

ASTAT Zusammenfassung

1.1 Daten

Eindimensionaler Datensatz: Liste / Messreihen

Zweidimensionale Datensätze: Tabellen (am häufigsten)

Quantitative Daten: gemessene Zahlen (Größe und Gewicht)

Qualitative Daten: nehmen nur bestimmte Anzahl Werte an (müssen keine Zahlen sein)
z.B. Geschlecht oder Nationalität

1.2 Deskriptive Statistik

Deskriptive Statistik: Darstellung von Datensätzen

Datensätze:

- durch gewisse Zahlen charakterisieren (z.B. Mittelwert)
- und graphisch darstellen

1.2.1 Ziele der Deskriptiven Statistik

- Daten zusammenfassen durch nummerische Kennwerte
- Graphische Darstellung der Daten
- Messreihen zusammenzufassen
- Interpretation und darauffolgende statistische Analyse dieser Daten vereinfachen

Warnung!!!

Wann immer ein Datensatz „reduziert“ wird (durch Kennzahlen oder Graphiken), geht *Information verloren!*

1.2.2 Kennzahlen

Legeparameter, Streuungsparameter

- **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

1.2.3 Arithmetisches Mittel (Durchschnitt, Mittelwert) (z.T. mu genannt)

Arithmetisches Mittel

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

Arithmetisches Mittel sagt nicht alles über Messreihe aus.

Der **Mittelwert** wird berechnet, indem alle Werte summiert werden und danach die Summe durch die Anzahl der Werte dividiert wird. Durchschnitt der Verteilung. **Nicht robust.**

1.2.4 Mittlere absolute Abweichung

Streumass, mit Beträgen rechnen. Wie Durchschnitt ausrechnen und Vorzeichen weglassen.

- Nächste Idee: Unterschiede durch die *Absolutwerte* ersetzen

Mit Beträgen rechnen, Vorzeichen weglassen

- 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 grösser gleich 0), desto mehr unterscheiden sich die Daten bei gleichem Mittelwert voneinander

- Dieser Wert für die Streuung: *Mittlere absolute Abweichung*

- Aber: Theoretische Nachteile

Nachteil: Beträge lassen sich in der Analysis nicht ableiten.

Durch das Quadrieren fallen die Vorzeichen ebenfalls weg. Somit verwendet man die empirische Varianz und empirische Standardabweichung.

1.2.5 Empirische Varianz

Mass für Variabilität/Streuung der Messwerte. Standardabweichung σ^2 . Grosse Varianz = grosse Streuung.

Empirische Varianz $\text{var}(x)$ und Standardabweichung s_x

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

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

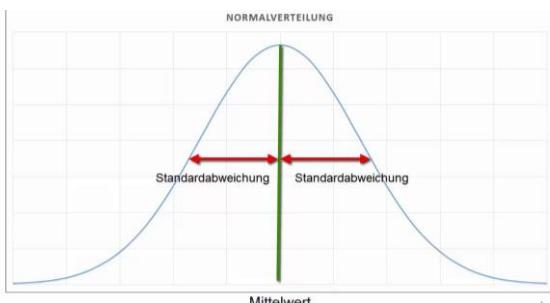
$$\text{Var}(X) = \sigma^2$$

1.2.6 Standardabweichung



Standardabweichung ist die Wurzel der Varianz. Streuung der Werte um den Durchschnitt. Standardfehler/abweichung vom Mittelwert.

- ▶ Empirische Standardabweichung: s_x aus konkreten Daten
berechnet: Aus Messwerte x_1, \dots, x_n wird nach Formel oben s_x berechnet
- ▶ Standardabweichung σ_X : Theoretischer Wert, der sich aus Modell der W'keitsverteilung ergibt



1.2.7 Median (Zentralwert, mittlerer Wert)

Ein weiteres Lagemaß für die „Mitte“: Median

Median wird durch kleine Fehler nicht beeinflusst, ist **robust**. Das arithmetische Mittel wird beeinflusst, ist nicht robust.

$$\tilde{x} = \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n}{2} + 1} \right)$$

Sehr vereinfacht: Wert, bei dem die Hälften der Messwerte unter oder gleich diesem Wert sind. Die andere Hälfte ist über diesem Wert. Der **Median** kann berechnet werden, indem alle Zahlen in aufsteigender Reihenfolge aufgelistet werden und dann die Zahl in der Mitte dieser Verteilung ausgewählt wird. Macht keine Aussage über Streuung.

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

1.2.8 Quartile

- 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

Beispiel: Waage $0.25 \cdot 13 = \frac{3}{4} \cdot 13 = 3.25$

- 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 2.0 , bei Median mit $n=13$
 $0.75 \cdot 13 = 6.75 \sim 7$
- 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, 80.03, 80.03, 80.03, 80.04, 80.04, 80.05

