell: Preis für Bücher mit Seitenzahlen berechnen, die in der Tabelle nicht vorkommen
- Wieviel würde nach diesem Modell ein Buch von 375 Seiten kosten?
- $x = 375$ in die Geradengleichung oben einsetzen:

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

- Das Buch dürfte also etwa CHF 23.60 kosten
- Dieses Modell ist allerdings nur begrenzt gültig
- Vor allem bei *Extrapolationen* muss man vorsichtig sein
- Möglich: Was kostet ein Buch mit einer Million Seiten?
- Oder ein Buch mit -100 Seiten? $\rightarrow$ Nicht realistisch!

Bestimmung der Parameter a und b

- *Frage:* Wie berechnet der Computer die Parameter a und b?
- Parameter a, b minimieren (Methode der Kleinsten-Quadrate)

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

Die Lösung dieses Optimierungsproblem ergibt:

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

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

wobei $\overline{x}$ und $\overline{y}$ die Mittelwerte der jeweiligen Daten

- Diese Gerade $y = a + bx$ wird auch *Regressionsgerade* genannt

Lineare Regression mit 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 Befehl `lm(y~x)` passt R ein Modell von der Form $y = a + bx$ an die Daten an
- R findet also $a = 6.04$ und $b = 0.0467$

Plotten der Regressionsgerade

- Diese Gerade wird in R wie folgt gezeichnet:

```
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")
```

Beispiel: Körpergrösse Vater-Sohn

- Vermutung: Zusammenhang zwischen der Körpergrösse der Väter und der Grösse der Söhne
- Der britische Statistiker Karl Pearson trug dazu um 1900 die Körpergrösse von 10 (in Wahrheit waren 1078) zufällig ausgewählten Männern gegen die Grösse ihrer Väter auf

Grösse des Vaters	152	157	163	165	168	170	173	178	183	188
Grösse des Sohnes	162	166	168	166	170	170	171	173	178	178

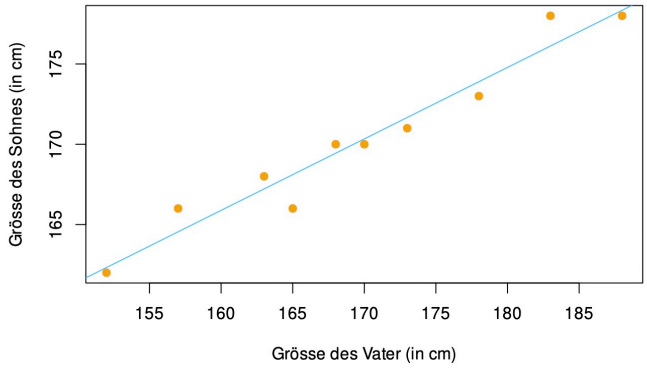
- Es *scheint* einen Zusammenhang zu geben: Je grösser der Vater, desto grösser der Sohn
- Streudiagramm: Möglicher linearer Zusammenhang besteht

- Die Punktwolke „folgt" der Geraden

$$y = 0.445x + 94.7$$

(mit der Methode der Kleinsten Quadrate aus den Daten)

- Streudiagramm:



- Möglich: In Tabelle nicht vorkommende Grösse von 180 cm des Vater, den zu erwartenden Wert für die Grösse seines Sohnes berechnen:

$$y = 0.445 \cdot 180 + 94.7 \approx 175\,\text{cm}$$

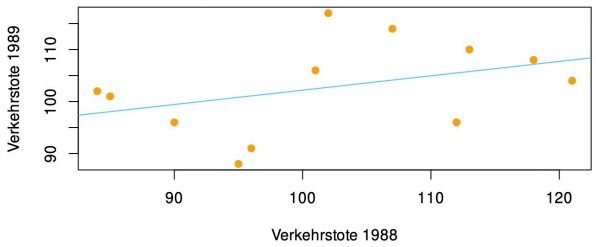
- Achtung: Formel nicht dort anwenden, wo man es *nicht* darf
- Für $x = 0$ erhält man einen Wert von 94.7
- Was heisst dies aber? Wenn der Vater 0 cm gross ist, so ist der Sohn ungefähr 95 cm gross $\rightarrow$ Macht keine Sinn!

Beispiel: Autounfälle

- Tabelle: Zusammenhang zwischen den Zahlen der Verkehrstoten her, die es 1988 und 1989 in zwölf Bezirken in den USA geben hat

Bezirk	1	2	3	4	5	6	7	8	9	10	11	12
Verkehrstote 1988	121	96	85	113	102	118	90	84	107	112	95	101
Verkehrstote 1989	104	91	101	110	117	108	96	102	114	96	88	106

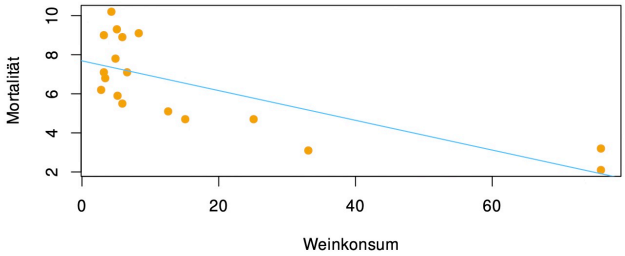
- Es besteht kein offensichtlicher Zusammenhang
- Streudiagramm: kein offensichtlicher Zusammenhang



- Zu erwarten, da es zwischen den Verkehrstoten der einzelnen Bezirke keinen Zusammenhang gibt
- In Abbildung ist noch die Regressionsgerade eingezeichnet
- Können sie zwar berechnen/einzeichnen, *aber diese macht hier gar keinen Sinn*
- *Immer* Berechnung und Plot vergleichen

Beispiel: Weinkonsum

- Schon gesehen: Sterblichkeit vs. Weinkonsum



- Regressionsgerade

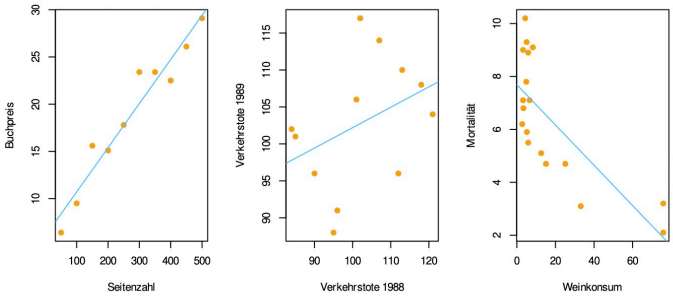
$$y = 7.68655 - 0.07608x$$

- Zusammenhang der Daten nicht linear ist (folgt eher einer Hyperbel)
- Regressionsgerade sagt wenig über den wahren Zusammenhang aus

SW04 – 1. Korrelation, 2. Wahrscheinlichkeitsmodelle

Wie gut passt die Regressionsgerade?

- Regressionsgerade kann (fast) immer bestimmt werden
- Beispiele:



- Letzten beiden Beispiele: Regressionsgerade sagt sehr wenig über die wirkliche Verteilung der Punkte im Streudiagramm aus
- Dafür gibt es zwei Gründe:
  - Punkte folgen scheinbar gar keiner Gesetzmässigkeit
  - 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

- Numerischer Wert der *linearen* Abhängigkeit von zwei Grössen:

Empirische Korrelation

$$r = \frac{(x_1-\overline{x})(y_1-\overline{y}) + \ldots + (x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+\ldots+(x_n-\overline{x})^2}\cdot\sqrt{(y_1-\overline{y})^2+\ldots+(y_n-\overline{y})^2}}$$

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

- Empirische Korrelation: Dimensionslose Zahl zwischen -1 und $+1$
- Misst Stärke und Richtung der *linearen Abhängigkeit* zwischen Daten x und y
- $r = +1$: Punkte liegen auf steigender Geraden : $y = a + bx$ mit $a \in \mathbb{R}$ und ein $b > 0$
- $r = -1$: Punkte liegen auf fallender Geraden : $y = a + bx$ mit $a \in \mathbb{R}$ und ein $b < 0$
- Sind x und y unabhängig (d.h. kein Zusammenhang), so ist $r = 0$

Begründung

- Eigenschaften nicht herleiten, sondern nur graphisch veranschaulichen
- Betrachten Zähler der Korrelation (Kovarianz):

$$(x_1-\overline{x})(y_1-\overline{y}) + \ldots + (x_n-\overline{x})(y_n-\overline{y}) = \sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})$$

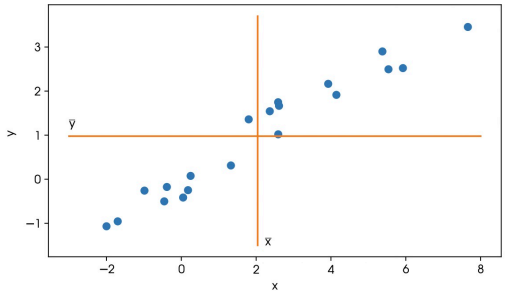
- Werden zeigen, dass diese Grösse:
  - grösser als 0 ist, wenn die Punktwolke einer steigenden Geraden folgt
  - kleiner als 0 ist, wenn die Punktwolke einer fallenden Geraden folgt
  - 0 ist, wenn kein Zusammenhang vorhanden ist
  - 0 sein kann, wenn kein linearer Zusammenhang vorhanden ist
- Nenner der Korrelation:

$$\sqrt{(x_1-\overline{x})^2+\ldots+(x_n-\overline{x})^2}\cdot\sqrt{(y_1-\overline{y})^2+\ldots+(y_n-\overline{y})^2}$$

- Positiv und *normalisiert* Kovarianz: Werte liegen zwischen -1 und 1

Beispiel

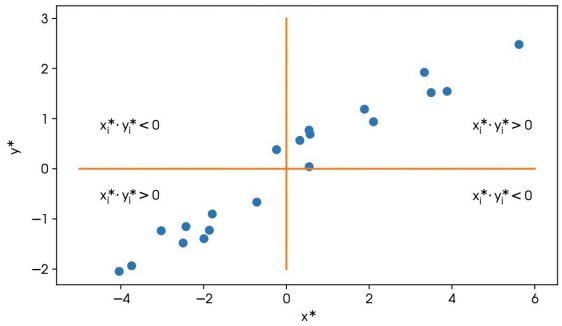
- Abbildung:



- Datenpunkte folgen mehr oder weniger einer Geraden
- Zu den Koordinatenachsen parallele Geraden $\overline{x}$ und $\overline{y}$ eingezeichnet
- Zähler der Korrelation:

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

- Von x-Koordinaten der Punkte wird der Durchschnitt $\overline{x}$ subtrahiert
- Von y-Koordinaten der Punkte wird der Durchschnitt $\overline{y}$ subtrahiert
- Ursprung der Abb. Folie ?? wird in Punkt $(\overline{x}, \overline{y})$ verschoben:



... (Nächste Seite)

...

- Zähler der Korrelation sieht dann wie folgt aus:

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

- 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*
- Abb. Folie ??: Fast alle Punkte im I. und III. Quadranten
- $x_i^* y_i^* > 0$
- Somit auch

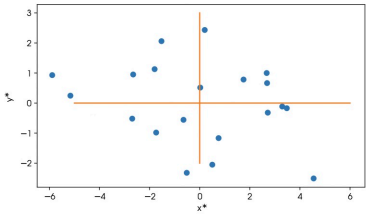
$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^* > 0$$

- Punkte folgen fallender Gerade: Fast alle Punkte im II. und IV. Quadranten
- $x_i^* y_i^* < 0$
- Somit auch

$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^* < 0$$

Beispiel

- Was passiert, wenn Punkte keinen Zusammenhang aufweisen?



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

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

Berechnung von Korrelation mit R

- Seitenzahl-Preis-Beispiel mit R

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

- Wert sehr nahe bei 1: Starker linearer Zusammenhang
- Wert positiv: „Je mehr, desto mehr"-Zusammenhang

Empirische Korrelation: Beispiele

- Beispiel Körpergrösse von Vater und Sohn: Erwarten hohen Korrelationskoeffizient, da Daten nahe Regressionsgerade

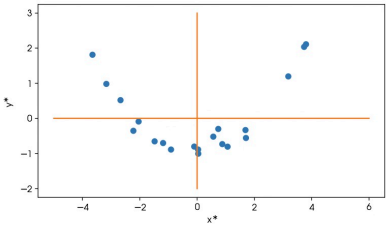
  $\rightarrow$ 0.973
- Verkehrsunfälle: Kein Zusammenhang: Korr.koeff. „nahe" 0

  $\rightarrow$ 0.386
- Weinkonsum:
  - Kein allzu grosser Korrelationskoeffizient (keine Gerade)
  - Sollte negativ sein, da mit steigendem Weinkonsum Mortalität sinkt

  $\rightarrow$ -0.746

Beispiel

- Wie sieht es für *quadratischen* Zusammenhang aus?

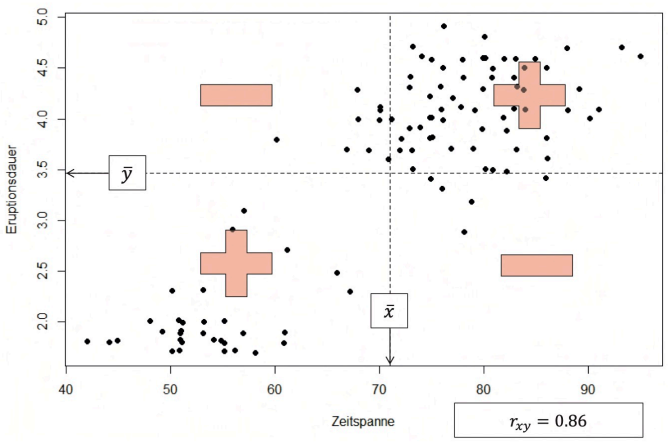


- Beträge der Produkte links und rechts von der y-Achse auf
- Es gilt

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

- Korrelationskoeffizient erkennt nur *lineare* Zusammenhänge

Abbildung: Old Faithful



Wahrscheinlichkeit

- Alle haben intuitives Gefühl, was W'keit ist
  - W'keit mit einem fairen Würfel eine 4 zu würfeln, ist ein Sechstel
- Aber: Interpretation der W'keit überraschend schwierig
- Aussage „Es regnet morgen mit einer W'keit von 80 %"
- Alles andere als klar, was damit gemeint ist

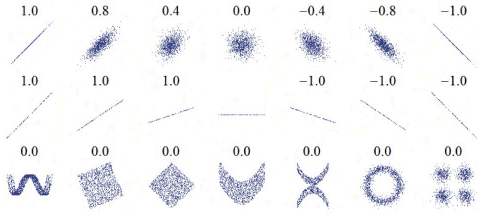
Wahrscheinlichkeitsmodell

- *Zufallsexperimente*: Ausgang nicht exakt vorhersagbar ist:
  - Würfelwurf
  - Münzenwurf
  - Anzahl Anrufe in einem Callcenter
- *Wahrscheinlichkeitsmodell* besteht aus:
  - Ereignissen, die in einem solchen Experiment möglich sind
  - W'keiten, die die verschiedenen Ergebnisse haben
- Beispiel: Würfel werfen
  - Mögliche Ergebnisse: $1, 2, 3, 4, 5, 6$
  - W'keit eine dieser Zahlen zu werfen: $\frac{1}{6}$ (sofern Würfel fair)
- Ein W'keitsmodell hat folgende Komponenten:
  - *Grundraum* Ω: Enthält alle möglichen Elementarereignisse ω
  - *Ereignisse* A, B, C: Teilmengen des Grundraums
  - *W'keiten* P, die zu den Ereignissen A, B, C gehören
- *Elementarereignisse*: *Mögliche* Ergebnisse (Ausgänge) des Experiments
- Zusammen bilden diese den Grundraum:

$$\Omega = \{\underbrace{\text{mögliche Elementarereignisse } \omega}_{\text{mögliche Ausgänge/Resultate}}\}$$

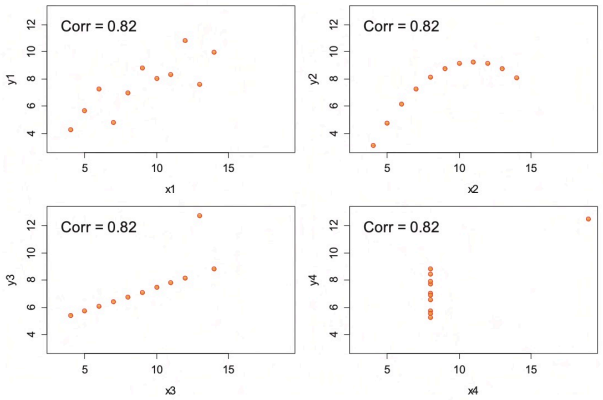
Empirische Korrelation: Bemerkungen

- Was heisst eigentlich „nahe" bei 0 oder „nahe" bei 1?
  - Das lässt sich allgemein nicht sagen
  - Hängt vom Problem und dem Gebiet ab
  - In Physik kann 0.9 sehr schlecht sein, in der Soziologie sehr gut
- Korrelation misst „nur" den *linearen Zusammenhang*
- Daher: Daten immer auch graphisch betrachten, statt sich „blind" auf Kennzahlen zu verlassen



Empirische Korrelation: Anscombe Plots

- Abbildung:



- Streudiagramme sehen sehr unterschiedlich aus, aber gleicher Korrelationskoeffizient

Beispiel: Würfelwurf

- Grundraum (die möglichen Ergebnisse)

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

- Element $\omega = 2$ ist ein Elementarereignis
- Bedeutung: Beim Würfeln wurde die Zahl 2 geworfen
- Zahl 7: *Kein* Elementarereignis, da nicht im Grundraum Ω

Weitere Beispiele auf nächster Seite

Beispiel: Anrufe in 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 Experimentes (Grundraum)

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

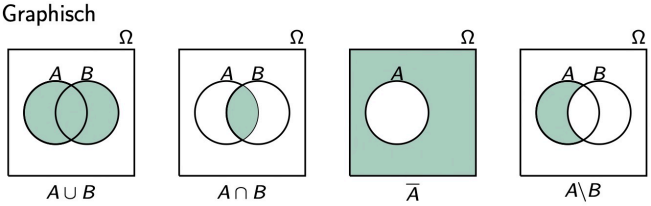
- Elementarereignis ist z.B. $\omega = KZ$

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	$\overline{A}$	*nicht* A
Differenz	$A \backslash B = A \cap \overline{B}$	A *ohne* B

- Graphisch



Beispiel: Würfelwurf

- Ereignis A: Die geworfene Zahl ist gerade:

$$A = \{2, 4, 6\}$$

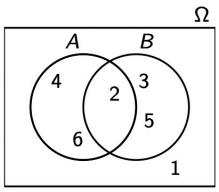
- Ereignis B: Die geworfene Zahl ist Primzahl:

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

- Ω wie gewohnt:

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

- Abbildung:



Weiter auf nächste Seite ...