- Knapp drei Viertel der Messwerte sind kleiner oder gleich 80.04

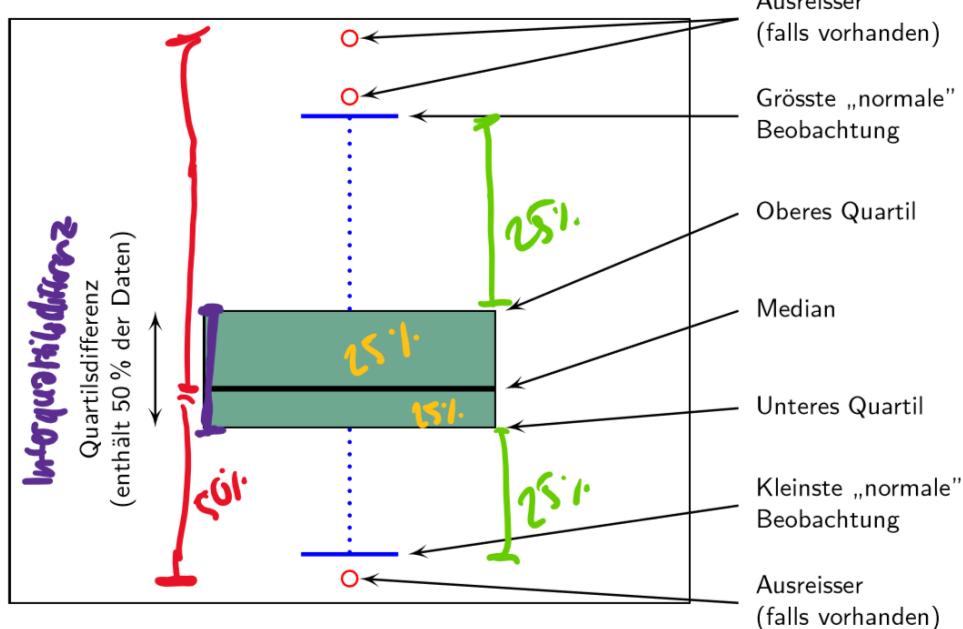
1.2.9 Quartaldifferenz

Quartilsdifferenz ist ein Streuungsmaß für die Daten oberes Quartil – unteres Quartil.

Unteres und oberes Quartil werden abgeschnitten, 50% bleiben übrig (Median). Ist auch robust gegenüber Ausreisern, kann sich jedoch ändern, wenn man den Wert, wo die Quartilsdifferenz beginnt, ändert.

1.2.10 Boxplot

Boxplot: Schematischer Aufbau



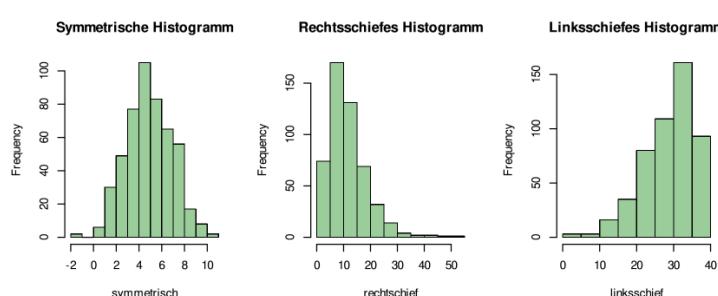
- Rechteck, dessen Höhe vom empirischen 25 %- und vom 75 %-Quantil begrenzt wird: Box
- Horizontaler Strich für den Median in Box (schwarz)
- Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten „normalen“ Wert führen (blau eingezzeichnet)
 - ▶ Definition: „Normaler“ Wert höchstens 1.5 mal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

Damit ein Boxplot normalverteilt ist, muss der Median exakt in der Mitte der Quartalsdifferenz sein.

2 Zweidimensionale Deskriptive Statistik

2.1 Schiefe von Histogrammen

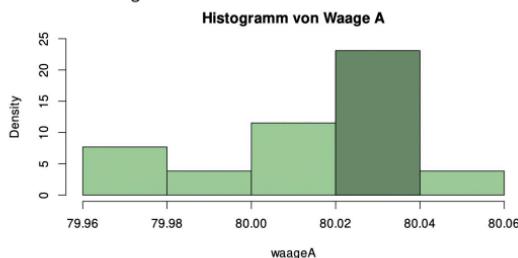
Aufpassen, ist das Gegenteil, dort wo weniger Daten sind



2.2 Normiertes Histogramm

Hier ist die Dichte (Destiny) gegeben nicht mehr die Häufigkeit (Frequency). Höhere Aussagekraft.

- Normiertes Histogramm:



- Dichte der Klasse von 80.02 – 80.04 ist etwa 23

- Fläche dieses Balkens (dunkelgrüne Fläche in Abbildung):

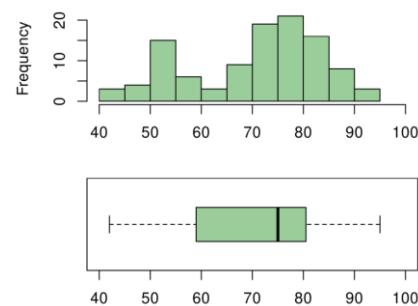
$$(80.04 - 80.02) \cdot 23 = 0.46$$

- 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

2.3 Multimodale Verteilung

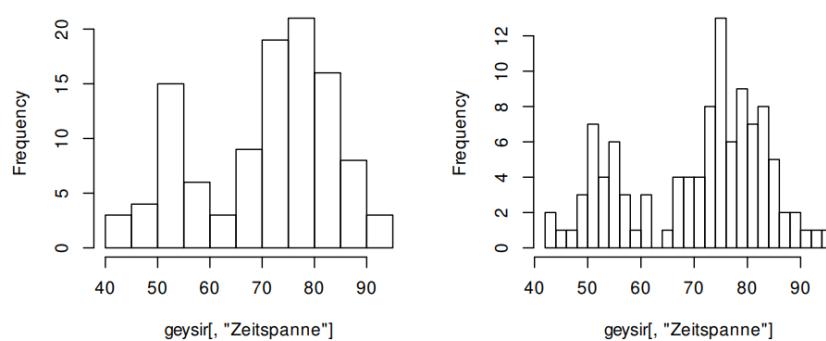
Wenn es eine Multimodale Verteilung gibt, dann ist der Boxplot nicht mehr genügend aussagekräftig. Dann soll man ein Histogramm machen. Dann gehen Daten verloren.



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

2.4 Verteilung über zwei Gipfel

So eine Verteilung mit zwei Gipfeln heisst auch bimodal.

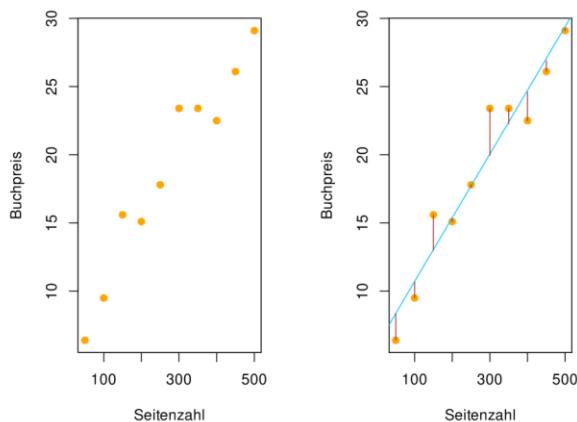


2.5 Streudiagramme (engl. Scatterplot)

Zwei Messungen als Koordinaten von Punkten in einem Koordinatensystem interpretiert und dargestellt. Bei Streudiagrammen aufpassen: Abhängigkeit nicht mit Kausalität (Ursache & Wirkung) verwechseln. Gesetzmässigkeit vorhanden, heisst dies noch lange nicht, dass diese Gesetzmässigkeit auch kausal erklärt werden kann.

2.6 Streudiagramm und Regressionsgerade

Regressionsgerade: Gerade wird erstellt, indem die Datenpunkte auf einer Ebene betrachtet und eine Linie gezeichnet wird, die so nah wie möglich an allen Punkten vorbeiführt.



2.7 Residuum

Abstände von Messpunkten zu Geraden (siehe oben).

$$y = a + bx$$

- Abstände von Messpunkten zu Geraden → 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:

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

- Minimierung von $\sum_i r_i$ hat aber eine **gravierende Schwäche**: Falls Hälften der Punkte weit über der Geraden, die andere Hälften weit unter der Geraden liegen: Summe der Abstände etwa null

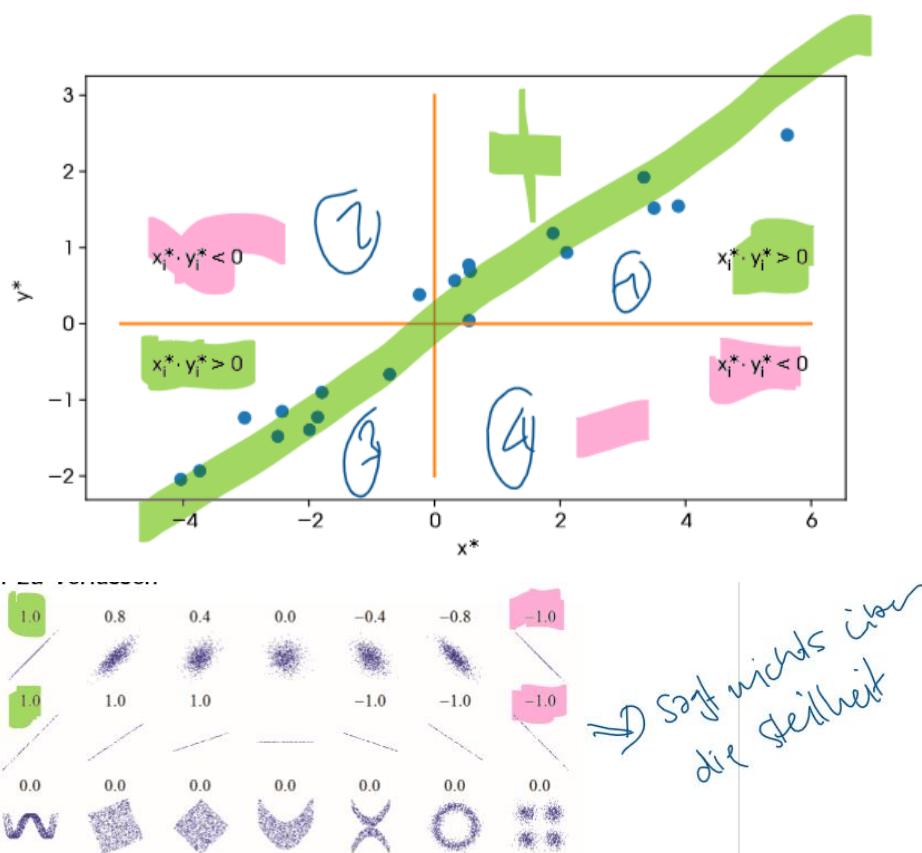
3 Korrelation und Wahrscheinlichkeitsmodelle

3.1 Empirische Korrelation

Misst Stärke und Richtung der **linearen** Abhängigkeit zwischen Daten x und y . Ist immer zwischen **-1 und +1** (Korrelationskoeffizient immer zwischen -1 und 1).

- $r = +1$: 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 $r = 0$

Erkennt nur lineare Zusammenhänge, quadratische nicht! Sagt auch nichts über die Steilheit aus. Wenn es eine quadratische Abhängigkeit ist, hat dies aber trotzdem Einfluss auf den Korrelationskoeffizienten, der nur lineare Abhängigkeit erkennt.