Ereignis

- *Ereignisse* allgemeiner und wichtiger als Elementarereignisse, bestehen aber aus diesen
- *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 K geworfen wird
- Ereignis A besteht aus Elementarereignissen KZ und ZK
- Ereignis A ist dann die Menge

$$A = \{KZ, ZK\}$$

- Werfen ZZ: Ereignis A tritt *nicht* ein
- W'keit, dass A eintritt (falls Münze fair):

$$P(A) = \frac{2}{4} = \frac{1}{2}$$

- Statistik: W'keiten oft mit P oder p bezeichnet

Beispiel: Würfeln

- Ereignis A: „Eine ungerade Zahl würfeln"
  - Dann ist

$$A = \{1, 3, 5\}$$

  - Ereignis A tritt ein, wenn z.B. Zahl 5 gewürfelt wird
  - W'keit, dass A eintritt (fairer Würfel):

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

- Ereignis B: „Eine Zahl kleiner als 7 würfeln"
  - Das ist natürlich immer der Fall und somit ist

$$B = \Omega$$

  - B heisst dann *sicheres* Ereignis
  - W'keit, dass B eintritt (fairer Würfel):

$$P(B) = \frac{6}{6} = 1$$

- Ereignis C: „Die Zahl 7 würfeln"
  - Dies ist unmöglich

$$C = \{\}$$

  - Leere Menge $\{\}$ (oder $\emptyset$) enthält kein Element
  - Heisst *unmögliches* Ereignis
  - W'keit, dass C eintritt (fairer Würfel):

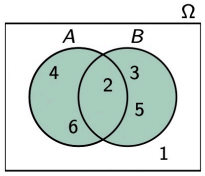
$$P(C) = \frac{0}{6} = 0$$

...

- *Vereinigung*: Alle Elemente, die entweder in A *oder* in B *oder* in beiden Mengen vorkommen:

$$A \cup B = \{2, 3, 4, 5, 6\}$$

  - Abbildung:

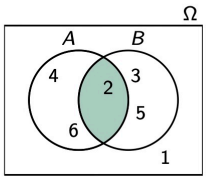


  - Element 2 kommt in Menge A *und* in Menge B vor

- *Schnittmenge*: Alle Elemente, die in A *und* in B vorkommen:

$$A \cap B = \{2\}$$

  - Abbildung:

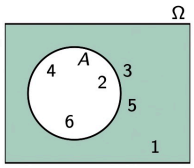


  - Element 2 einziges Element, das sowohl in Menge A wie auch in B vorkommt

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

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

  - Abbildung:

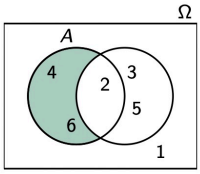


  - Menge $\overline{A}$ sind die ungeraden Zahlen

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

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

  - Abbildung:



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

Axiome der Wahrscheinlichkeit

- Eigenschaften von W'keiten

Kolmogorov Axiome der Wahrscheinlichkeitsrechnung

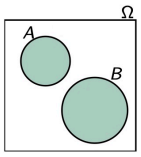
Jedem Ereignis A wird eine *W'keit* $P(A)$ zugeordnet, mit:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \{\}$

- Bezeichnung $P(A)$: W'keit, dass das Ereignis A eintritt
- Ereignis A: „ungerade Zahl würfeln" (bei fairem Würfel)

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

- Buchstabe P steht für *probability*
- A1: Wahrscheinlichkeit kann nicht negativ sein
- A2: Mit $P(\Omega) = 1$: W'keiten eines Ereignisses zwischen 0 und 1
- Mathematik (Statistik): W'keiten praktisch nie in Prozenten angeben
- A3: Für zwei *disjunkte* Ereignisse:



  - W'keit, dass eines der beiden eintritt, gleich Addition W'keiten 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:

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

  - Kann nicht sein, da $P(A \cup B) = \frac{5}{6}$
  - Grund: $A \cap B = \{2\} \neq \{\}$

Beispiel

- Wurf zweier Münzen:

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

- Plausibel: Alle 4 Elemente gleich wahrscheinlich (faire Münze)
- Wegen $P(\Omega) = 1$ müssen sich die W'keiten zu eins aufaddieren:

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

- Da alle Elementarereignisse gleichwahrscheinlich:

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

Rechenregeln aus Axiomen

Rechenregeln

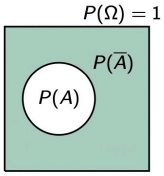
Sind A, B und $A_1, \ldots A_n$ Ereignisse, dann gilt

$P(\overline{A}) = 1 - P(A)$	für jedes A
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	für beliebige A und B
$P(A_1 \cup \ldots \cup A_n) \leq P(A_1) + \ldots + P(A_n)$	für beliebige $A_1, \ldots, A_n$
$P(B) \leq P(A)$	für beliebige A und B mit $B \subseteq A$
$P(A \backslash B) = P(A) - P(B)$	für beliebige A und B mit $B \subseteq A$

- W'keiten als Flächen im Venn-Diagramm vorstellen
- Totalfläche von Ω gleich 1 oder $P(\Omega) = 1$
- Regeln klar

1. Regel

- Abbildung:



- $P(A)$: Flächeninhalt der Fläche A
- $P(\overline{A})$: Flächeninhalt der restlichen Fläche in Ω
- Es gilt also offensichtlich:

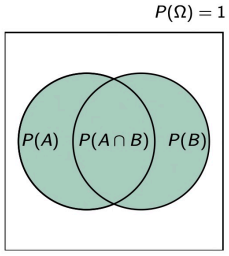
$$P(A) + P(\overline{A}) = P(\Omega) = 1$$

- Und damit:

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

2. Regel

- Abbildung:



- Schnittmenge $A \cap B$ 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:

$$A \cap B = \{\}$$

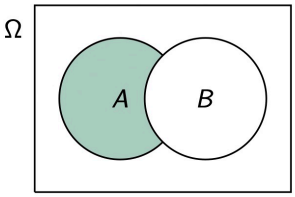
- Mit der Regel oben:

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

- Dies ist genau Axiom A3, da die W'keit des unmöglichen Ereignisses $\{\}$ gleich 0 ist

Knobelaufgabe

$P(A \setminus B) = \ldots$



„A ohne B"

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

Diskrete Wahrscheinlichkeitsmodelle

- Heisst: Grundraum endlich oder unendlich und diskret
- Begriff „diskret": *Endliche* Menge, wie:

$$\Omega = \{0, 1, \ldots, 10\}$$

- *Unendliche*, aber trotzdem *diskrete* Menge, wie

$$\Omega = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$$

- Menge $\Omega = \mathbb{R}$ (Menge aller Dezimalbrüche, Zahlengerade) ist *nicht* diskret
- Wird später für Messdaten eine sehr wichtige Rolle spielen

Wahrscheinlichkeiten für diskrete Modelle

- Berechnung von W'keiten für diskrete Modelle

Im diskreten Fall ist die W'keit eines Ereignisses

$$A = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

durch die W'keiten der zugehörigen Elementarereignisse $P(\omega)$ festgelegt:

$$P(A) = P(\omega_1) + P(\omega_2) + \ldots + P(\omega_n) = \sum_{\omega_i \in A} P(\omega_i)$$

- Alle W'keiten der Elementarereignisse aus Ereignis A werden addiert

Beispiele auf der nächsten Seite

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*: $P(\overline{B})$
- 1. Rechenregel: Komplement $\overline{B}$ von B:

$$\overline{B} = \{6\}$$

- Dann gilt:

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

Stochastische Unabhängigkeit

- Gesehen:

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

- Frage: Wie berechnet man $P(A \cap B)$?
  - Leider keine allgemeine Regel
  - Wenn W'keiten $P(A)$ und $P(B)$ bekannt, so ist W'keit $P(A \cap B)$ i. A. nicht aus $P(A)$ und $P(B)$ berechenbar
- Wichtiger Spezialfall: Berechnung von $P(A \cap B)$ aus $P(A)$ und $P(B)$ mit Produktformel:

> Sind Ereignisse A und B *stochastisch unabhängig*, so gilt
>
> $$P(A \cap B) = P(A) \cdot P(B)$$

- Was heisst hier aber „stochastisch unabhängig"?
- Ausgang des Ereignisses A keinen Einfluss auf den Ausgang des Ereignisses B hat und umgekehrt

Weitere Beispiele auf der nächsten Seite

Modell von Laplace

- Annahme: Jedes Elementarereignis hat die gleiche W'keit
- Ereignis $E = \{\omega_1, \omega_2, \ldots, \omega_g\}$;
- Grundraum m Elemente
- W'keiten addieren sich zu 1 und deshalb:

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

> Für ein Ereignis E im Laplace Modell gilt also
>
> $$P(E) = \frac{g}{m} = \sum_{k:\ \omega_k \in E} P(\{\omega_k\})$$

- Man teilt die Anzahl der *„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:

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

- 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:

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

Beispiel

- 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:

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

Beispiel

- 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

Beispiel

- 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

Beispiel

- 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

Beispiel

- 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

Achtung

- 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 $P(A \cap B)$

SW05 – 1. Zufallsvariable, 2. Wahrscheinlichkeitsverteilung

Zufallsvariable

1.

Zufallsvariable: Beispiel

- Zentraler Begriff in Statistik: *Zufallsvariable*
- Pack Spielkarten (Schweiz): 36 verschiedene Karten, mit 4 Farben und je den Werten $6, 7, 8, 9, 10$, Bube, Dame, König und Ass
- Ziehen nacheinander drei Karten, nach jedem Ziehen wieder zurücklegen
- 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

2.

- 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 $\rightarrow$ $0 + 3 + 4 = 7$
  2. 8, Bube, Ass $\rightarrow$ $0 + 2 + 11 = 13$
- Zweite Ziehung mit diesen Werten besser als die erste

3.

- Beispiel vorher: Situation kommt in der Stochastik häufig vor
- Zufallsexperiment mit dem Grundraum Ω
- Allen Elementarereignissen von Ω wird eine Zahl zugeordnet
- Zu jedem Elementarereignis ω gehört demnach eine Zahl

$$X(\omega) = x$$

- X: *Funktion*, die jedem Elementarereignis ω den Zahl x zugeordnet
- Diese Funktion wird *Zufallsvariable* genannt

4.

Beispiel

- Ziehen Karten aus einem Stapel Spielkarten
- Jeder Karte wird eine Zahl zugeordnet:

$$\begin{array}{lcl} \omega = \text{As} & \mapsto & X(\omega) = 11 \\ \omega = \text{König} & \mapsto & X(\omega) = 4 \\ \vdots & & \vdots \\ \omega = \text{Sechs} & \mapsto & X(\omega) = 0 \end{array}$$

- Mit Zahlen $X(\omega)$ kann z.B. „Durchschnitt" der gezogenen 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

5.

Beispiel

- Werfen gemeinsam einen blauen und roten Würfel
- Grundraum Ω (siehe früher): Augenzahlen der Würfel:

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

- Auf Ω sind verschiedene Zufallsvariablen definierbar
- Sei X die Zufallsvariable für die Summe der Augenzahlen:
  - Dann gilt:

$$X(16) = 7 \qquad \text{oder} \qquad X(31) = 4$$

  - Die Werte, die die Zufallsvariable annehmen kann, wird als *Wertemenge* bezeichnet:

$$\mathbb{W}_X = \{2, 3, 4 \ldots, 11, 12\}$$

6.

- Sei Y die Augenzahl des roten Würfels:
  - Dann gilt:

$$Y(16) = 6 \qquad \text{oder} \qquad Y(31) = 1 \qquad \text{oder} \qquad Y(13) = 3$$

  - Wertemenge:

$$\mathbb{W}_Y = \{1, 2, \ldots, 6\}$$

- Sei Z gleich 0 für alle Elementarereignisse:
  - Dann gilt:

$$Z(16) = 0 \qquad \text{oder} \qquad Z(31) = 0 \qquad \text{oder} \qquad Z(13) = 0$$

  - Wertemenge:

$$\mathbb{W}_Z = \{0\}$$

- Letztes Beispiel ist eine völlig legitime Zufallsvariable
- Wie sinnvoll diese ist, ist eine andere Frage

7.

Beispiel

- Wählen zufällig eine Person aus
- Grundraum Ω: Alle 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

Wahrscheinlichkeitsverteilung einer Zufallsvariable

- Schon gesehen: Berechnung W'keit $P(E)$ eines Ereignisses E
- Entsprechend: W'keit einer allgemeinen Realisierung x einer Zufallsvariable X

Beispiel: Spielkarten

- Zufallsvariable X: Wert einer gezogenen Spielkarte
- Wie gross ist W'keit ist, dass die gezogene Karte den Wert 4 hat
- Realisierung ist $X = 4$
- Zugehörige W'keit:

$$P(X = 4)$$

- 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

Allgemein

- Definition:

> Die Werte einer Zufallsvariablen X (die möglichen Realisierungen von X) treten mit gewissen W'keiten auf. Die W'keit, dass X den Wert x annimmt, berechnet sich wie folgt:
>
> $$P(X = x) = P(\{\omega \mid X(\omega) = x\}) = \sum_{\omega; X(\omega)=x} P(\omega)$$

- Im Spielkartenbeispiel ist $x = 4$ und ω alle möglichen Könige, deren entsprechende W'keiten aufaddiert werden

Definition Zufallsvariable

> **Zufallsvariable**
>
> Eine *Zufallsvariable* X ist eine *Funktion*:
>
> $$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

Bemerkungen

- $\mathbb{R}$: Reelle Zahlen (Punkte der Zahlengeraden, alle Dezimalbrüche)
- Stochastik X (oder $Y, Z, \ldots$): Notation für *Zufallsvariable* (*Funktion*)
- Funktionen in der Mathematik oft in der Form:

$$y = f(x)$$

- Folgende Unterscheidung ist sehr wichtig:

> - *Zufallsvariable* werden mit *Grossbuchstaben* X (oder Y, Z) bezeichnet
> - Entsprechender *Kleinbuchstabe* x (oder y, z) stellt *konkreter Wert* dar, den die Zufallsvariable annehmen kann
> - Ereignis, bei dem die Zufallsvariable X den Wert x annimmt:
>
> $$X = x$$

- Bei einer Zufallsvariable ist nicht die Funktion X zufällig, sondern nur das Argument ω:
  - Je nach Ausgang des Zufallsexperiments ω anderer Wert $x = X(\omega)$
- x heisst auch eine *Realisierung* der Zufallsvariablen X

Beispiele

- Spielkartenbeispiel: Realisierung $X = 11$ entspricht dem Ziehen eines Asses
- Würfelbeispiel: Realisierung $X = 8$ entspricht dem Würfeln der Augensumme 8

Wahrscheinlichkeitsverteilung

- Vorher: W'keit einer Realisierung berechnet
- Jetzt: W'keiten *aller* Realisierungen berechnen
- Definition:

> **Wahrscheinlichkeitsverteilung**
>
> Für *jede* Realisierung einer Zufallsvariable wird die zugehörige W'keit berechnet $\to$ *W'keitsverteilung* dieser Zufallsvariablen

Beispiele auf nächster Seite...

Beispiel

- Zufallsvariable X: Wert einer gezogenen Jasskarte
- Schon berechnet: W'keit
$$P(X=4)=\frac{1}{9}$$
- W'keit $P(X=0)$ mit Laplace-W'keit:
$$P(X=0)=\frac{16}{36}=\frac{4}{9}$$
- W'keit $P(X=2)$: Ziehen eines Buben:
$$P(X=2)=\frac{4}{36}=\frac{1}{9}$$
- W'keitsverteilung von X als Tabelle:

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 Tabelle nicht aufgeführt
- Grund: Diese Werte können nicht gezogen werden
- Trotzdem eine W'keit zuordnen, nämlich die Zahl 0:
$$P(X=1)=0 \quad \text{oder} \quad P(X=178)=0$$
- Addieren alle Werte der W'keitsverteilung $\rightarrow$ 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

- W'keitsverteilung für die Zufallsvariable X:

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

- Brüche sind hier ungekürzt: Man sieht besser, dass die Summe der W'keiten 1 geben muss

Beispiel: Augensumme zweier Würfel

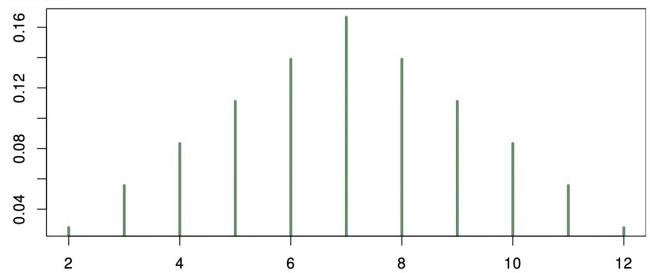
- X: Zufallsvariable der geworfenen Augensumme
- Wie gross ist die W'keit, genau die Augensumme 6 zu würfeln?
  - Gesucht: $P(X=6)$:
$$P(X=6)=\frac{5}{36}$$
- Wie gross ist die W'keit, die Augensumme 6 oder 8 zu würfeln?
  - Gesucht: $P(X=6)+P(X=8)$:
$$P(X=6)+P(X=8)=\frac{5}{36}+\frac{5}{36}=\frac{10}{36}=\frac{5}{18}$$
- Wie gross ist die W'keit, mindestens die Augensumme 3 zu würfeln?
  - Gesucht:
$$P(X\geq 3)=P(X=3)+\ldots+P(X=12)$$
  - Einfacher:
$$P(X\geq 3)=1-P(X=2)$$
  - Also:
$$1-P(X=2)=1-\frac{1}{36}=\frac{35}{36}$$

Beispiel: Augensumme zweier Würfel

- Schon gesehen:

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

- Skizze:



- Wie gross ist die W'keit, höchstens die Augensumme 3 zu würfeln?
  - Gesucht:
$$P(X\leq 3)=P(X=2)+P(X=3)$$
  - Also:
$$P(X=2)+P(X=3)=\frac{1}{36}+\frac{2}{36}=\frac{3}{36}=\frac{1}{12}$$
- Wie gross ist die W'keit, eine Augensumme von 3 bis 5 zu würfeln?
  - Gesucht:
$$P(3\leq X\leq 5)=P(X=3)+P(X=4)+P(X=5)$$
  - Also:
$$P(3\leq X\leq 5)=\frac{2}{36}+\frac{3}{36}+\frac{4}{36}=\frac{9}{36}=\frac{1}{4}$$

„Allgemein“ auf nächer Seite

Allgemein

- Es gilt:

Wahrscheinlichkeitsverteilung

„Liste" von $P(X = x)$ für alle *möglichen* Werte $x_1, x_2, \ldots, x_n$ heisst *diskrete W'keitsverteilung* der diskreten Zufallsvariablen X

Es gilt immer:

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

Mit Summenzeichen:

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

Alle W'keiten einer W'keitsverteilung ergeben 1

Beispiel

- Wurf eines fairen Würfels: Alle 6 möglichen Zahlen gleiche W'keit geworfen zu werden
- Zufallsvariable X sei die geworfene Zahl
- Erwartungswert $\mathrm{E}(X)$:

$$\begin{aligned}\mathrm{E}(X) &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_6 \cdot P(X = x_6)\\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}\\ &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6)\\ &= 3.5\end{aligned}$$

- Dieser Erwartungswert 3.5 nicht anderes als der Durchschnitt der Augenzahlen
- Wie lässt sich dieser Wert nun interpretieren?
- Würfeln fairen Würfel 100mal: Durchschnitt meistens *nicht* exakt 3.5
- Aber: Durchschnitt in der *Nähe* von 3.5
- Annäherung sollte immer besser werden, je mehr wir würfeln
- 100 Milliarden mal würfeln: Durchschnitt nicht *exakt* 3.5, aber sehr nahe dran
- Interpretation: Für sehr viele Würfe liegt Durchschnitt sehr nahe bei Erwartungswert
- Standardabweichung mit 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
```

- D.h.: Abweichung „durchschnittlich" 1.7 von 3.5

Kennzahlen einer Verteilung

- Beliebige (diskrete) Verteilung: Vereinfachend durch 2 Kennzahlen zusammengefasst:
  - *Erwartungswert* $\mathrm{E}(X)$: Mittlere Lage der Verteilung
  - *Standardabweichung* $\sigma(X)$: Streuung der Verteilung
- Diskrete Zufallsvariable X: Mögliche Werte $x_1, x_2, \ldots, x_n$
- Definition:

Erwartungswert und Standardabweichung

- *Erwartungswert*:

$$\begin{aligned}\mathrm{E}(X) &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n)\\ &= \sum_{\text{alle möglichen } x} xP(X = x)\end{aligned}$$

- *Varianz* und *Standardabweichung*:

$$\begin{aligned}\mathrm{Var}(X) &= (x_1 - \mathrm{E}(X))^2 \cdot P(X = x_1) + \ldots + (x_n - \mathrm{E}(X))^2 \cdot P(X = x_n)\\ &= \sum_{\text{alle möglichen } x} (x - \mathrm{E}(X))^2 P(X = x)\end{aligned}$$

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

Beispiel Spielkarte auf nächster Seite

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 $\mathrm{E}(X)$:

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

- Zahlen untereinander in Tabelle werden multipliziert und dann addiert
- R:

```
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
```

- Durchschnittlicher Wert, der zu erwarten ist, wenn Karte sehr oft gezogen und wieder in Stapel zurücklegt wird
- Viele Karten mit Wert 0 $\rightarrow$ Erwartungswert eher tief
- Varianz und die Standardabweichung:

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

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

Bemerkungen

- Erwartungswert einer diskreten Zufallsvariable: Gewichtetes arithmetisches Mittel von allen möglichen Werten, wobei die Werte mit ihrer W'keit gewichtet werden
- Erwartungswert: Oft auch mit μ_X bezeichnet:
  - Index X wird oft weglassen, falls Zufallsvariable klar
- W'keiten für alle Werte $x_1, x_2, \ldots, x_n$ gleich: Erwartungswert gerade arithmetisches Mittel der Werte
- Varianz das Quadrat der Abweichung eines Wertes der Zufallsvariable vom Erwartungswert mit der W'keit des Wertes gewichtet
- Standardabweichung hat dieselbe Einheit wie X: Einheit der Varianz deren Quadrat ist:
  - Z. B. X in Metern (m) gemessen $\rightarrow$ $\mathrm{Var}(X)$ in Quadratmeter (m^2)
  - $\sigma(X)$ wiederum die Dimension Meter (m)

Unterschied empirischer und theoretischer Kennzahlen

- Gesehen: Mittelwert:

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

- *Empirische* Varianz:

$$\mathrm{Var}(x) = \frac{(x_1 - \overline{x}_n)^2 + (x_2 - \overline{x}_n)^2 + \cdots + (x_n - \overline{x}_n)^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2$$

- Empirische Standardabweichung s_x:

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

- Wie hängen diese Definition und die Definition für die Standardabweichung for eine Zufallsvariable X zusammen?
- Sehr genau unterscheiden:
  - Arithmetische Mittelwert $\overline{x}$: Aus *konkreten* Daten berechnet: Aus Messwerten $x_1, \ldots, x_n$ wird nach der Formel oben $\overline{x}_n$ berechnet
  - Erwartungswert $\mathrm{E}(X)$: *Theoretischer Wert*, der sich aus dem Modell der W'keitsverteilung ergibt
- Hoffnung: Arithmetisches Mittel $\overline{x}$ nähert für immer mehr Versuche den theoretischen Wert $\mu_X = \mathrm{E}(x)$ immer besser an, sofern die Daten der W'keitsverteilung von X folgen
- Ist dies nicht der Fall, so stimmt etwas am Modell nicht (z.B. alle Seiten beim Würfel haben gleiche W'keit geworfen zu werden)

1.

Beispiel: Fairer 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
```

2.

- 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

3.

- 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

4.

- 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

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

6.

- Analoges gilt für die Standardabweichung:
  - Empirische Standardabweichung: s_X aus *konkreten* Daten berechnet: Aus Messwerte $x_1, \ldots, x_n$ wird nach Formel oben s_X berechnet
  - Standardabweichung σ_X: Theoretischer Wert, der sich aus Modell der W'keitsverteilung ergibt
- Hoffnung: Empirische Standardabweichung s_X nähert für immer mehr Versuche den theoretischen Wert σ_X immer besser an, falls Daten der W'keitsverteilung von X folgen
- Ist dies nicht der Fall, so stimmt etwas am Modell nicht (z.B. alle Seiten beim Würfel haben gleiche W'keit geworfen zu werden)

SW06 – Bedingte Wahrscheinlichkeit

1.

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, $\overline{R}$: Nichtraucher
- Tabelle:

$|R \cap M|$ $|R|$

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

2.

- Es hat:
  - 4 Raucher und 16 Nichtraucher
  - 8 Frauen und 12 Männer
- Wert 3 links oben: Anzahl Personen, die Männer sind *und* rauchen
- Schreibweise:

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

3.

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

$P(R \cap M)$ $P(R)$

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

4.

- 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:

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

5.

- Betrachten nur ein Teil der Tabelle $\rightarrow$ Raucher

	M	F	
R	0.15	0.05	0.2
$\overline{R}$	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:

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

6.

- 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:

$$P(M \mid R)$$

- Begriff „bedingt": Nicht die gesamte Grundmenge betrachten, sondern nur ein Teil davon

7.

- Neue Grundmenge hier: Raucher R
- Dies ist in $P(M \mid R)$ die Grösse nach dem Längsstrich
- Es gilt dann:

$$P(M \mid R) = \frac{P(R \cap M)}{P(R)} \qquad (*)$$

- Formel wird als Definition der bedingte W'keit verwendet

8.

- Berechnen bedingte W'keit:

$$P(R \mid M)$$

- 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
$\overline{R}$	0.45	0.35	0.8
	0.6	0.4	1

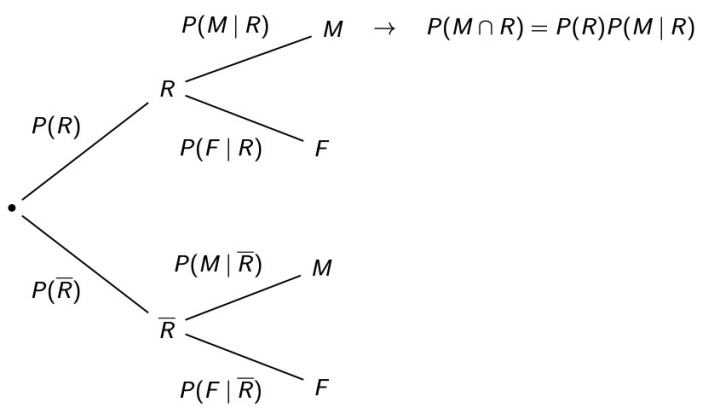
9.

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

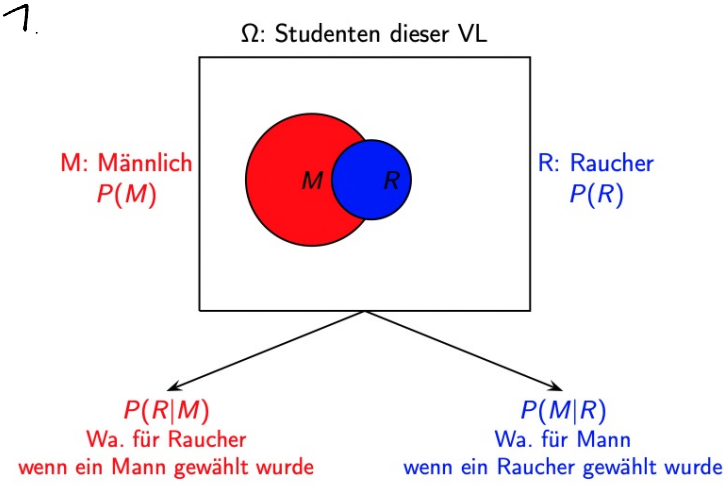
$$P(R \mid M) = \frac{P(M \cap R)}{P(M)} = \frac{P(R \cap M)}{P(M)} = \frac{0.15}{0.6} = 0.25$$

Bedingte Wahrscheinlichkeit

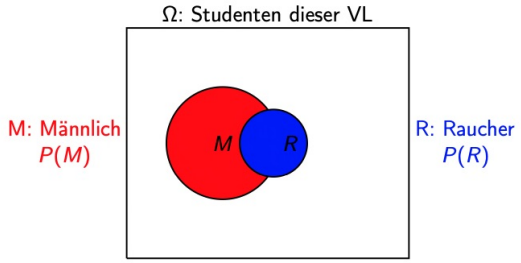
Baumdiagramm



1.



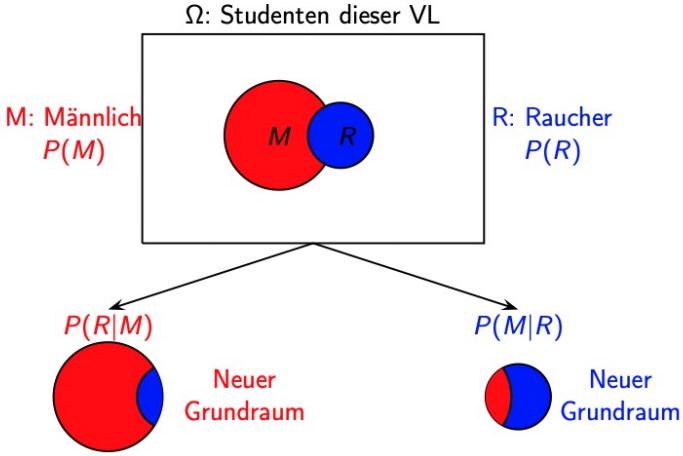
2.



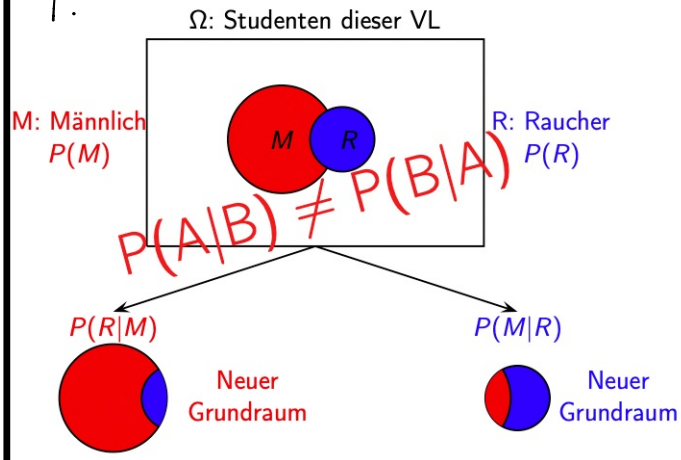
Welche Aussagen sind korrekt?

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

3.



4.



5.

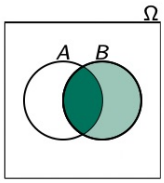
Bedingte Wahrscheinlichkeit: Definition

- Die *bedingte W'keit* ist die W'keit, dass das Ereignis A eintritt, wenn man schon weiss, dass B eingetreten ist
- Bezeichnung:
$$P(A \mid B)$$
- Längsstrich wird als „unter der Bedingung" gelesen
- Bedingte W'keit $P(A \mid B)$ wird definiert durch
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$
- Interpretation: $P(A \mid B)$ ist die W'keit für das Ereignis A, wenn man weiss, dass das Ereignis B schon eingetroffen ist

6.

Verdeutlichung der Formel mit Flächen

- Graphisch:



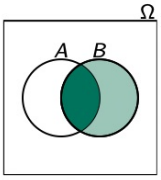
- Es ist $|\Omega| = 1$
- $P(A \mid B)$ Flächeninhalt der dunkel gefärbten Flächen
- $P(B)$ Flächeninhalt der gesamten gefärbten Fläche B
- Anteil der dunkelgefärbten Fläche zur gefärbten Fläche ist:
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Weiter auf nächster Seite

7.

Verdeutlichung der Formel mit Flächen

- Graphisch:



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

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

1.

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; $\overline{D}$: Krankheit ist nicht vorhanden
  - $+$: Test zeigt Krankheit an; $-$: Test zeigt Krankheit nicht an
- W'keiten in Tabelle sind durch Versuche bekannt

	D	$\overline{D}$
$+$	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
  - $P(-|\overline{D})$: 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 benützt

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

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

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

- Scheinbar ist dieser Test recht genau

2.

- 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

Bayes' Theorem

Nützlicher Zusammenhang zwischen $P(A \mid B)$ und $P(B \mid A)$:

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

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:
    $$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \quad \Rightarrow \quad P(B \cap A) = P(B \mid A)P(A)$$
  - und:
    $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A \mid B)P(B)$$
- Da $A \cap B = B \cap A$, gilt:
  $$P(A \cap B) = P(B \cap A)$$
- Und somit gilt auch
  $$P(B \mid A)P(A) = P(A \mid B)P(B)$$
- Dividieren beide Seiten durch $P(B)$ und erhalten Bayes Theorem:
  $$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

2.

Gesetz der totalen Wahrscheinlichkeit

- Weiterer nützlicher Begriff: *Totale Wahrscheinlichkeit*
- Menge A in Mengen $A_1, \ldots, A_k$ 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:
  $$A_1 = \{1\}, \qquad A_2 = \{2, 4\}, \qquad A_3 = \{3, 5, 6\}$$
- Es gilt also:
  $$A_1 \cap A_2 = \{\}; \quad A_1 \cap A_3 = \{\}; \quad A_2 \cap A_3 = \{\}$$
  und
  $$A_1 \cup A_2 \cup A_3 = A$$
- Es gilt dann das

Gesetz der totalen Wahrscheinlichkeit

Für Partionierung $A_1, \ldots, A_k$ und jedes beliebige Ereignis B gilt:

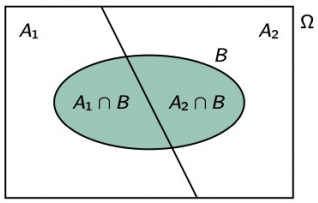
$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_k)P(A_k) \\ &= \sum_{i=1}^{k} P(B|A_k)P(A_k) \end{aligned}$$

- $k = 2$:
  $$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$$
- $k = 3$:
  $$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

2.

Herleitung

- Fall $k = 2$
- Graphische Darstellung:



- Mengen A_1 und A_2 bilden eine Partition von Ω
- Es gilt also:
  $$A_1 \cup A_2 = \Omega \qquad \text{und} \qquad A_1 \cap A_2 = \{\}$$

3.

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

4.

- Rechte Seite: Definition der bedingten W'keiten anwenden:
  $$P(B \mid A_1) = \frac{P(A_1 \cap B)}{P(A_1)} \quad \Rightarrow \quad P(A_1 \cap B) = P(B \mid A_1)P(A_1)$$
- Entsprechende Formel gilt für $P(A_2 \cap B)$
- Oben einsetzen: Gesetz der totalen W'keit für $k = 2$:
  $$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) \end{aligned}$$

Beispiel: Spam-Mail

- Teilen Emails in drei Kategorien ein:

$$A_1 : \text{„spam"}, \quad A_2 : \text{„niedrige Priorität"}, \quad A_3 : \text{„hohe Priorität"}$$

- Aus früheren Beobachtungen bekannt:

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

- 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 Worten 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*)

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

- 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

$$P(X < 170)$$

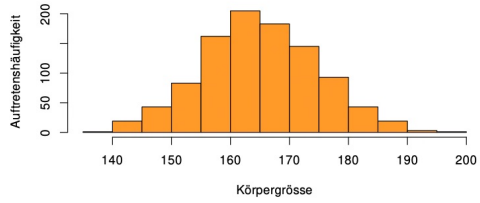
  entspricht
- Dies ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine gerundete Körpergrösse *kleiner als* 170 cm hat. Die Körpergrösse 170 cm gehört hier *nicht* dazu.
- Wichtig: Es gilt z. B.

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

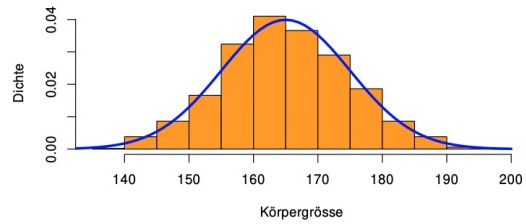
- W'keit eine zufällig eine Person auszuwählen, die kleiner als 160 cm ist, ist kleiner gleich, als dass sie kleiner als 160 cm ist
- Wichtig: Alle W'keiten der Verteilung aufaddiert, ergibt 1:

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

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



- *Form* des Histogrammes sehr typisch $\rightarrow$ 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 $\rightarrow$ Fläche von Histogramm 1
- Blaue Kurve heisst *Normalverteilungskurve*

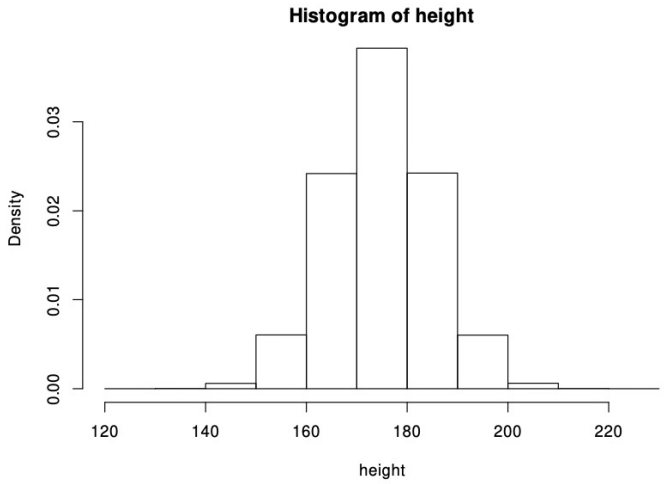
Von diskreter zur kontinuierlicher W'keitsverteilung

1.