3.2 Wahrscheinlichkeitsmodell

- 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:

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

3.2.1 Mengen aus Bekannten

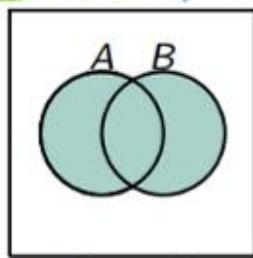
Bei **unabhängiger** Wahrscheinlichkeit (*Klammern immer setzen*):

- Operationen der Mengenlehre für Ereignisse:

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

- Graphisch

Nachweis ein Ereignis Ω

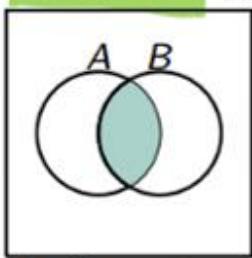


$$P = P(A \cup B)$$

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

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

*Schnittmenge
beide Ereignisse* Ω



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

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

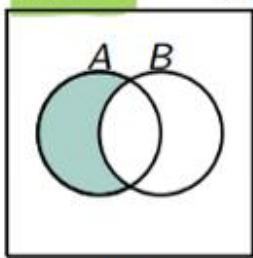
$$P(\Omega) = 1$$



$$\bar{A}$$

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

Differenz Ω

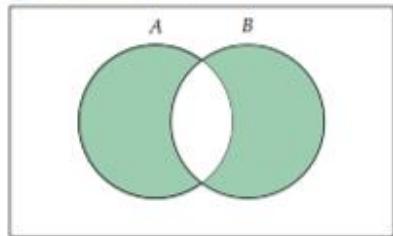


$$A \setminus B$$

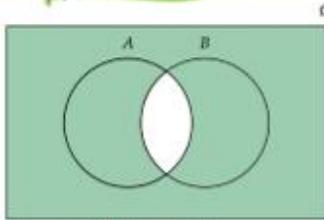
$$P(A) - P(A \cap B)$$

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

genau eines Ω



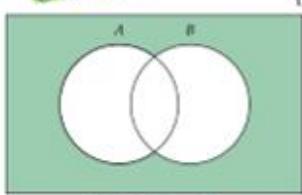
genau eins Ω



$$P = 1 - P(A \cap B)$$

$$P = 1 - (P(A) + P(B))$$

keines Ω



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

$$P = 1 - P(A \cup B)$$

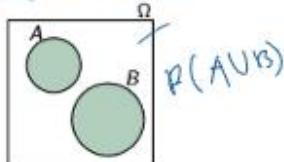
$$P = 1 - (P(A) + P(B) - (P(A) \cdot P(B)))$$

$$P(\text{genau ein Ereignis}) = P(A \cup B) - P(A \cap B)$$

$$= P(A) + P(B) - 2P(A) \cdot P(B)$$

$$= A + B - 2 \cdot A \cdot B$$

Disjunkte Ereignisse:



$$P(A) + P(B)$$

Disjunkte Ereignisse = haben Einfluss aufeinander, Schnittmenge = 0

3.2.2 Axiome der Wahrscheinlichkeit (Kolmogorov)

Allgemeine Aussagen über die Wahrscheinlichkeit (unabhängig oder abhängig).

Kolmogorov Axiome der Wahrscheinlichkeitsrechnung

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

- A1 $P(A) \geq 0$
- A2 $P(\Omega) = 1$
- A3 $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \{\}$

- Bezeichnung $P(A)$: 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*

Folgendes gilt (Algorithmus muss dies erfüllen):

- **Axiom 1: Wahrscheinlichkeit muss immer grösser als 0 sein!**
- **Axiom 2: Wahrscheinlichkeit muss immer 1 geben (100%)**
- **Axiom 3: Wahrscheinlichkeiten (von Fall A und Fall B) von verschiedenen Fällen können zusammenaddiert werden, wenn diese sich nicht überlappen (disjunkt)!**

Wahrscheinlichkeit wird nie in Prozent angegeben.

3.2.3 Rechenregel aus Axiomen

Prüfungsrelevant, die markierten Regeln:

Rechenregeln

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

| | |
|---|--|
| $P(\bar{A}) = 1 - P(A)$ | für jedes A <i>DGegenwahrscheinlichkeit</i> |
| $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ | für beliebige A und B <i>Allgemeine Additionsgesetze</i> |
| $P(A_1 \cup \dots \cup A_n) \leq P(A_1) + \dots + P(A_n)$ | für beliebige A_1, \dots, A_n |
| $P(B) \leq P(A)$ | für beliebige A und B mit $B \subseteq A$ |
| $P(A \setminus B) = P(A) - P(B)$ | für beliebige A und B mit $B \subseteq A$ |

3.3 Diskrete Wahrscheinlichkeitsmodelle

Grundraum ist endlich/unendlich und diskret.

- Begriff „diskret“: Endliche Menge, wie:

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

- Unendliche, aber trotzdem diskrete Menge, wie

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

- Menge $\Omega = \mathbb{R}$ (Menge aller Dezimalbrüche, Zahlengerade) ist nicht diskret



- Berechnung von W'keiten für diskrete Modelle

Im diskreten Fall ist die W'keit eines Ereignisses

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

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

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

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

3.4 Gegenwahrscheinlichkeit

- Einfacher mit Gegenw'keit: $P(\bar{B})$

- 1. Rechenregel: Komplement \bar{B} von B :

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

- Dann gilt:

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

3.5 Modell von Laplace

- Annahme: Jedes Elementarereignis hat die gleiche W'keit
- Ereignis $E = \{\omega_1, \omega_2, \dots, \omega_g\}$;
- 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

3.6 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

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

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

- 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 $P(A \cap B)$

Abhängigkeit: wenn das Eintreten eines einen Ereignisses die Wahrscheinlichkeit für das Eintreten des anderen Ereignisses beeinflusst.

4 Zufallsvariable & Wahrscheinlichkeitsverteilung

4.1 Zufallsvariable

Ist eine Zahl (keine Beschreibung wie Geschlecht oder Nationalität).

Gezinkte Münze: manipulierte Münze, die so verändert wurde, dass sie beim Münzwurf oder in anderen Glücksspielen dazu neigt, immer auf eine bestimmte Seite zu fallen.