- 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

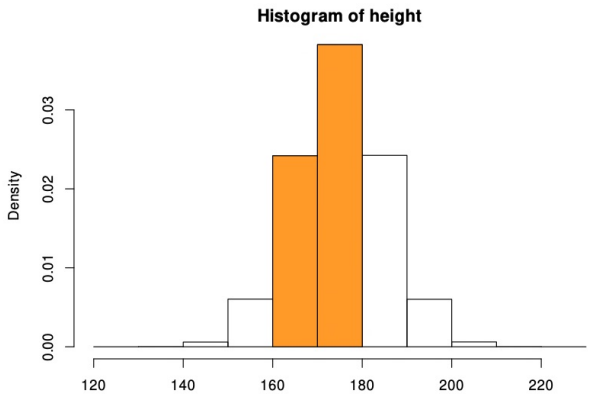
2.

- Histogramm:



3.

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

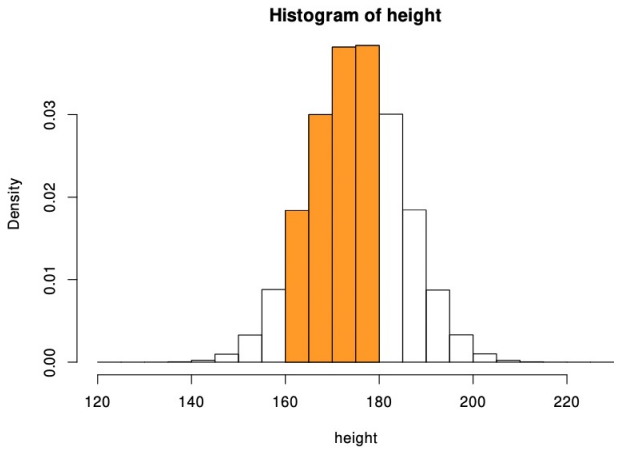


4.

- 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 $\rightarrow$ 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

5.

- Histogramm mit Balkenbreite 5 cm

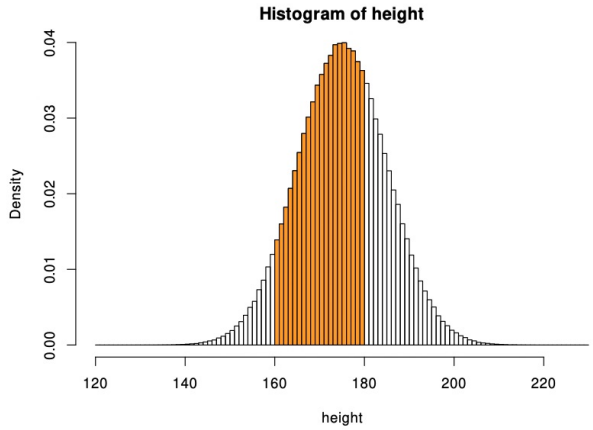


6.

- 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

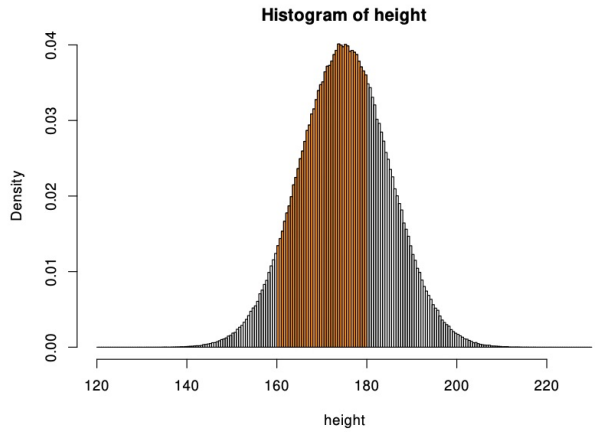
7.

- Histogramm mit Balkenbreite 1 cm



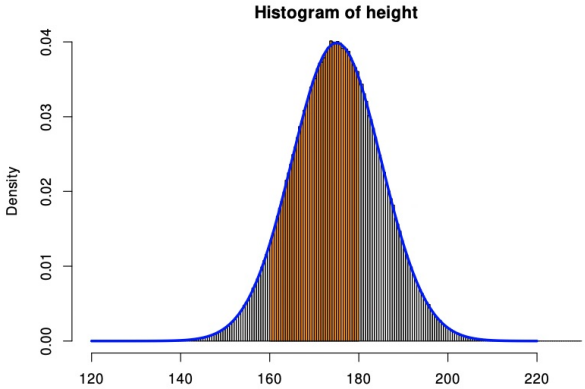
8.

- Histogramm mit Balkenbreite 0.5 cm



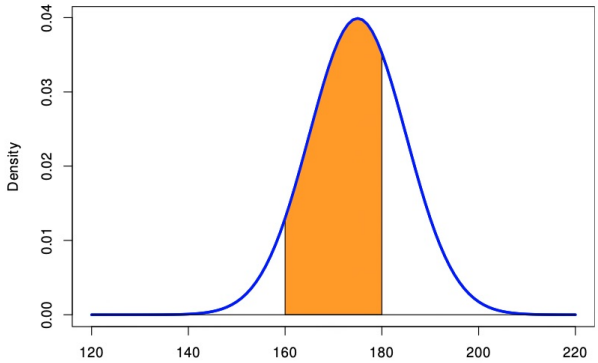
9.

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



10.

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



11.

- 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

SW07 – Normalverteilung

Kontinuierliche Messdaten

- In vielen Anwendungen: Keine diskreten Daten, sondern *Messdaten*
- Messdaten können jeden Wert in einem bestimmten Bereich annehmen
- Bsp: Gemessene Körpergrössen (in cm) können *jeden* Wert im Intervall $[0, 500]$ annehmen
- Also auch:
$$145.325\,986\,54\ldots$$
- Voraussetzung: Beliebig genaue Messung möglich

Intervalle

- Intervall, wo die Grenzen innerhalb oder ausserhalb des Intervalls sein sollen $\rightarrow$ 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 den einen Punkt 1.2 der Zahlengeraden mehr
- Statistik: Spielt keine Rolle, ob 1. oder 2. Intervall verwendet wird

Definitionen

- *Wertebereich W_X einer Zufallsvariable* $\rightarrow$ 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 Wertebereich:
$$W_X = \mathbb{R},\ \mathbb{R}^+ \quad \text{oder} \quad [0, 1]$$
- Letzter Fall: Zahlen 0 und 1 *und* alle Zahlen dazwischen

Punktwahrscheinlichkeit 0

- W'keitsverteilung einer *diskreten* Zufallsvariablen: „Punkt"-W'keiten $P(X = x)$ für alle möglichen x im Wertebereich
- Vorher: X Körpergrösse auf cm gerundet $\rightarrow$ $P(X = 174)$ nicht 0
- Stetige Zufallsvariable X: Für alle $x \in W_X$ gilt:
$$P(X = x) = 0$$
- Folgerung: W'keitsverteilung von X kann *nicht* mittels „Punkt"-W'keiten beschrieben werden

Beispiel: Körpergrösse

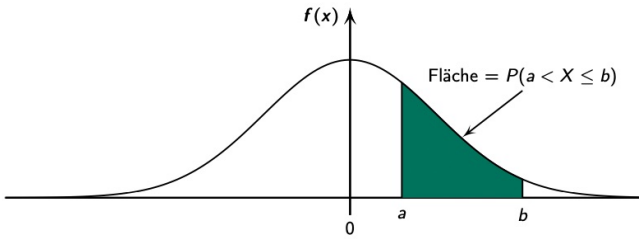
- Messen Körpergrösse von Personen
- W'keit *genau* eine Körpergrösse von $182.254\,680\,895\,434\ldots$ cm zu messen ist gleich 0:
$$P(X = 182.254\,680\,895\,434\ldots) = 0$$
- Verwendung der W'keit einen exakten Messwert zu messen, bringt nichts für W'keitsverteilung der Körpergrösse
- Summe aller W'keiten müsste 1 ergeben $\rightarrow$ 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 W'keitsdichte $f(x)$ gelten folgende Eigenschaften:

- Es gilt:
$$f(x) \geq 0$$
Das heisst, die Kurve liegt überhalb der x-Achse
- W'keit
$$P(a < X \leq b)$$
entspricht der Fläche zwischen a und b unter $f(x)$
- Die gesamte Fläche unter der Kurve ist 1:
  - Dies ist die W'keit, dass *irgendein* Wert gemessen wird

- Skizze:



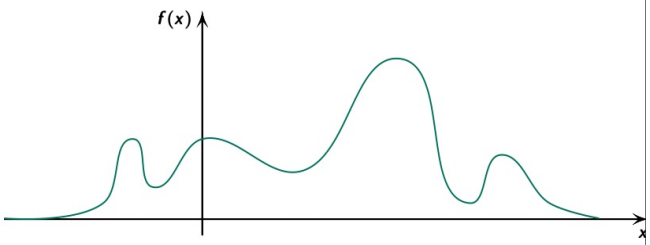
- Wichtig: Zusammenhang zwischen W'keit und Flächen:

Merkregel

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

Beispiel: Wahrscheinlichkeitsdichtefunktion

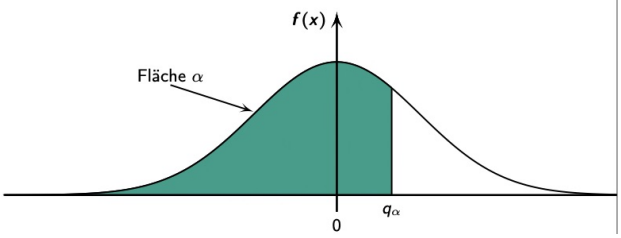
- Skizze:



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

Quantile

- Stetige Verteilungen: α-Quantil q_α derjenige Wert, wo die Fläche (W'keit) unter der Dichtefunktion von $-\infty$ bis q_α gerade α entspricht
- 50 %-Quantil: *Median*
- Skizze:



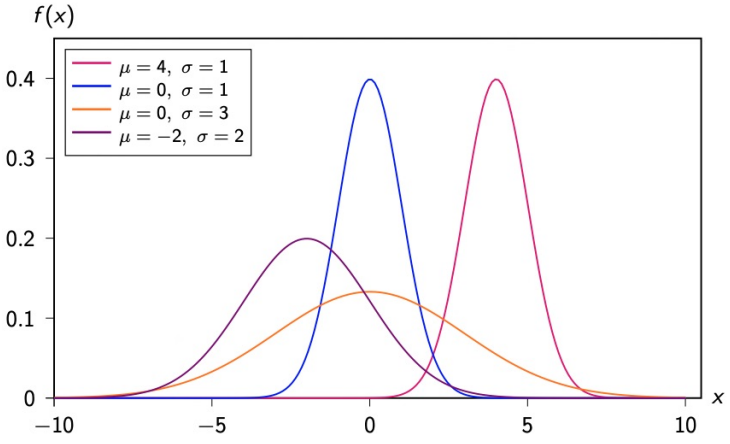
Beispiel: Körpergrösse

- 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): $X \sim \mathcal{N}(\mu, \sigma^2)$

- 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*
$$\mathrm{E}[X] = \mu$$
- *Varianz*
$$\mathrm{Var}(X) = \sigma^2$$

Normalverteilung: Illustration Dichten



Eigenschaften der Normalverteilung

- 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

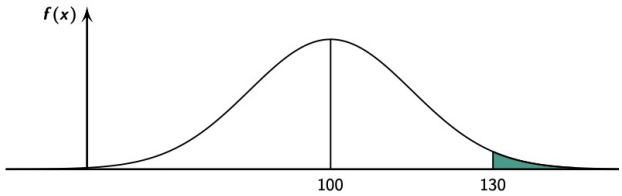
1. Beispiel mit R: Verteilung von IQ

- 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:

$$X \sim \mathcal{N}(100, 15^2)$$

2. Beispiel

- 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:



3.

- 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 $P(X > 130)$

4.

- Aber: Gesamtfläche unter der Kurve 1 $\rightarrow$ Gesuchte W'keit wie folgt schreiben:

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

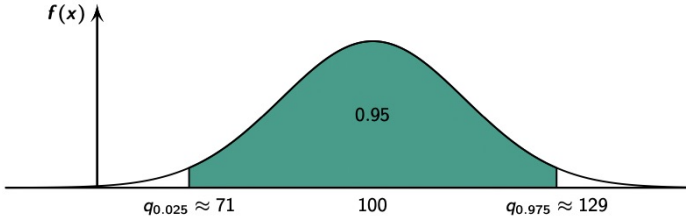
- Berechnung mit R:

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

- Also rund 2 % der Bevölkerung ist hochbegabt

5.

- 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.

6.

- W'keiten gegeben $\rightarrow$ 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
```

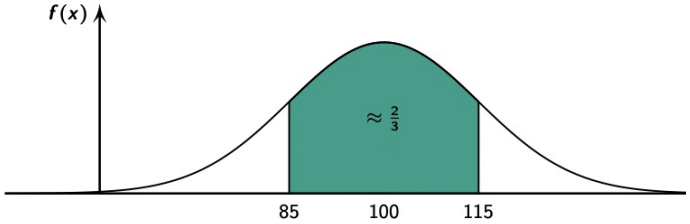
- 95 % der Menschen haben einen IQ zwischen ungefähr 70 und 130
- Entspricht Abstand von etwa 2 Standardabweichungen vom Mittelwert $\mu = 100$.

7.

- Wieviel Prozent der Bevölkerung liegen innerhalb einer Standardabweichung vom Mittelwert liegen?
- Gesucht W'keit:

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

- W'keit als Fläche:



8.

- 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

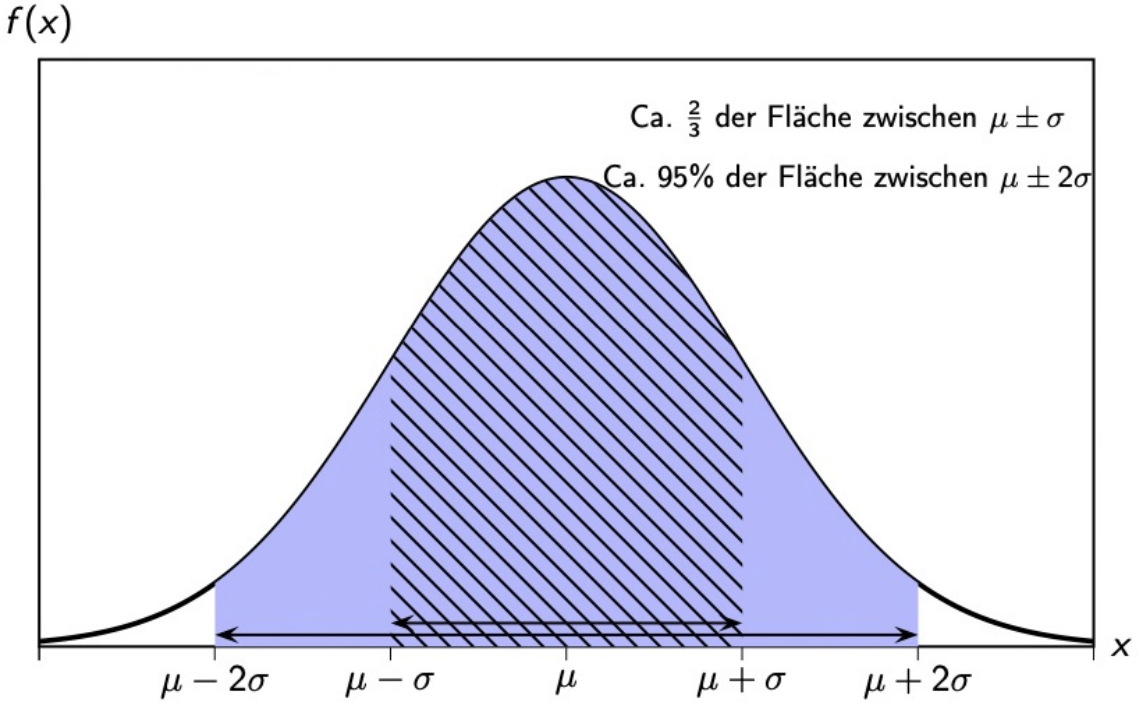
- Letzte Resultat aus Beispiel gilt für alle Normalverteilung $\mathcal{N}(\mu, \sigma^2)$
- Die W'keit, dass eine Beobachtung eine höchstens Standardabweichung vom Erwartungswert abweicht, ist etwa $\frac{2}{3}$:

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

- 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



SW08 – Gesetz der grossen Zahlen, zentraler Grenzwertsatz

Funktionen von mehreren Zufallsvariablen

- Bis jetzt: Verteilung *einer* Zufallsvariable (ZV)
- Aber: üblicherweise wird die *gleiche* Grösse mehrmals gemessen
- Bsp: Misst Gewicht mehrmals
- Allgemein: Messungen $x_1, x_2, \ldots, x_n$ als Realisierungen der ZV auffassen:
$$X_1, \ldots, X_n$$
- X_i: Die i-te Wiederholung von Zufallsexperiment

Beispiel

- 20 Messungen der Wasserverschmutzung in einem See
- Messungen (konkrete Werte):
$$x_1,\ x_2,\ \ldots,\ x_{20}$$
- Realisierungen der ZV:
$$X_1,\ X_2,\ \ldots,\ X_{20}$$
- Annahme: 20 ZV mit gleichen W'keitsverteilungen
- Wasserproben: Alle aus demselben See mit identischer Methode gemessen
- Interessant: *Durchschnitt* dieser Messungen und die Verteilung der zugehörigen ZV

Summe und Durchschnitt

- Gegeben ZV:
$$X_1, \ldots, X_n$$
- *Summe*:
$$S_n = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$$
- *Arithmetisches Mittel*:
$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} S_n$$

Beispiel: Warum ist der Durchschnitt wichtig?

- Untersuchen, ob Angabe 500 ml Inhalt einer Pet-Flasche gilt
- Kaufen *eine* Flasche: Messen Inhalt 495.21 ml
- Das ist weniger als 500 ml, aber ist es zuwenig?
- Bei einer Flasche möglich
- Idee: Kaufen 100 Flaschen und messen Inhalt
- Durchschnitt 463.21 ml
- Scheint eindeutig zu wenig: Angabe 500 ml kann nicht stimmen
- Genaues Vorgehen: Siehe Hypothesentest

1.

Kennzahlen von S_n und $\overline{X}_n$

- Annahme:
$$X_1, \ldots, X_n \text{ i.i.d.}$$
- Zweites „i" in i.i.d.: X_i dieselbe Verteilung mit denselben Kennzahlen:
$$\mathrm{E}(X_i) = \mu \quad \text{und} \quad \mathrm{Var}(X_i) = \sigma_X^2$$
- Gesucht: Erwartungswert und Varianz für:
  - Summe S_n:
  $$S_n = X_1 + X_2 + \ldots + X_n$$
  - Durchschnitt $\overline{X}_n$:
  $$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

3.

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

2.

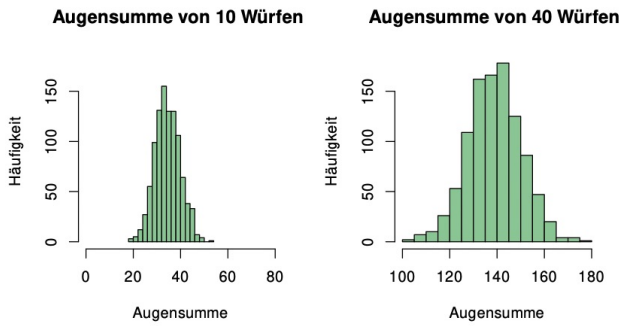
Graphisches Beispiel

- Werfen einen fairen Würfels
- X: ZV für geworfene Augenzahl
- Erwartungswert:
$$\mathrm{E}(X) = \mu = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$
- Varianz:
$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\right) \\ &= 2.92\end{aligned}$$
- R:

```
x <- c(1, 2, 3, 4, 5, 6)

ave <- mean(x)

var <- mean((x - ave)^2)

var
## [1] 2.916667
```

1

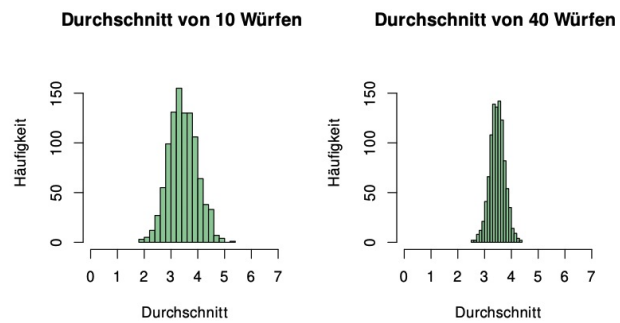
Histogramme



2

Feststellungen

- 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:
$$\mathrm{E}(S_n) = n\mu$$
- Stimmt auch (ohne Beweis)
- Varianz/Standardabw. nimmt mit zunehmender Anzahl Würfen *zu*
- Gesetz (ohne Beweis) angegeben:
$$\mathrm{Var}(S_n) = n\,\mathrm{Var}(X), \qquad \sigma_{S_n} = \sqrt{n}\sigma_X$$
- Dasselbe mit dem Durchschnitt $\overline{X}_n$
- Histogramme:



3

Bemerkungen

- Standardabweichung von $\overline{X}_n$: *Standardfehler* des arithmetischen Mittels
- Standardabweichung der Summe: Wächst mit wachsendem n, aber langsamer als die Anzahl Beobachtungen n
- D. h.: Kleinere Streuung für wachsendes n
- Erwartungswert von $\overline{X}_n$: Gleich demjenigen einer einzelnen ZV X_i, die *Streuung nimmt jedoch ab mit wachsendem n*

Standardfehler

Standardfehler

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

$$\sigma_{\overline{X}_n} = \frac{1}{\sqrt{n}}\sigma_X$$

Um *Standardfehler* zu halbieren, braucht man also *viermal* so viele Beobachtungen

Dies nennt man auch das $\sqrt{n}$-*Gesetz*

Feststellungen

- Beide Histogramme: Grösste Häufigkeit bei 3.5, also μ
- Vermutung (stimmt, aber ohne Beweis):
$$\mathrm{E}(\overline{X}_n) = \mu$$
- Varianz/Standardabw. nimmt mit zunehmender Anzahl Würfen *ab*
- Gesetz (ohne Beweis):
$$\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X)}{n}; \qquad \sigma_{\overline{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$
- Oben gemachte Beobachtungen gelten allgemein

Zentraler Grenzwertsatz

- Bekannt: Kennzahlen von S_n und $\overline{X}_n$
- Unbekannt: Verteilung von S_n und $\overline{X}_n$
- 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 $\overline{X}_n$ verteilt?
- Vermutung wegen Slides 9 und 12: Beide normalverteilt
- Dies ist die Aussage des *Zentralen Grenzwertsatzes*
- Simulation der Aussage, kein Beweis

Allgemein

- Annahme:
$$X_1, \ldots, X_n \text{ i.i.d.}$$
- Es gilt:

Kennzahlen von S_n

$$\mathrm{E}(S_n) = n\mu$$
$$\mathrm{Var}(S_n) = n\,\mathrm{Var}(X_i)$$
$$\sigma(S_n) = \sqrt{n}\sigma_X$$

Kennzahlen von $\overline{X}_n$

$$\mathrm{E}(\overline{X}_n) = \mu$$
$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma_X^2}{n}$$
$$\sigma(\overline{X}_n) = \frac{\sigma_X}{\sqrt{n}}$$

1.

Simulation von $\overline{X}_n$

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

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

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

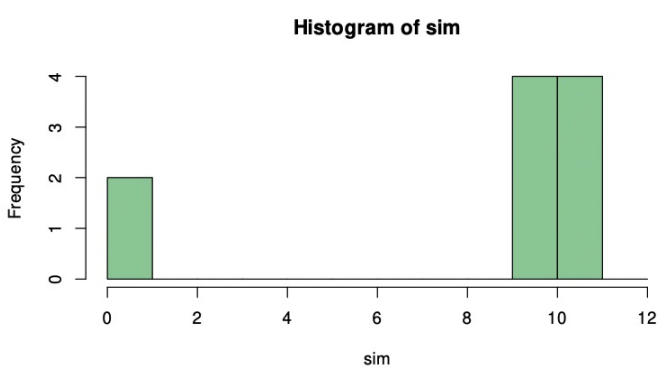
```
var.X <- sum((werte - ew)^2 * 1/3)

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
```

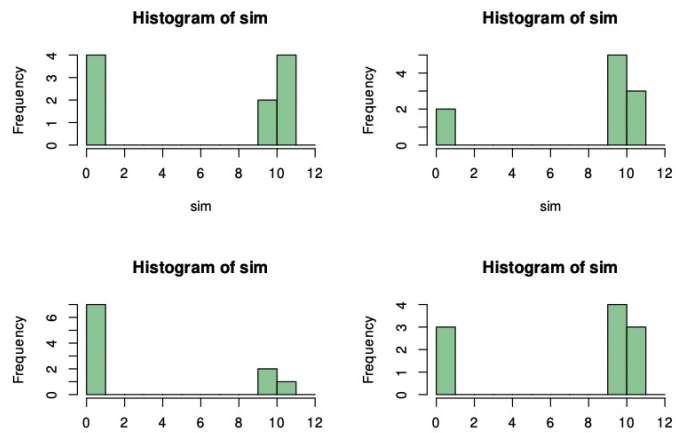
- `sample`: Zieht zufällig Zahlen aus `werte`
- `replace = T`: Legt die Zahl nach dem Ziehen wieder zurück

```
# Histogramm mit diesen 10 Werten
hist(sim, col = "darkseagreen3", breaks = 0:12)
```



2.

- 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
```

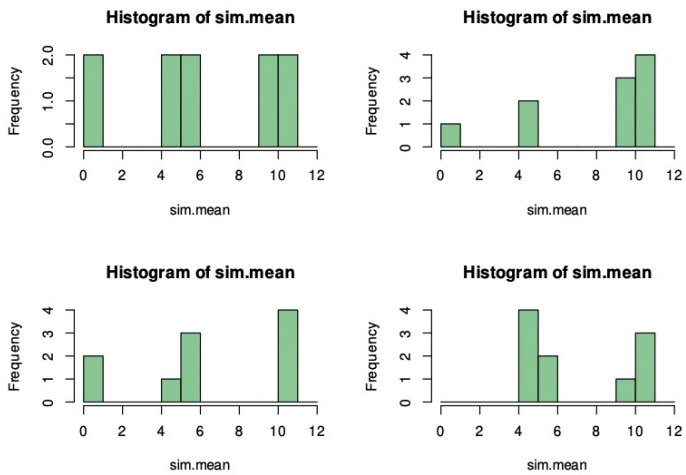
- Neben Zahlen 0, 10, 11: Auch Zahlen 5, 5.5 und 10.5 können vorkommen
- Histogramm:

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

Weiter auf nächster Seite

3.

- 4 Histogramme: Alle verschieden



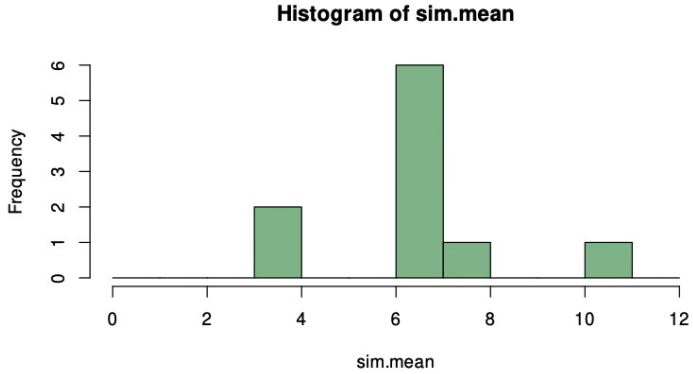
- Jeder Versuch sieht anders aus
- Aber: Tendenzen zeichnen sich ab
- 0 weniger oft vertreten, da doppelte 0 nur mit W'keit $\frac{1}{9}$ vorkommt

- 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
```

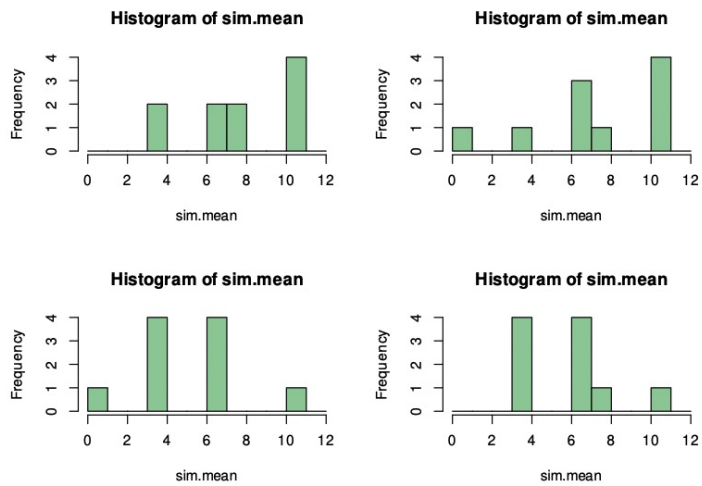
- Histogramm:

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

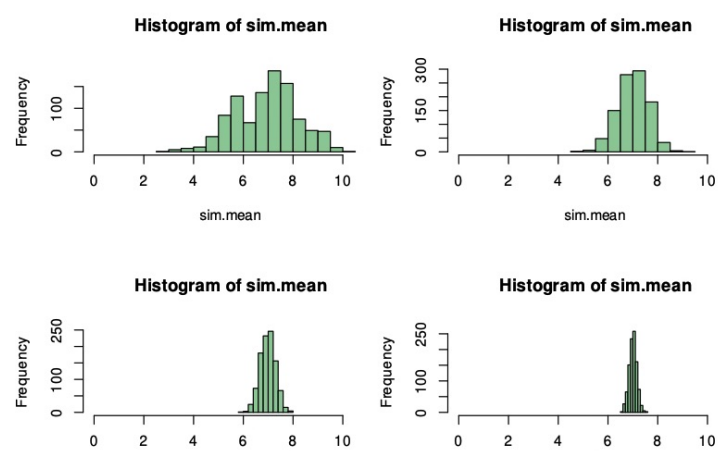


4.

- 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:

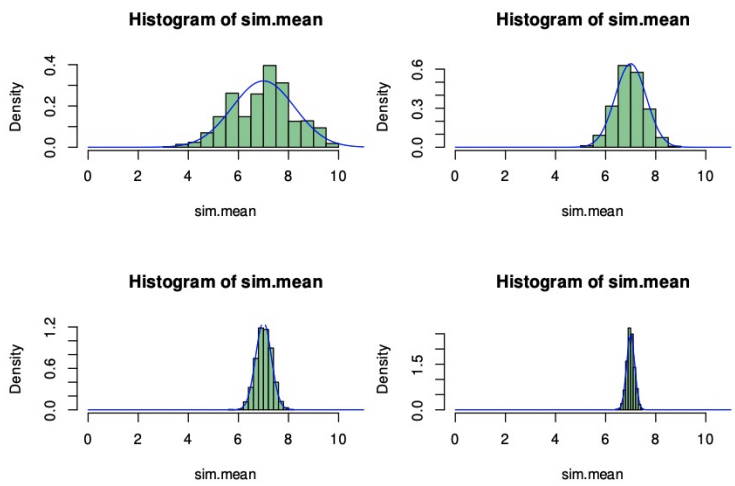


- Bei genauerem Hinsehen fällt auf:
  - Werte häufen sich um den Erwartungswert 7
  - Standardabweichung wird kleiner: Halbiert sie sich etwa beim Vervierfachen der Anzahl Versuche
  - Histogramme scheinen einer Normalverteilung zu folgen
- Zeichnen noch die jeweiligen Dichtekurven für

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

5.

Histogramme



6.

- 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* $\overline{X}_n$ (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

$X_1, \ldots, X_n$ i.i.d. mit irgendeiner Verteilung mit Erwartungswert μ und Varianz σ^2, dann gilt (ohne Beweis):

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

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

- 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?

Lösung

- 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:

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

- Gesucht:

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

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

Beispiel

- 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?

Lösung

- X_i: Zufallsvariable für die Lebensdauer des Teils i
- Es gilt $\mu = 100$ und $\sigma_X = 20$
- Annahme i.i.d.
- Betrachten durchschnittliche Lebensdauer $\overline{X}_{16}$
- Annähernd verteilt wie:

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

- Gesucht:

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

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

- Gesucht:

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

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

SW09 – Hypothesentest, z-Test, t-Test

Hypothesentest: Problemstellung, Beispiele

- Hypothesentests: Wichtiges statistisches Mittel um zu entscheiden, ob eine Messreihe zu einer gewisse Grösse „passt"
- Brauerei bestellt neue Abfüllmaschine für 500 ml Büchsen
- Abfüllmaschine füllt *nie genau* 500 ml ab, sondern nur *ungefähr*
  - Mal einen Tropfen mehr, mal einer weniger
- Für Brauerei wichtig, dass Abfüllmaschine möglichst genau abfüllt:
  - Füllt Maschine zuviel ab, so ist dies schlecht für Brauerei, da sie zuviel Bier für den angegebenen Preis verkauft
  - Füllt sie zuwenig ab, sind Kunden unzufrieden, da sie für den angegebenen Preis zuwenig 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 499.57 ml
- 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"?

Allgemein

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

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

- Zwei Kennzahlen der Zufallsvariablen X_i sind:

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

- Typischerweise sind diese (und andere) Kennzahlen unbekannt
- Ziel: Rückschlüsse darüber aus den Daten

Beispiele

- 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

- Ziel: Standardisiertes, reproduzierbares Verfahren einzuführen, mit dem wir entscheiden können, ob der Mittelwert einer Messreihe zu einem bestimmten „wahren" Mittelwert μ 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!

Beispiel

- 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

Schätzungen

1.

- (Punkt-) Schätzungen für den Erwartungswert und die Varianz sind:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

2.

Beispiel: Waage A

- Schätzungen für den Mittelwert μ und die Varianz σ_X^2:

$$\widehat{\mu} = 80.02 \qquad \text{und} \qquad \widehat{\sigma}_X^2 = 0.024^2$$

- 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
```

3.

- 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:

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

- 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

4.

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

8.

- 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

5.

- 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

6.

- Beispiel vorher: geschätzte Mittelwerte alle sehr nahe bei $\mu = 80$
- Allerdings sind auch folgende Fälle möglich:

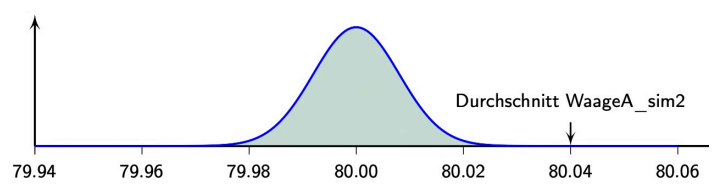
```
set.seed(1450070)
waageA.sim2 <- rnorm(n = 6, mean = 80, sd = 0.02)

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
```

7.

- Mittelwert dieser Messreihe verteilt wie (ZGWS):

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

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



9.

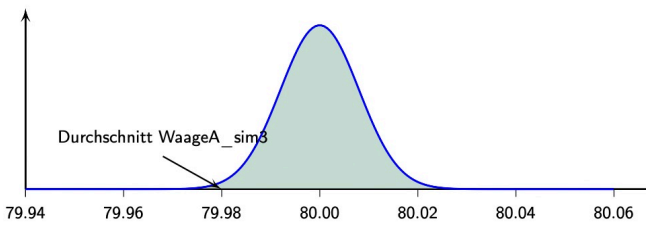
- Ein weiteres Beispiel:

```
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
```

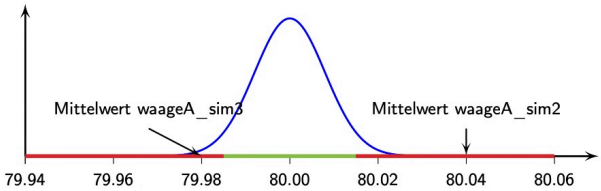
- Mittelwert etwa 3 Standardabweichungen unter 80
- Dies ist zwar immer noch weit weg, aber nicht so stark wie im vorher

10.

- 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" μ ist
- Skizze:



1. 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 $X_1, X_2, \ldots X_6$, wobei X_i eine kontinuierliche Messgrösse ist. Es soll gelten:

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

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

Nullhypothese

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

Alternativhypothese

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

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

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

2.

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

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

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

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

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

Teststatistik

Verteilung der Teststatistik T unter der Nullhypothese H_0:

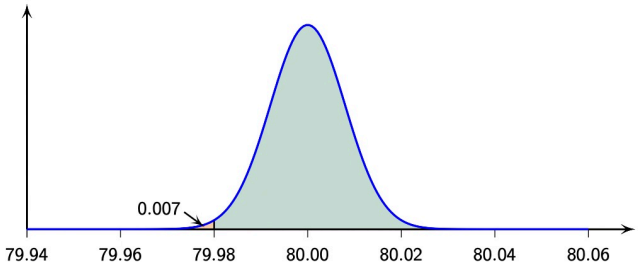
$$T = \overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

- Erhalten für die W'keit

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

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

- Skizze:



Weiter auf nächster Seite

3.

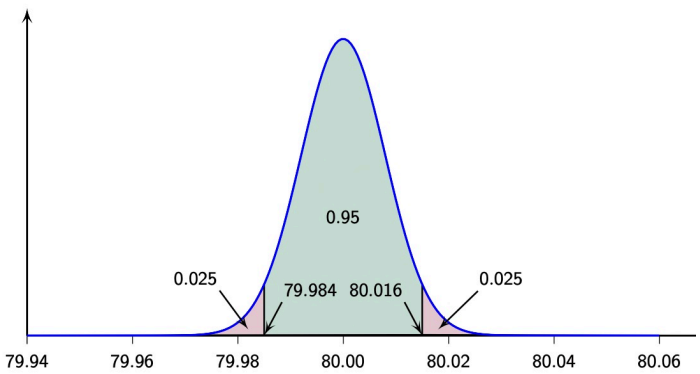
- Diese W'keit ist klein: 0.7 %
- Ist sie aber *zu* klein?
- Nun kommt eine *Abmachung*: Es hat sich als praktisch erwiesen, diese Grenze, was zu klein ist und was nicht bei 2.5 % festzulegen
- Warum dies 2.5 % sind, kommt gleich
- Gemäss dieser Abmachung ist:

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

- Geschätzer Mittelwert $\widehat{\mu} = 79.98$ 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 Teilen 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 α

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

$$\alpha = 0.05$$

4.