- 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

Zufallsvariable

Eine Zufallsvariable X ist eine Funktion:

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

- 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:

$$X = x$$

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

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

$$P(\overline{X} = 4) = P(\text{König}) + P(\text{König}) + P(\text{König}) + P(\text{König})$$

$$\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{4}{36} = \frac{1}{9}$$

→ 9 Könige

4.2 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 → W'keitsverteilung dieser Zufallsvariablen

Tabelle ist eine Wahrscheinlichkeitstabelle, wenn die Summe 1 gibt!

- W'keitsverteilung von X als Tabelle:
- | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $P(X = x)$ | $4/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ | $1/9$ |

Wahrscheinlichkeitsverteilung

„Liste“ von $P(X = x)$ für alle möglichen Werte x_1, x_2, \dots, x_n heißt diskrete W'keitsverteilung der diskreten Zufallsvariablen X

Es gilt immer:

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

Mit Summenzeichen:

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

Alle W'keiten einer W'keitsverteilung ergeben 1

!

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

- Wie gross ist die W'keit, die Augensumme 6 oder 8 zu würfeln?

► Gesucht: $P(X = 6) + P(X = 8)$:

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

4.3 Kennzahlen der Verteilung (Erwartungswert und Standardabweichung)

μ

- Beliebige (diskrete) Verteilung: Vereinfachend durch 2 Kennzahlen zusammengefasst:

► **Erwartungswert**

$$E[X] = \mu$$

► Erwartungswert: Mittlere Lage der Verteilung

► Standardabweichung $\sigma(X)$: Streuung der Verteilung

- Erwartungswert: Oft auch mit μ_X bezeichnet:

► Index X wird oft weglassen, falls Zufallsvariable klar

- Definition:

Erwartungswert und Standardabweichung

► **Erwartungswert:**

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

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

► **Varianz und Standardabweichung:**

$$\text{Var}(X) = (x_1 - E(X))^2 \cdot P(X = x_1) + \dots + (x_n - E(X))^2 \cdot P(X = x_n)$$

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

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

Hat zwei Möglichkeiten f. 3-Versuche
z.B. Kopf oder Zahl

(3) = 8 → Waren.

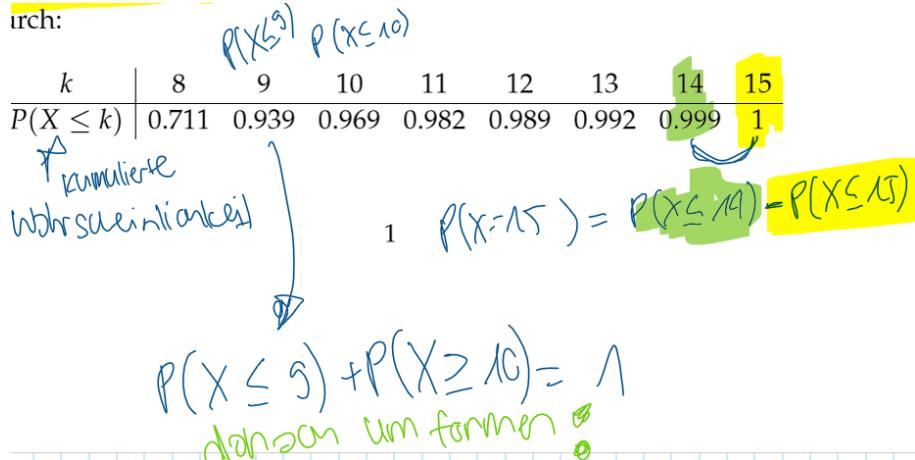
- Arithmetische Mittelwert \bar{x} : Aus **konkreten Daten** berechnet: Aus **Messwerten** x_1, \dots, x_n wird nach der Formel oben \bar{x}_n berechnet
- Erwartungswert $E(X)$: **Theoretischer Wert**, der sich aus dem Modell der W'keitsverteilung ergibt

$P \geq 2$ bis $P \leq 4$; wird aber mit X anders geschrieben!

$$P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) = 0.2 + 0.2 + 0.1 = 0.5$$

Kumulierte Wahrscheinlichkeit berechnen:

durch:



Die Wahrscheinlichkeit $P(X > 11)$ ist 0.989

$$P(X > 11) = 1 - P(X \leq 10)$$

Aufpassen, bei dazwischen, schauen was man abschneidet!

- b) ... genau 5 ...
- c) ... mindestens 1 ...
- d) ... höchstens 6 ...
- e) ... zwischen 3 und 10 ...
- f) ... mehr als 2 ...

$$P(3 \leq X \leq 10) = F_{50;0,1}(10) - F_{50;0,1}(2)$$

↑ Intervall [0; 10] ↑ Intervall [0; 2]

(1) Erwartungswert; (2) Varianz & (3) Standardabweichung in R. Elementarereignis

```
E <- sum(x*p)
E
```

Wir verwenden zur Berechnung R:

```
x <- 2:12
p <- c(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1) / 36
```

Wir berechnen noch die Standardabweichung.

```
E <- sum(x*p)
var.X <- sum((x-E)^2*p)  $\rightarrow$  Varianz?
var.X

## [1] 5.833333

sigma <- sqrt(var.X)  $\rightarrow$  Stdabweichung
sigma

## [1] 2.415229
```

| x_i | Elementarereignis | abs. Häufigkeit | p_i |
|-------|-------------------|-----------------|----------------|
| 2 | 11 | 1 | $\frac{1}{36}$ |
| 3 | 12,21 | 2 | $\frac{2}{36}$ |
| 4 | 13,22,31 | 3 | $\frac{3}{36}$ |
| 5 | 14,23,32,41 | 4 | $\frac{4}{36}$ |
| 6 | 15,24,33,42,51 | 5 | $\frac{5}{36}$ |
| 7 | 16,25,34,43,52,61 | 6 | $\frac{6}{36}$ |
| 8 | 26,35,44,53,62 | 5 | $\frac{5}{36}$ |
| 9 | 36,45,54,63 | 4 | $\frac{4}{36}$ |
| 10 | 46,55,64 | 3 | $\frac{3}{36}$ |
| 11 | 56,65 | 2 | $\frac{2}{36}$ |
| 12 | 66 | 1 | $\frac{1}{36}$ |