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

$$H_0: \quad \mu = 80$$

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

Verwerfungsbereich

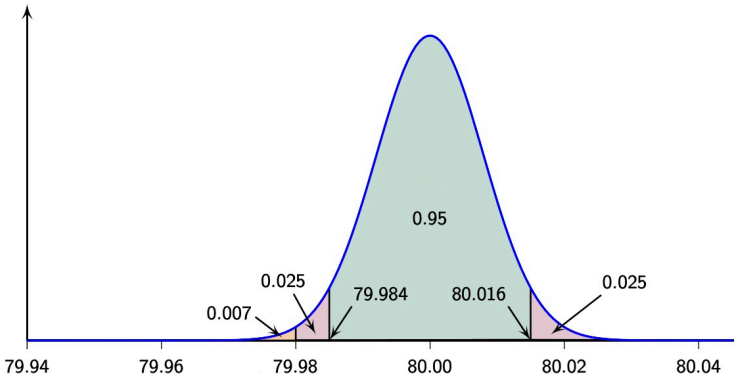
$$K = (-\infty, 79.984] \cup [80.016, \infty)$$

- 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 μ nicht 80 ist
- Mit Messreihe überprüfen, ob deren Mittelwert im Verwerfungsbereich liegt oder nicht
- Machen den sogenannten

Testentscheid

- In Beispiel oben
$$\overline{X}_6 = 79.98 \in K$$
- 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:



5

- 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

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

```
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 $P(\overline{X}_6 > 80.04)$ 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

Beispiel: Abfüllanlage

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

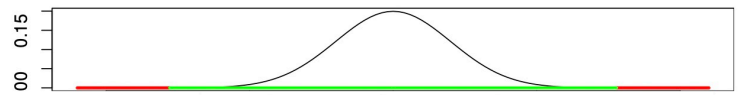
Weiter auf nächster Seite

Bemerkungen

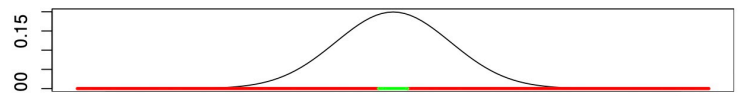
- Warum wurde Verwerfungsbereich nach oben und nach unten aufgeteilt, wenn man schon weiss, dass der gemessene Mittelwert kleiner als $\mu = 80$ ist?
- *Vor* der Messung war dies nicht bekannt
- Der gemessene Mittelwert hätte also durchaus auch grösser als $\mu = 80$ sein können
- Man spricht in diesem 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

Wahl von Signifikanzniveau α

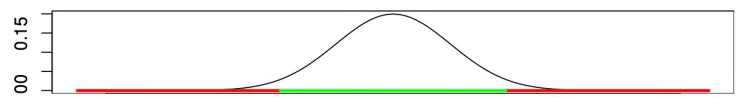
- Graphik: $\alpha = 0.0001$ (nahe bei 0)



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



- Graphik: $\alpha = 0.05$



- Ist α 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 α 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

- *Modell*
  X_i: Inhalt der i-ten Büchse

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

- *Nullhypothese*

$$H_0: \qquad \mu_0 = 500$$

- *Alternativhypothese*

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

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

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

- *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

2.

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

$$\overline{X}_8 \sim \mathcal{N}\left(180, \frac{10^2}{8}\right)$$

- 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 $P(\overline{X}_6 < 171.54)$:

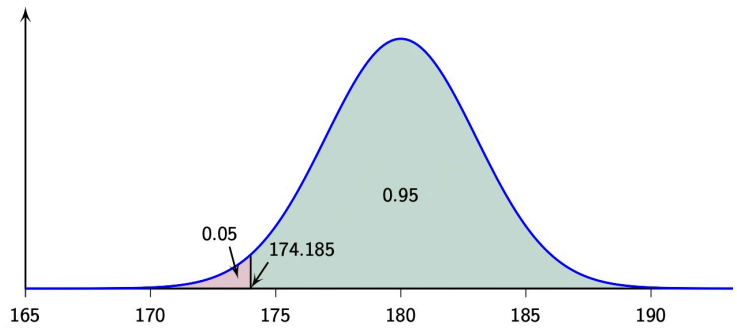
$$P(\overline{X}_6 < 171.54) = 0.008$$

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))
## [1] 0.008359052
```

Weiter auf nächster Seite…

1. 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



- Erwartung, dass der wahre Mittelwert tiefer als $\mu = 180$ ist
- *Einseitiger* Test
- 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

$$X_1, \ldots, X_8 \text{ i.i.d. } \sim \mathcal{N}(\mu, 10^2)$$

- Gehen davon aus, dass der wahre Mittelwert wirklich 180 cm ist
- *Nullhypothese*

$$H_0: \qquad \mu_0 = 180$$

- *Alternativhypothese*

$$H_A: \qquad \mu < \mu_0 = 180$$

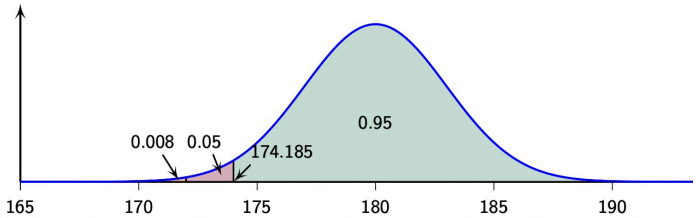
- Testen ob jetzt der Wert

$$P(\overline{X}_8 < \overline{x}_8) < 0.05$$

  ist oder nicht
- Verwerfungsbereich ist hier also einseitig nach unten
- Abbildung oben: Verwerfungsbereich für $n = 8$ pink eingezeichnet

...

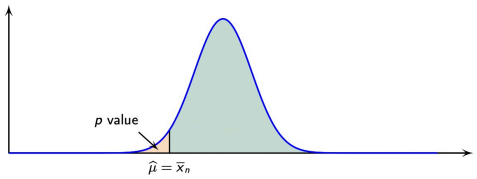
- Abbildung:



- 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.

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

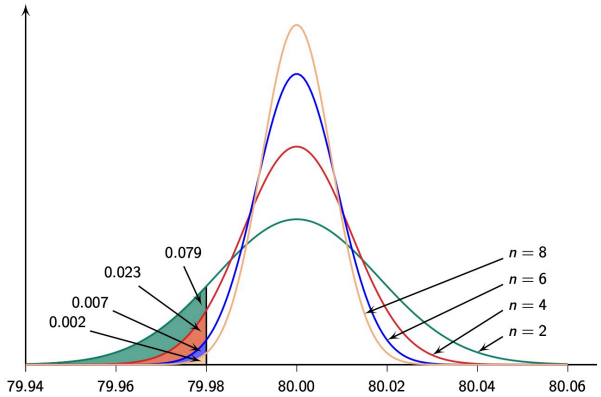
Bei einem vorgegebenen Signifikanzniveau α (z.B. $\alpha = 0.05$) gilt aufgrund der Definition des p-Werts für einen einseitigen Test:

- 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 ≈ 0.05 : schwach signifikant, "."
p-Wert ≈ 0.01 : signifikant, "$*$"
p-Wert ≈ 0.001 : stark signifikant, "$**$"
p-Wert $\leq 10^{-4}$: äusserst signifikant, "$***$"

- Abbildung:



Einfluss der Anzahl Messungen auf Verwerfungsbereich

- Beispiel Waage von früher
- Messreihen verschiedener Länge n, alle mit geschätztem Mittelwert $\widehat{\mu} = 79.78$
- Bestimmen für alle Messreihen den Wert

$$P(\overline{X}_n \leq 79.98) \qquad \text{mit} \qquad \overline{X}_n \sim \mathcal{N}\left(80, \frac{0.02^2}{n}\right)$$

- Ist dieser Wert grösser als 0.025, dann wird die Nullhypothese nicht verworfen, ansonsten schon
- $n = 2$:

$$P(\overline{X}_2 \leq 79.98) = 0.079 > 0.025$$

```
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
- $n = 4$:

$$P(\overline{X}_4 \leq 79.98) = 0.022 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(4))
## [1] 0.02275013
```

- Hier wird die Nullhypothese (knapp) verworfen
- $n = 6$:

$$P(\overline{X}_6 \leq 79.98) = 0.007 < 0.025$$

```
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 $n = 8$:

$$P(\overline{X}_6 \leq 79.98) = 0.002 < 0.025$$

```
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

$$P(\overline{X}_n \leq 79.98)$$

immer kleiner

- Grund: Standardabweichung mit grösser werdendem n kleiner wird
- Normalverteilungskurven werdem schmaler
- D.h.: je mehr Messungen wir haben, umso gewichtiger ist eine Abweichung von wahren Mittelwert

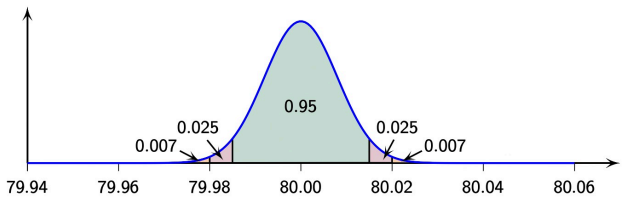
p-Wert für zweiseitigen Test

- Haben den p-Wert für einseitige Tests definiert
- Wie sieht nun aber der p-Wert für zweiseitige Tests aus?
- Beispiel von früher:

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

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:

$$p\text{-Wert} = 2 \cdot P(\overline{X}_6 \leq 79.98) = 0.014$$

- Dieser p-Wert dann mit dem Signifikanzniveau vergleichen
- Computersoftware gibt den p-Wert *immer* auf Signifikanzniveau an

2. R

- 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

Beispiel

- Datensatz aus normalverteilten Datenpunkten $x_1, \ldots, x_{20}$

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: $x_1, x_2, \ldots, x_{20}$ Realisierungen von

$$X_i \sim \mathcal{N}(5, \sigma_X^2)$$

- σ_X unbekannt $\rightarrow$ σ_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
```

- `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`

1. 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 $\rightarrow$ Nur andere Verteilung
- Wie vorher Annahme: Daten Realisierungen von

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

- Praxis: Annahme, dass σ_X bekannt ist, meist unrealistisch
- Können aber σ_X aus Daten schätzen $\rightarrow$ $\widehat{\sigma}_X^2$

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

- Zusätzliche Unsicherheit $\rightarrow$ Verteilung der Teststatistik ändern

t-Verteilung

Die Verteilung der Teststatistik beim t-Test unter der Nullhypothese

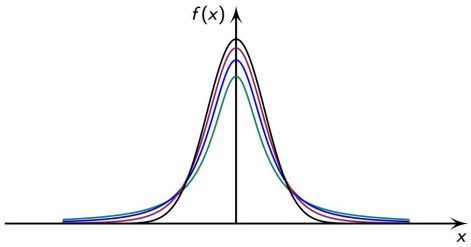
$$H_0: \quad \mu = \mu_0$$

ist gegeben durch

$$T = \overline{X}_n \sim t_{n-1}\left(\mu, \frac{\widehat{\sigma}_X}{\sqrt{n}}\right)$$

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

Beispiel: Waage A

- Schätzen Standardabweichung σ_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

$$H_0: \quad \mu_0 = 180$$

  verwerfen

- Alternativhypothese

$$H_A: \quad \mu_0 < 180$$

  annehmen

- Aussage des Bundesamtes für Statistik stimmt also statistisch signifikant (sehr wahrscheinlich) *nicht*

SW10 – Vertrauensintervall, Zweistichprobentest, Wilcoxon-Test

Vertrauensintervall

1.

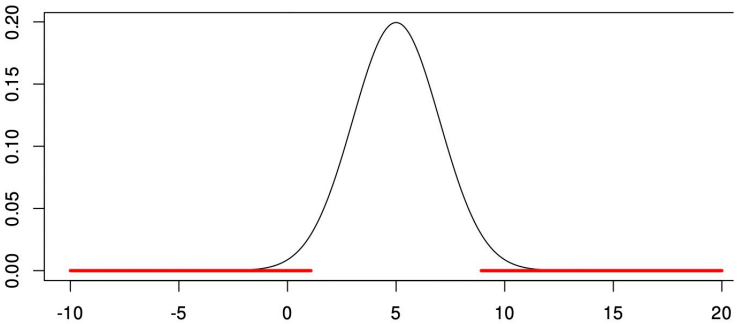
- Bei Punktschätzung für Mittelwert μ einer Messreihe $\rightarrow$ *Einziger* Zahlwert
- Wissen aber nicht, wie nahe dieser geschätzte Mittelwert beim wahren, aber meist unbekannten, Mittelwert der Verteilung der Messreihe liegt
- Vertrauensintervall: Intervall, das angibt , wo, grob gesagt, der wahre Mittelwert mit einer bestimmten vorgegebenen W'keit liegt
- Wollen mit Beispiel Vertrauensintervall verschaulichen

Beispiel

- Bestimmung Verwerfungsbereiches beim z-Test: Gehen vom einem wahren (aber unbekannten) Wert μ 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
```

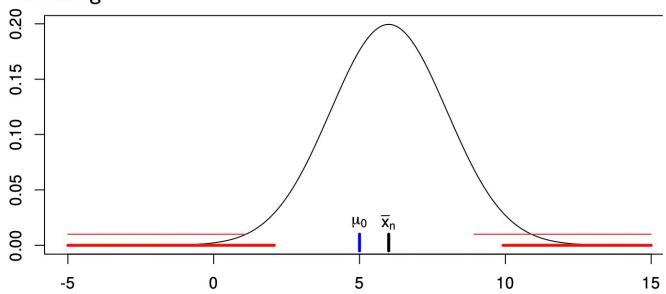
- Abbildung: Normalverteilungskurve mit dem Verwerfungsbereich rot:



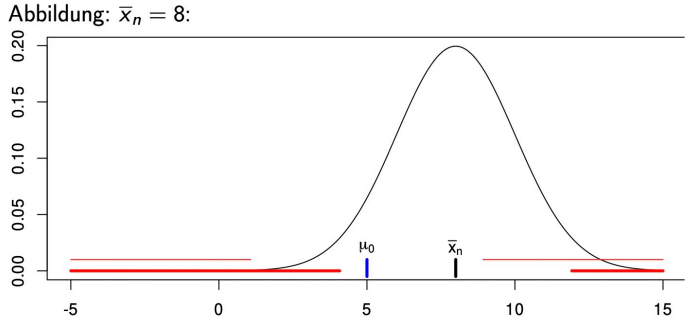
- Liegt nun $\overline{x}_n$ im Verwerfungsbereich (roter Bereich), dann wird die Nullhypothese H_0 verworfen
- Wahres μ praktisch immer unbekannt: Wert einfach angenommen
- Frage umkehren: Kennen $\overline{x}_n$ und fragen, für welche μ wird H_0 *nicht* verworfen
- Kann man rechnerisch herleiten, machen es aber nur graphisch
- Annahme $\mu_0 = 5$
- Gegeben $\overline{x}_n = 6$ und zeichnen Verwerfungsbereich für diesen Wert ein

2.

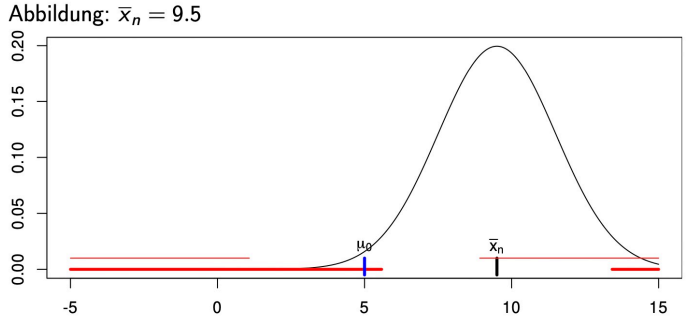
- Abbildung:



- Linien:
  - Dicke roten Linien: Verwerfungsbereich für $\overline{x}_n = 6$
  - Dünne roten Linien: Verwerfungsbereich für $\mu_0 = 5$
  - Vertikaler schwarze Strich: $\overline{x}_n = 6$
  - Vertikaler blauer Strich: $\mu_0 = 5$
- Stellen fest: $\overline{x}_n$ und μ_0 nicht innerhalb in einem der beiden Verwerfungsbereiche
- Idee: $\overline{x}_n$ vergrössern und $\mu_0 = 5$ konstant lassen
- Abbildung: $\overline{x}_n = 8$:



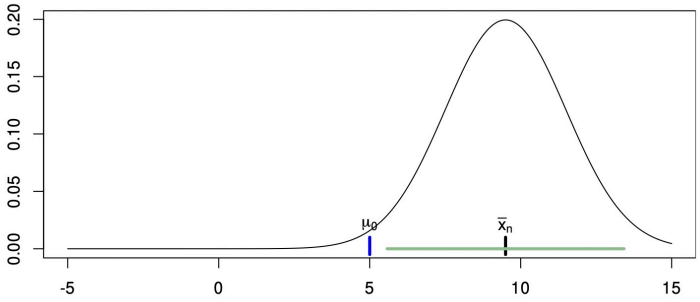
- Auch hier: $\overline{x}_n$ und μ_0 nicht innerhalb der beiden Verwerfungsbereiche
- Abbildung: $\overline{x}_n = 9.5$



- Andere Situation: $\overline{x}_n$ (schwarze Linie) im Verwerfungsbereich von $\mu_0 = 5$ (dünne blaue Linien)
- Nullhypothese H_0 nun verworfen
- Aber: $\mu_0 = 5$ im Verwerfungsbereich für $\overline{x}_n = 9.5$.

Weiter auf nächster Seite

- Andere Darstellung: Bereich, der *nicht* zum Verwerfungsbereich gehört
- Dieses Intervall heisst *Vertrauensintervall*



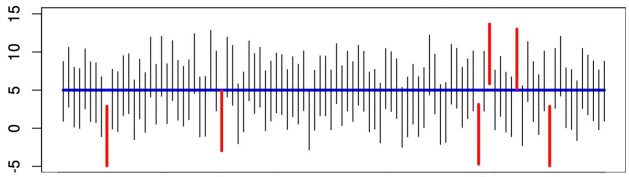
- Wert 5 liegt nicht im Vertrauensintervall:

```
qnorm(p = c(0.025, 0.975), mean = 9.5, sd = 2)
## [1]  5.580072 13.419928
```

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 μ also aus dem Vertrauensintervall, dann wird Nullhypothese verworfen
- Weiteren Interpretation des Vertrauensintervalls: Enthält alle μ_0's für die Nullhypothese *nicht* verworfen wird
- Es sagt uns also in welchem Intervall sich das wahre μ_0 befindet
- Gilt nicht absolut: Mit einer bestimmten W'keit liegt wahres μ_0 in diesem Intervall
- Hier: Wahres μ_0 liegt zu 95 % im Vertrauensintervall
- Sprechen deswegen auch von einem 95 %-Vertrauensintervall

Weiter auf nächster Seite…

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 $\mu_0 = 0$ nicht im Vertrauensintervall der Zufallszahlen
- 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 $\mu_0 = 0$ nicht im Vertrauensintervall der Zufallszahlen
- Das sind etwa 5 %

...