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

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

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

5 Bedingte Wahrscheinlichkeit

- Bezeichnungen:

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

- Tabelle:

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

- Tabelle mit W'keiten:

| | M | F | $P(R)$ |
|-----------|------|------|--------|
| R | 0.15 | 0.05 | 0.2 |
| \bar{R} | 0.45 | 0.35 | 0.8 |
| | 0.6 | 0.4 | 1 |

- Betrachten nur ein Teil der Tabelle → Raucher

| | M | F | |
|-----------|------|------|-----|
| R | 0.15 | 0.05 | 0.2 |
| \bar{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:

$$\frac{|R \cap M|}{|R|} = \frac{3}{4} = 0.75 \quad \frac{0.15}{0.2} = 0.75$$

Von allen Rauchern \textcircled{R} , dass dieser ein Mann (M) ist. Von Rechts ausgehen! (unten)

ODER: Männer unter der Bedingung, dass diese Raucher sind

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

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

- Formel wird als Definition der bedingte W'keit verwendet

- Die **bedingte W'keit** ist die W'keit, dass das Ereignis A eintritt, wenn man schon weiss, dass B eingetreten ist

- Bezeichnung:

$$P(A | B)$$

- Längsstrich wird als „unter der Bedingung“ gelesen

- Bedingte W'keit $P(A | B)$ wird definiert durch

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

- Interpretation: $P(A | B)$ ist die W'keit für das Ereignis A , wenn man weiss, dass das Ereignis B schon eingetroffen ist

- Berechnen bedingte W'keit:

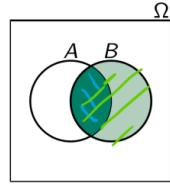
$$P(R | M)$$

- W'keit, dass ein zufällig ausgewählter Mann ein Raucher ist

- Tabelle: Nur Männer werden berücksichtigt werden

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

=
Wahrscheinlichkeit
 $P(M | R)$
Wahrscheinlichkeit hier
Bedingung / Anforderung



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

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

5.1.1 Bayes Theorem

Bayes' Theorem

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

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

5.2 Totale Wahrscheinlichkeit

Gesetz der totalen Wahrscheinlichkeit

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

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

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

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

- Entsprechende Formel gilt für $P(A_2 \cap B)$

- Oben einsetzen: Gesetz der totalen W'keit für $k = 2$:

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

$$1 = P(K) + P(\bar{K})$$

$$P(T|D) = 1 - P(\bar{T}|D) \rightarrow \text{nur vorne, Gegenteil}$$

5.3 Unterschied Gross- und Kleinbuchstaben

- Unterschied: Gross- und Kleinschreibung:
 - $x = 174$ ist eine Zahl / Realisierung
 - $X = 174$ ist eine Menge (Personen mit gerundeter Körpergrösse 174 cm)

5.4 Stetig vs diskret

diskret : $W_X = \{0, 1, \dots, 500\}$
stetig : $W_X = [0, 500]$ - alle Zahlen

6 Normalverteilung

6.1 Wertebereich

- 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:

$$W_X = \mathbb{R}, \mathbb{R}^+ \text{ oder } [0, 1]$$

- Letzter Fall: Zahlen 0 und 1 und alle Zahlen dazwischen

6.2 Intervall

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

- 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

Bei stetiger Zufallsvariable ist die Punktwahrscheinlichkeit immer 0.

Bei einer diskreten Zufallsvariable ist die Punktwahrscheinlichkeit NICHT 0.

6.3 Eigenschaften von 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

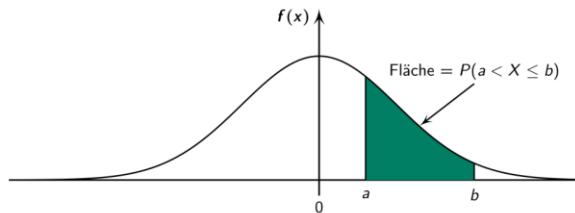
$$P(a < X \leq b)$$

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

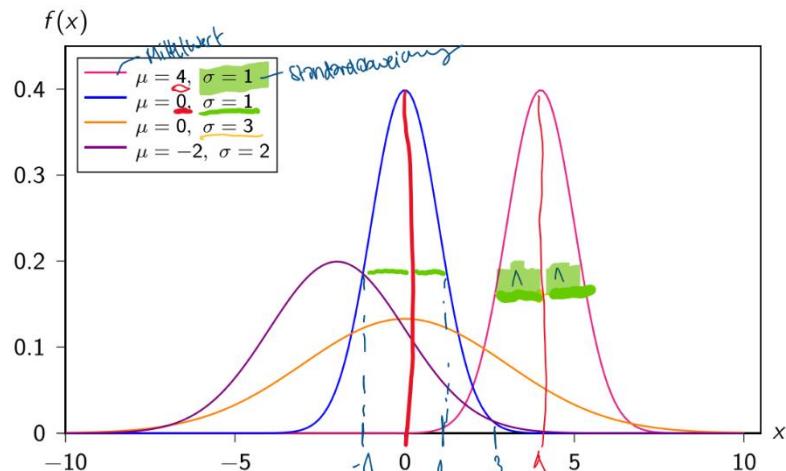
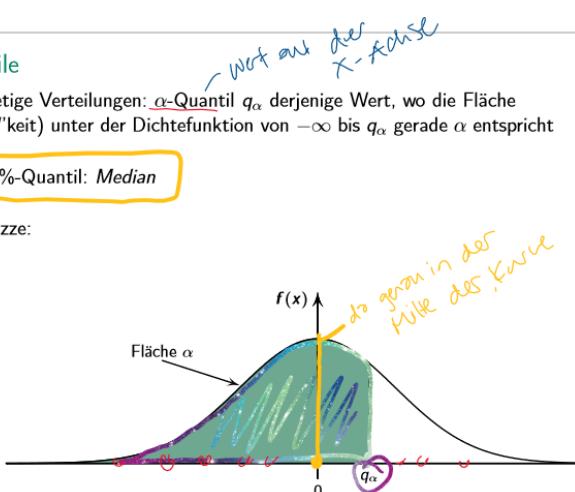
Für stetige W'keitsverteilungen entsprechen W'keiten Flächen unter der Dichtefunktion.

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:



Normalverteilung (Gaussverteilung), Wertebereich, Dichte, Erwartungswert, Varianz

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

- Definition muss man einmal gesehen haben

- Wertebereich

$$W = (-\infty, \infty)$$

- Dichte

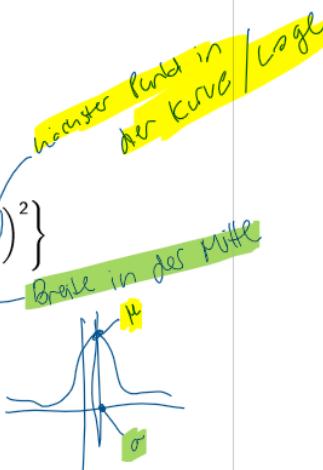
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$$

- Erwartungswert

$$E[X] = \mu$$

- Varianz

$$\text{Var}(X) = \sigma^2$$

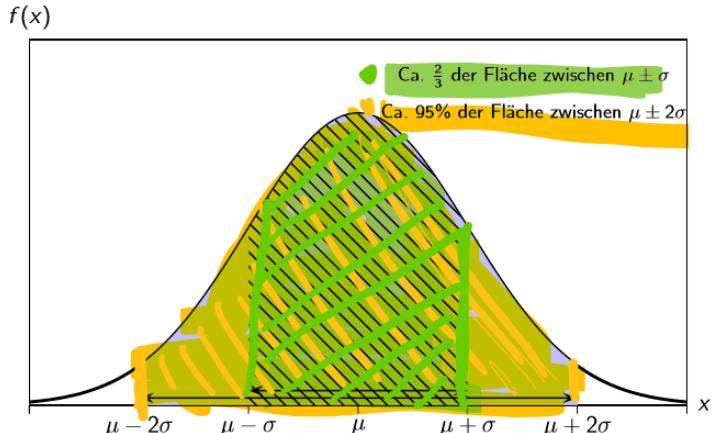


6.4 Eigenschaften der Normalverteilung

Konkrete Aussage über die Streuung als mittlere Abweichung vom Erwartungswert. Die Wahrscheinlichkeit, dass eine Beobachtung eine höchstens Standardabweichung vom Erwartungswert hat, ist immer ca. 2/3.

- 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

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

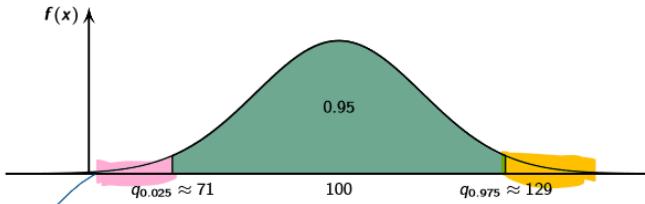
$$\mathbb{E}[X] = \mu$$

- Varianz

$$\text{Var}(X) = \sigma^2$$

- 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.

- W'keiten gegeben → Suchen die zugehörigen Werte

- Bestimmung der Quartile $q_{0.025}$ und $q_{0.975}$

- R:

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

- Oder kürzer:

```
qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)
## [1] 70.60054 129.39946
```

- 95 % der Menschen haben einen IQ zwischen ungefähr 70 und 130

7 Zentraler Grenzwertsatz (durchschnittliche Wahrscheinlichkeit)

7.1 Standardfehler

Kennzahlen von S_n

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

Kennzahlen von \bar{X}_n

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

Varianz/Standardabweichung nimmt mit zunehmender Anzahl zu!

Standardfehler

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

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

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

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

7.2 Zentraler Grenzwertsatz

Durchschnittliche Wahrscheinlichkeit, näherungsweise Wahrscheinlichkeit i.i.d. Annahme.

Zentraler Grenzwertsatz

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

- summe $S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$ pnorm(q = , mean = n * mu, std = $\sqrt{n} \cdot \sigma_X$) !
- durchschnitt, Mittelwert $\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$ pnorm(q = , mean = mu, std = $\frac{\sigma_X}{\sqrt{n}}$) !
- ▶ 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

Zufallsvariable = (Erwartungswert, Varianz)

$N = (\mu, \sigma^2)$

Varianz = sigma ^2

Standardabweichung = sigma (=bei pnorm) → **immer std/sqrt(n)**

Mit den Kennzahlen von X arbeiten, außer es ist die Summe gesucht (selten).

Erwartungswert, sigma, Varianz, Standardabweichung.