- Weitere Möglichkeit für Testentscheid:
  - Liegt μ_0 der Nullhypothese im Vertrauensintervall, so wird die Nullhypothese *nicht* verworfen
  - Liegt μ_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 μ zu 95 % in diesem Intervall liegt
- Mit Vertrauensintervall kann man ebenfalls Testentscheid durchführen

Beispiel: Waage A

- Nullhypothese
$$H_0: \quad \mu_0 = 80$$
- R-Output: Vertrauensintervall:
$$[80.00629,\ 80.03525]$$
- Zu 95 % liegt das wahre μ in diesem Intervall
- Aber $\mu_0 = 80$ *nicht* in diesem Intervall
- Zu 95 % Sicherheit ist das wahre μ *nicht* 80
- Nullhypothese wird verworfen und Alternativhypothese angenommen

Beispiel: Körpergrösse Frauen

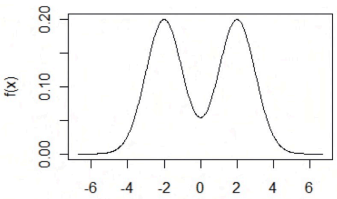
- Nullhypothese:
$$H_0: \quad \mu_0 = 180$$
- R-Output: Vertrauensintervall:
$$(-\infty,\ 169.382]$$
- Zu 95 % liegt das wahre μ in diesem Intervall
- $\mu_0 = 180$ *nicht* in diesem Intervall
- Mit 95 % Sicherheit ist das wahre μ *nicht* 80
- Nullhypothese verwerfen; Alternativhypothese annehmen

Bemerkung

- Je schmaler 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 μ 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. μ_0
- Annahme:
$$X_i \sim F \text{ iid}, \quad F \text{ ist symmetrisch}$$



- 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*

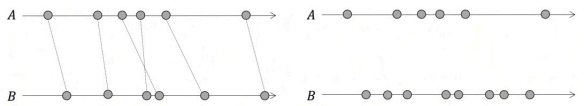
Unterscheidung gepaart versus ungepaarte Stichproben

Gepaarte Stichproben

- Jede Beobachtung einer Gruppe kann eindeutig einer Beobachtung der anderen Gruppe zugeordnet werden
- Stichprobengrösse ist in beiden Gruppen zwangsläufig gleich

Ungepaarte Stichproben

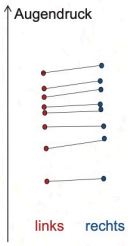
- Keine Zuordnung von Beobachtungen möglich
- Stichprobengrössen können verschieden sein (müssen aber nicht!)
- Man kann die eine Gruppe vergrössern, ohne dass man die andere vergrössert



Gepaarte versus ungepaarte Stichproben

- Beispiel: Augeninnendruck; ein Auge behandelt, das andere nicht (gepaarter Test ist angebracht)
- Gemäss Vorraussetzungen dürfte auch ein ungepaarter Test angewendet werden

$H_0: \mu_X = \mu_Y$



Ungepaart:
Intuition Teststatistik: $T = \frac{\overline{X}-\overline{Y}}{\widehat{\sigma_{\overline{X}}}}$

Gepaart:
Differenz $D_i = X_i - Y_i$
Teststatistik $T = \frac{\overline{D}}{\widehat{\sigma_{\overline{D}}}}$

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

$$\mathrm{E}(D) = \mu_D = 0$$

- *Kein* Unterschied!
- Falls Daten nicht normalverteilt $\rightarrow$ 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

Statistischer t-Test für ungepaarte Stichproben

1.

- *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

$$\mu_X = \mu_Y$$

2.

- 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

3.

- 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`

- 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

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
```

SW11 – Lineare Regression

- Fortsetzung von Block 3: Jetzt mit Hypothesentest
- Lineare Regression ist einer der Startpunkte in Machine Learning

Einführung, Beispiel

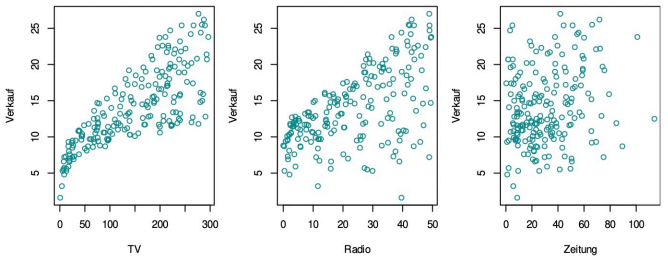
1.

- 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

2.

- Mathematische Sichtweise: Gesucht Funktion f, die Werbebudgets X_1 (`TV`), X_2 (`Radio`) und X_3 (`Zeitung`) den Verkauf Y ermittelt:

$$Y \approx f(X_1, X_2, X_3)$$

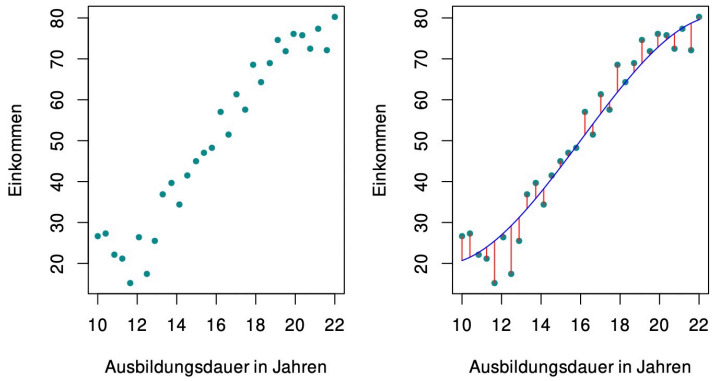
- Beziehung oben: Kein Gleichheitszeichen, da Streudiagramme keine Graphen einer Funktion darstellen
- Funktion f kann Zusammenhang zwischen X_1, X_2, X_3 und Y nur *approximativ* darstellen
- Bezeichnung:
  - Variable Y: *Zielgrösse*, *Outputvariable*,
  - X_1, X_2 und X_3: *Prädiktoren*,*erklärende Variable*
- Allgemein: Quantitative Zielgrösse Y und p verschiedene Prädiktoren $X_1, X_2, \ldots, X_p$
- Annahme: Es besteht *irgendein* Zusammenhang zwischen Y und $X_1, X_2, \ldots, X_p$
- Allgemeine Form:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- f irgendeine feste, aber *unbekannte* Funktion von $X_1, X_2, \ldots, X_p$
- Grösse ε: *Zufälliger Fehlerterm* unabhängig von $X_1, X_2, \ldots, X_p$ mit Mittelwert 0
- Bedeutung Fehlerterm ε $\rightarrow$ 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



Weiter auf nächster Seite…

...

- 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 ε
- 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 $\widehat{\ }$ gekennzeichnet
- $\widehat{Y}$: Schätzung der unbekannten Grösse Y
- $\widehat{f}$: 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 $X_1, X_2, \ldots, X_p$ einfach verfügbar, aber die Zielgrösse nicht
- In so einem Fall: Y schätzen durch

$$\widehat{Y} = \widehat{f}(X_1, X_2, \ldots, X_p)$$

- Fehlerterm im Mittel 0

Rückschlüsse auf f: Fragestellungen

- *Welche Inputvariablen werden mit dem Output assoziert?*
  - 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)

Weiter auf nächster Seite...

Beispiel

- Prädiktoren $X_1, X_2, \ldots, X_p$ 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 $X_1, X_2, \ldots, X_p$ 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 $\widehat{Y}$ als Vorhersage von Y hängt von zwei Grössen ab:
  - *Reduzibler Fehler*
  - *Irreduzibler Fehler*
- Allgemein: $\widehat{f}$ 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

$$\widehat{Y} = f(X_1, X_2, \ldots, X_p)$$

- Vorhersage Y enthält immer noch Fehler
- Liegt am Fehlerterm ε: Hängt nicht von $X_1, X_2, \ldots, X_p$ ab
- Variablilität von ε 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 ε, 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 $\rightarrow$ Für die Vorhersage auch nicht verwendbar
- Grösse ε 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

...

- *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?

1.

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 $X_1, X_2, \ldots, X_p$:

$$f(X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

  - Nach Wahl des Modells: Verfahren, das die Daten in das Modell *passt*
  - Lineares Modell: Parameter $\beta_0, \beta_1, \ldots \beta_p$ schätzen
  - Parameter so bestimmen, dass

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

  - Häufigste Methode zur Bestimmung von $\beta_0, \beta_1, \ldots \beta_p$: *Methode der kleinsten Quadrate*

Beispiele

- Beispiel `Werbung`: Lineares Modell:

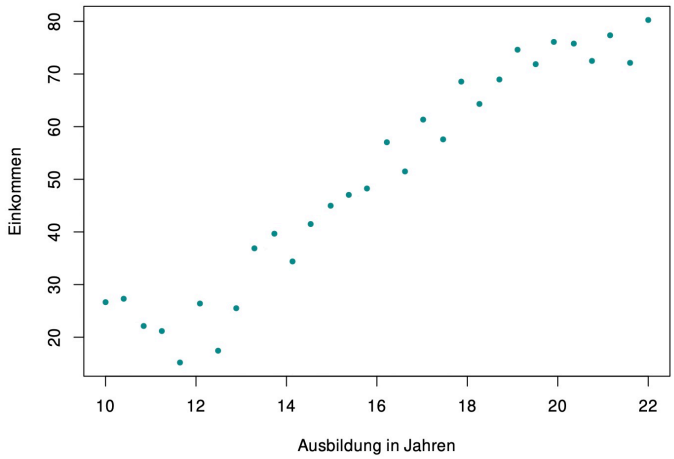
$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung}$$

- Beispiel `Einkommen`: Lineares Modell:

$$\texttt{Einkommen} \approx \beta_0 + \beta_1 \cdot \texttt{Ausbildung}$$

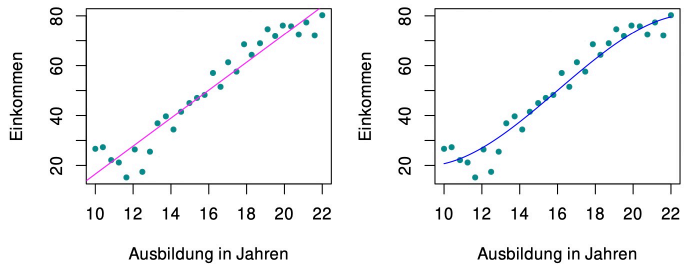
Beispiel

- Datensatz `Einkommen`:



2.

- 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):

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

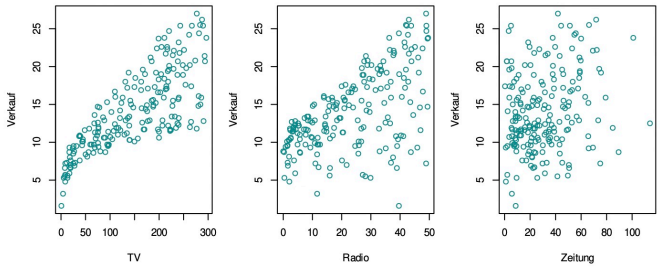
- Viele weitere Modelle denkbar
- Aber welches ist nun das „richtige"?
- Dies lässt sich in dieser Absolutheit nicht entscheiden
- Funktion f i. A. unbekannt: Liegt an uns „bestes" Modell zu wählen
- Statistik: Bei Entscheidungsfindung behilflich
- Welches Modell ist in unserem Beispiel das „bessere"?
- Kubisches Modell scheint besser zu passen, aber auch komplizierter
- Lineares Modell einfacher (etwas weniger genau) hat Vorteil: Die Parameter β_0 und β_1 lassen sich geometrisch interpretieren:
  - β_0 ist der y-Achsenabschnitt
  - β_1 die Steigung der Geraden

Bemerkungen

- 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

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV}$$

- Grössen β_0 und β_1 sind unbekannte Konstanten, die den y-Achsenabschnitt und die Steigung des linearen Modells darstellen
- β_0 und β_1 die *Koeffizienten* oder *Parameter* des Modells
- Koeffizienten werden aus den gegebenen Daten geschätzt
- Schätzungen $\widehat{\beta}_0$ und $\widehat{\beta}_1$ für die Modellkoeffizienten
- Sind diese Koeffizienten bekannt, so können zukünftige Verkäufe auf der Basis eines bestimmten Werbebudgets für TV vorhersagen
- Berechnung mittels:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

wobei $\widehat{y}$ die Vorhersage von Y auf Basis des Inputs $X = x$ bezeichnet.

Schätzung der Parameter

- Praxis: β_0 und β_1 unbekannt
- Bevor lineare Modell benützen $\rightarrow$ Koeffizienten schätzen
- Gehen von n Beobachtungspaaren aus:

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

- 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: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so zu bestimmen, dass die Gerade $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ möglichst gut zu den Daten passt
- Das heisst, dass

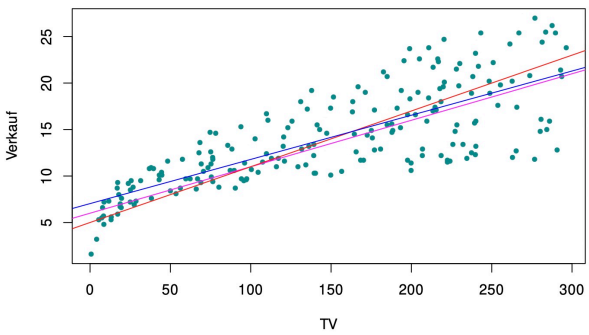
$$y_i \approx \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

für alle $i = 1, \ldots, n$

- 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:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

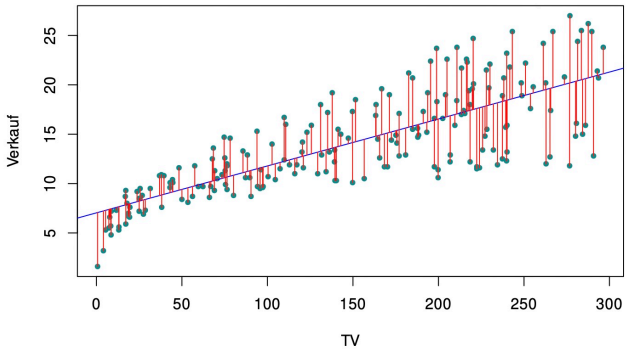
- i-tes *Residuum*:

$$r_i = y_i - \widehat{y}_i$$

- 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

Weiter auf nächster Seite...

Für die, die es interessiert

- Mit Differentialrechnung: Für $\widehat{\beta}_0$ und $\widehat{\beta}_1$ gilt:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

mit

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i \quad \text{und} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$$

- Mit Hilfe der Methode der kleinsten Quadrate geschätzte Koeffizienten für die einfache lineare Regression

Beispiel

- Beispiel `Werbung`: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ und die Regressionsgerade bestimmen:

```
lm(Verkauf ~ TV)
##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Coefficients:
## (Intercept)           TV
##     7.03259      0.04754
```

- Wert unter `Intercept`: $\widehat{\beta}_0$ $\rightarrow$ y-Achsenabschnitt
- Wert unter `TV`: $\widehat{\beta}_1$ $\rightarrow$ Steigung der Geraden
- Lineares Modell:

$$Y \approx 7.03 + 0.0475X$$

...

- Summe der *Quadrate* der Residuen (RSS genannt)
- Es gilt dann

$$\text{RSS} = r_1^2 + r_2^2 + \ldots + r_n^2$$

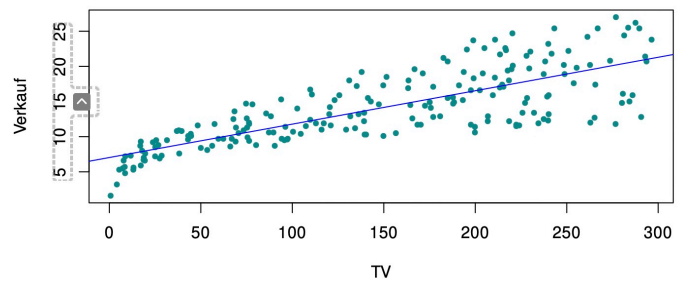
- Oder äquivalent:

$$\text{RSS} = (y_1 - \widehat{\beta}_0 - \widehat{\beta}_1 x_1)^2 + (y_2 - \widehat{\beta}_0 - \widehat{\beta}_1 x_2)^2 + \ldots + (y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n)^2$$

- Methode der kleinsten Quadrate: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so gewählt, dass RSS *minimal* wird

- 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 β_0:

$$[6.130, 7.935]$$

- Für β_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

Beispiel

- p-Wert von β_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
```

Hypothesentest: Statistische Signifikanz von β_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:

$$H_0: \quad \beta_1 = 0$$

- Gegen:

$$H_A: \quad \beta_1 \neq 0$$

- $\beta_1 = 0$, dann:

$$Y = \beta_0 + \varepsilon$$

- Y hängt *nicht* von X ab
- Nullhypothese testen: $\widehat{\beta}_1$ genügend weit von 0 weg, damit β_1 nicht 0
- Mit t-Statistik

Weiter auf nächster Seite...

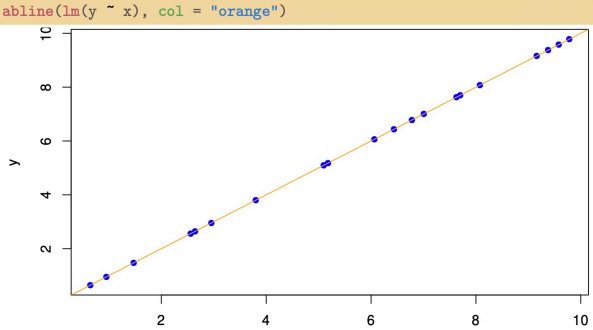
...

- Eintrag `Coefficients` unter `Pr(>|t|)`: p-Wert $2 \cdot 10^{-16}$
- Bei weitem kleiner als 0.05
- Nullhypothesen $\beta_1 = 0$ verwerfen: $\beta_1 \neq 0$
- Klarer Hinweise für Zusammenhang zwischen `TV` und `Verkauf`

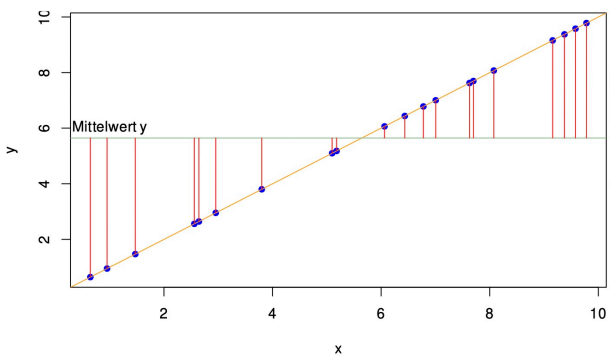
Punkte folgen linearem Modell

- 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 $\overline{y}$

- Output:
  - Korrelation:

```
cor(x, y)
## [1] 1
```

  - R^2:

```
summary(lm(y ~ x))$r.squared
## [1] 1
```

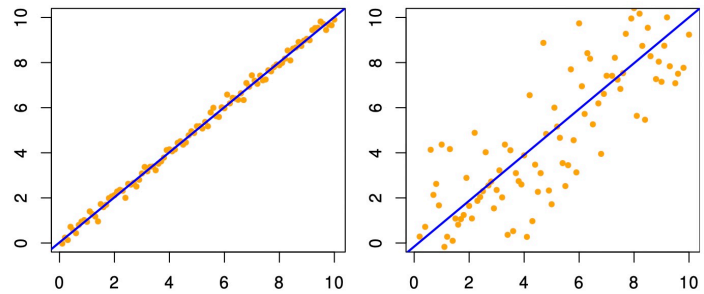
  - Varianz:

```
var(y)
## [1] 8.998626
```

  - 100% der Varianz von 9 wird durch das Modell erklärt

Abschätzung der Genauigkeit des Modells: R^2

- Nullhypothese verworfen: *In welchem Ausmass passt das Modell zu den Daten?*
- Abbildung:



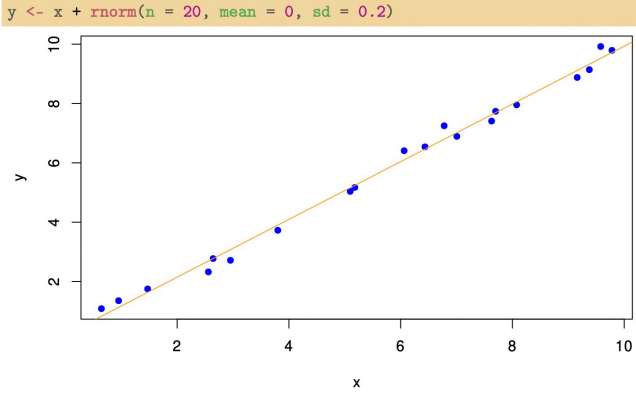
  - Links: Steigende Gerade passt sehr gut zu Punkten
  - Rechts: Steigende Gerade passt *nicht* gut zu Punkten

- 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 Variabiliät 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 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:

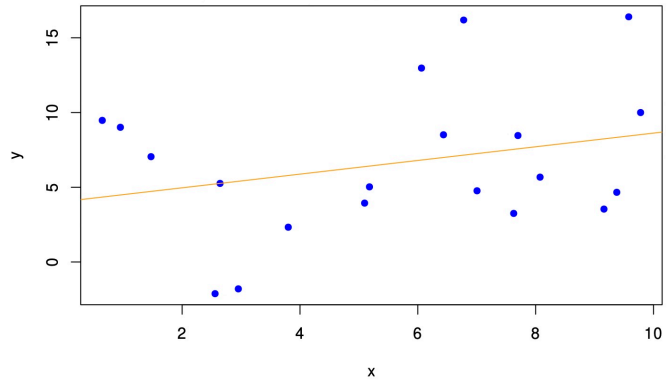
```
var(y)
## [1] 8.573793
```

  - 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:

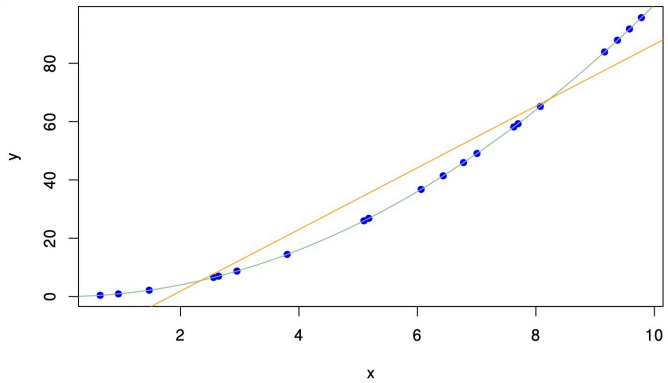
```
var(y)
## [1] 24.55976
```

  - 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
```

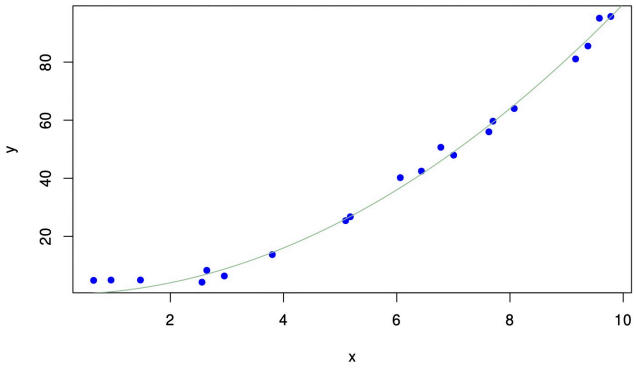
  - Varianz:

```
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
```

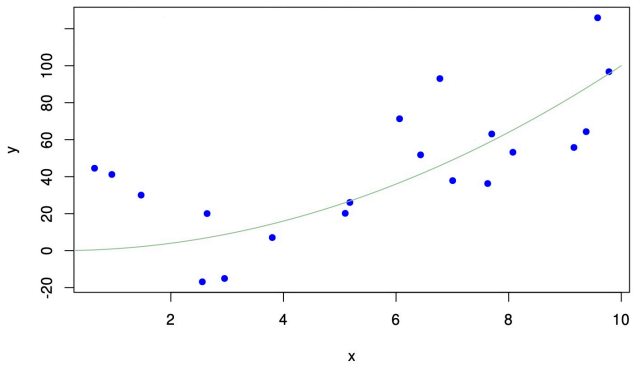
  - Varianz:

```
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

Beispiel

- Im Beispiel der TV-Werbung war der R^2-Wert 0.61

```
summary(lm(Verkauf ~ TV))$r.squared
## [1] 0.6118751
```

- Somit werden knapp zwei Drittel der Variabilität in `Verkauf` durch `TV` mit linearer Regression erklärt.

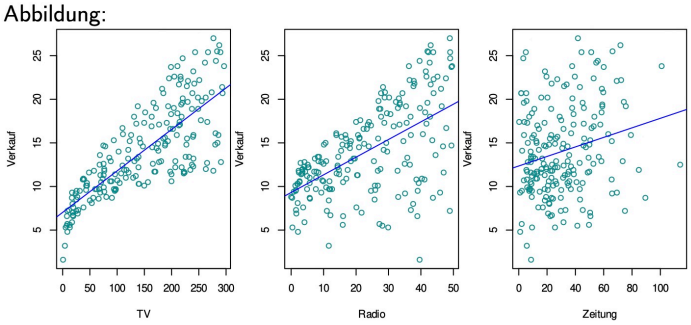
SW12 – Multiple lineare Regression

1.

- 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

2.

- 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 mitberücksichtigten
- Jeder erklärenden Variablen wird ein *eigener* Steigungskoeffizient in *einer* Gleichung zugeordnet
- Allgemein: p verschiedene erklärende Variablen
- *Multiples lineares Regressionsmodell*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- X_j: j-ter Input
- β_j: Zusammenhang zwischen *dieser* erklärenden Variablen und der Zielgrösse Y
- β_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`:

$$\texttt{Verkauf} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung} + \varepsilon$$

- Also

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung}$$

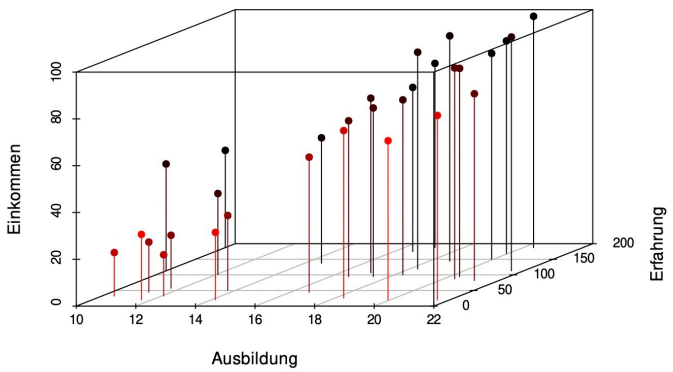
- Multiples lineares Modell verallgemeinert einfaches lineares Modell
- Berechnungen und Interpretationen für multiples Modell ähnlich, wenn auch meist komplizierter als beim linearen Modell
- Graphische Methoden: Entfallen für multiples lineare System praktisch vollends
- Datenpunkte für Beispiel vorher: Nicht darstellbar, da schon für erklärende Variablen drei Achsen gebraucht werden

Beispiel: `Einkommen`

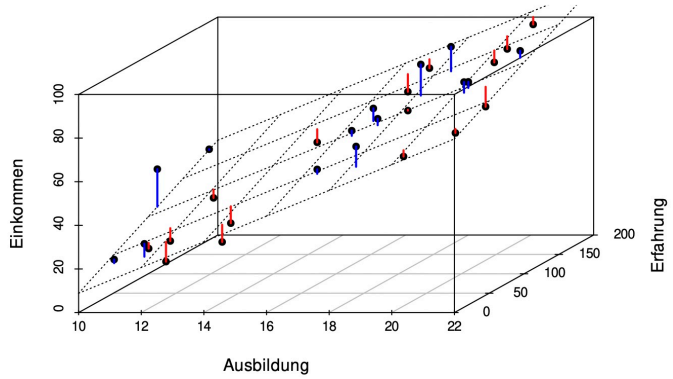
- Graphische Darstellung für zwei erklärende Variablen möglich
- Datensatz `Einkommen`
- Bis jetzt: `Ausbildung` einzige erklärende Variable
- Einkommen auch von `Erfahrung` (Anzahl Berufsmonate) abhängig
- Multiples lineares Modell:

$$\texttt{Einkommen} = \beta_0 + \beta_1 \cdot \texttt{Ausbildung} + \beta_2 \cdot \texttt{Erfahrung} + \varepsilon$$

- 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*

Interpretation der Koeffizienten

- $\widehat{\beta}_0 = -50.086$:
  - Wenn Person keine Ausbildung und keine Erfahrung hat, so „erhält" man CHF $-50\,086$
  - Interpretation macht praktisch natürlich keinen Sinn
- $\widehat{\beta}_1 = 5.896$:
  - Bei konstanter Erfahrung verdient man pro zusätzliches Ausbildungsjahr Ausbildung CHF 5896 mehr
- $\widehat{\beta}_2 = 0.173$:
  - Bei konstanter Ausbildung verdient man pro zusätzlichen Monat Arbeitserfahrung CHF 173 mehr

Allgemein: Schätzung der Regressionskoeffizienten

- Wie einfache linearer Regression: Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_p$ i. A. unbekannt
- Müssen sie aus Daten schätzen:

$$\widehat{\beta}_0, \quad \widehat{\beta}_1, \quad \ldots, \quad \widehat{\beta}_p$$

- Aufgrund der Schätzungen kann man Vorhersagen machen:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \ldots + \widehat{\beta}_p x_p$$

- Parameter wieder mit der Methode der kleinsten Quadrate schätzen

Beispiel auf der nächsten Seite

- Schätzung von β_0, β_1 und β_2 mit `R`:

$$\widehat{\beta}_0 = -50.086; \qquad \widehat{\beta}_1 = 5.896; \qquad \widehat{\beta}_2 = 0.173$$

```
coef(lm(Einkommen ~ Ausbildung + Erfahrung))
## (Intercept)  Ausbildung   Erfahrung
## -50.0856388   5.8955560   0.1728555
```

- Multiples lineares Modell:

$$\texttt{Einkommen} \approx -50.086 + 5.896 \cdot \texttt{Ausbildung} + 0.173 \cdot \texttt{Erfahrung}$$

Beispiel

1.

- 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:

$$\texttt{Verkauf} \approx 2.94 + 0.046 \cdot \texttt{TV} + 0.189 \cdot \texttt{Radio} - 0.001 \cdot \texttt{Zeitung}$$

- 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 $\widehat{\beta}_3$ 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)

2.

- 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:

	TV	Radio	Zeitung	Vekauf
TV	1.0000	0.0548	0.0567	0.7822
Radio		1.0000	0.3541	0.5762
Zeitung			1.0000	0.2283
Verkauf				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 $\rightarrow$ mehr Glaceverkäufe $\rightarrow$ 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 $X_1, \ldots, X_p$ nützlich, um die Zielgrösse vorherzusagen?*
- *Spielen alle erklärenden Variablen $X_1, \ldots, X_p$ 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 β_0 Null sind (keine Variable hat Einfluss):
$$\beta_1 = \beta_2 = \ldots = \beta_p = 0$$
- Nullhypothese
$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
- Alternativhypothese
$$H_A: \quad \text{mindestens ein } \beta_i \text{ ist ungleich } 0$$
- 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 β_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:

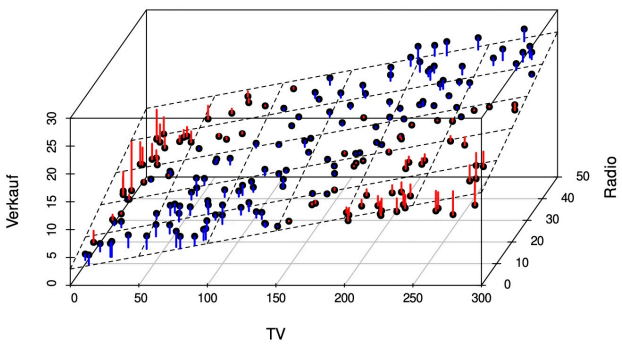
```
##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62976 -0.66857  0.00927  0.64462  2.81840
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.029669   0.047272  -0.628   0.5305
## X1          -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 Variabeln sind wichtig?
- Prozedere: *Variablenselektion* (nächstes Mal)

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

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 $\rightarrow$ 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`

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 Interaktionterm sehr klein ist, aber die p-Werte der Haupteffekte (hier `TV` und `Radio`) sind es nicht

SW13 – Variablenselektion (NUR Qualitative Variablen)

Qualitative erklärende Variablen

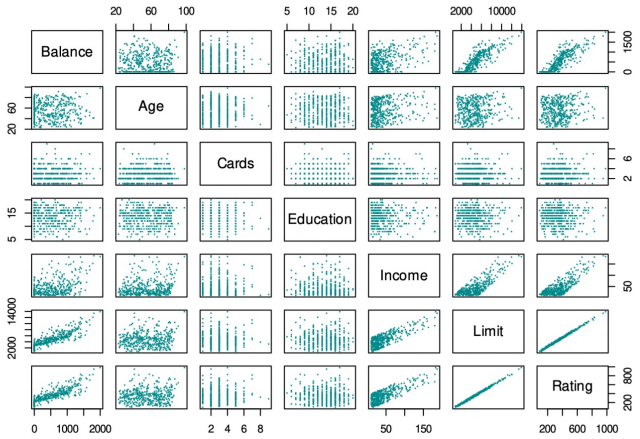
- Bisher angenommen: Alle Variablen *quantitativ* in linearem Regressionssystem
- Aber: Oft sind einige erklärenden Variablen *qualitativ*

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 auf nächster Seite

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:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- β_0: durchschn. Kreditkartenrechnungen der Männern
- $\beta_0 + \beta_1$: durchschn. Kreditkartenrechnungen der Frauen
- β_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 β_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 β_0 und β_1 \$ 529.53, resp. \$ -19.73
- Entspricht wiederum Rechnungen von:
  - Frauen: \$ 529.53
  - Männer: \$ 529.73 − \$ 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:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- β_0: Durchschn. Rechnungen ohne Berücksichtigung des Geschlechts
- β_1: Wert, mit welchem Frauen über dem Durchschnitt liegen und mit welchem Männer unter dem Durchschnitt liegen

- β_0 durch \$ 519.665 geschätzt: Durchschn. Rechnungen von \$ 509.80 für Männer und von \$ 529.53 für Frauen
- Schätzung \$ 9.865 für β_1: Hälfte vom Unterschied \$ 19.73 zwischen Männern und Frauen
- Wichtig: Vorhersagen für d 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:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch} \end{cases}$$

- β_0: Durchschn. Kreditkartenrechnungen von Afroamerikanern

- β_1: Differenz der durchschn. Rechnungen von Afroamerikanern und Asiaten
- β_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 $\rightarrow$ 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

$$H_0: \quad \beta_1 = \beta_2 = 0$$

- p-Wert dieser Statistik hängt *nicht* von der Kodierung ab
- p-Wert 0.96 $\rightarrow$ 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:

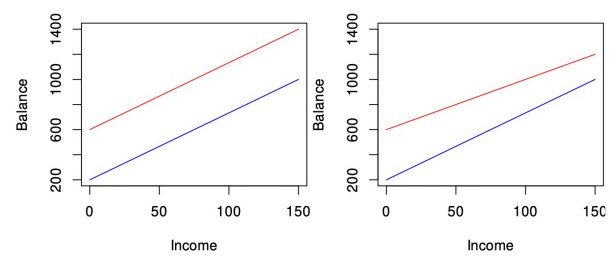
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 & \text{falls } i\text{-te Person Student} \\ 0 & \text{falls } i\text{-te Person kein Student} \end{cases}$$

$$= \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{falls } i\text{-te Person Student} \\ \beta_0 & \text{falls } i\text{-te Person kein Student} \end{cases}$$

- 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
```

- $\widehat{\beta}_0$:
  Ohne Einkommen und als Nichtstudent zahlt man $211 monatliche Kreditkartenrechnung
- $\widehat{\beta}_1$:
  Pro $1000 Einkommen mehr, zahlt man $6 mehr Kreditkartenrechnung (unabhängig vom Studentenstatus)
- $\widehat{\beta}_2$:
  Studierende zahlen $383 mehr Kreditkartenrechnung als Nichtstudierende (unabhängig vom Einkommen)
- Modell beschreibt zwei parallele Geraden: eine für Studierende und eine für Nichtstudierende
  - Steigung β_1 ist bei beiden gleich
  - y-Achsenabschnitte sind verschieden ($\beta_0 + \beta_2$ und β_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"

Weiter auf nächster Seite…

...

- Modell:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \texttt{income}_i & \text{falls studierend} \\ 0 & \text{falls nicht studierend} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \texttt{income}_i & \text{falls studierend} \\ \beta_0 + \beta_1 \cdot \texttt{income}_i & \text{falls nicht studierend} \end{cases}$$

- Zwei unterschiedliche Regressionsgeraden für Studierende und Nichtstudierende (Abbildung oben rechts):
  - Verschiedene Steigungen $\beta_1 + \beta_3$ und β_1
  - Unterschiedliche y-Achsenabschnitte $\beta_0 + \beta_2$ und β_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 Studenierende (rot) und Nichtstudierende (blau)

- Steigung für Studierende ist grösser als für Nichtstudierende

- Deutet an: Zunahme im Einkommen eines Studierenden eine grössere 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