- X_i : ZV für die gefallene Menge Schnee am Tag i
 - Annahme: i.i.d. → gerechtfertigt?
 - Es gilt $\mu = 1.5$ und $\sigma_X = 0.3$
 - Schneemenge (Summe) S_{50} der nächsten 50 Tage
 - Soll 80 nicht übersteigen $E(S_{50}) = 50 \cdot 1.5$
 - Es gilt annähernd:
- $$S_{50} \sim N(50 \cdot \mu, 50 \cdot \sigma_X^2) = N(75, 4.5)$$
- Gesucht:
- $$P(S_{50} \leq 80) = 0.991$$
- ```
> pnorm(q= 80, mean= 50*1.5, sd=(50*(0.3/sqrt(50))))
[1] 0.9907889
```

Pear Büchel (HSU I) Zentrale Grenzwertsatz ASTAT: Block 09 37 / 40

### 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?

Pear Büchel (HSU I) Zentrale Grenzwertsatz ASTAT: Block 09 38 / 40

### 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  $\bar{X}_{16}$
  - Annähernd verteilt wie:
- $$\bar{X}_{16} \sim N\left(\mu, \frac{\sigma_X^2}{n}\right) = N\left(100, \frac{20^2}{16}\right) = N(100, 25)$$

Pear Büchel (HSU I) Zentrale Grenzwertsatz ASTAT: Block 09 39 / 40

- Gesucht:

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

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

parametrische Verteilung, Erwartungswert, Varianz

### Zentraler Grenzwertsatz

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

- Summe:  $S_n \sim N(n\mu, n\sigma^2)$  (approx.  $\sum_{i=1}^n (X_i - \mu) \sim N(n(\mu - \mu), n(\sigma^2))$ )
  - durchschnitt / Mittelwert:  $\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$  (approx.  $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ )
- Approximation wird mit grösserem  $n$  i.A. besser  
► Approximation besser, je näher die Verteilung von  $X_i$  bei der Normalverteilung  $N(\mu, \sigma^2)$  ist

Kennzahlen von  $S_n$  → Standardfehler von empirischen Mitteln

Summe

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

Kennzahlen von  $\bar{X}_n$

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

für anschaulich:

## 8 Hypothesentest, z-Test, t-Test

Bei z-Test, Standardabweichung ist geben.  $H_A < H_0$  (auch möglich bei einseitigem Test)

- *Modell*

$X_i$ : Inhalt der  $i$ -ten Büchse

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

- *Nullhypothese*

$$H_0 : \mu_0 = 500$$

- *Alternativhypothese*

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

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

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

Bei t-Test

Falls Standardabweichung nicht bekannt.

- *Modell:*  $D_1, \dots, D_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ ,  $\sigma$  wird durch  $\hat{\sigma}$  geschätzt.

- *Nullhypothese:*

$$H_0 : \mu_D = \mu_0 = 0$$

*Alternative:*

$$H_A : \mu_D < \mu_0$$

- *Signifikanzniveau:*

$$\alpha = 5\%$$

- *Testentscheid:*

### 8.1 Schätzwerte

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

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

*N = sagt dass es Schätzwerte sind*

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

### 8.2 Hypothesentest

#### Modell

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

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

### 8.3 Nullhypotesen, Alternativhypothese

$80 = \text{Mittelwert} (\mu)$

**Einseitiger Test**, falls sich etwas erhöht (z.B. max); Wert «eh» zu hoch; **zweiseitiger Test**, falls sich etwas verändert (Kleiner/tiefer oder höher/grosser).

Dies erkennt man an dem R-Output bei der Alternativhypothese (ob ein- oder zweiseitig):

```

Paired t-test
##
data: t.1 and t.2
t = 5.6569, df = 9, p-value = 0.0001554
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.7976252 Inf
sample estimates:
mean of the differences
1.18

> t.test(x, mu = 180, alternative = "two.sided")
 one sample t-test

data: x
t = -193.99, df = 11, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 180
95 percent confidence interval:
69.00479 71.49521
sample estimates:
mean of x
70.25

```

*less /*  
*"nur" grösser als 0*  
*→ einseitig*

*zwischen*

Nullhypothese = *repräsentiert Annahme*

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

Alternativhypothese = *widerspricht Nullhypothese*

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

### 8.4 Signifikanzniveau

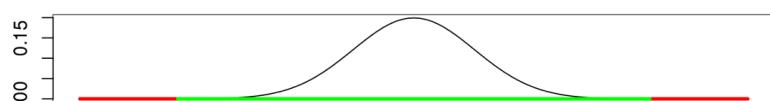
Signifikanzniveau  $\alpha$

Signifikanzniveau  $\alpha$ , gibt an, wie hoch das Risiko ist, das man bereit ist einzugehen, eine falsche Entscheidung zu treffen

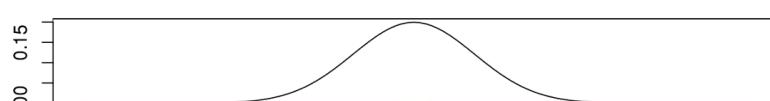
Für die meisten Tests wird ein  $\alpha$ -Wert von 0.05 bzw. 0.01 verwendet. Hier

$$\alpha = 0.05$$

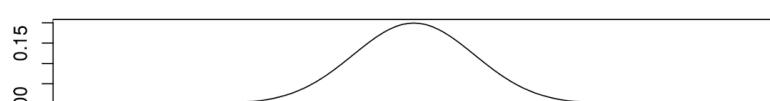
- Graphik:  $\alpha = 0.0001$  (nahe bei 0)



- Graphik:  $\alpha = 0.8$  (gross) *roter Bereich 80%*



- Graphik:  $\alpha = 0.05$  *→ Roter Bereich 5%*



## 8.5 Verwerfungsbereich

Bereich, wo Nullhypothese angenommen wird. Immer 5%:

- **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

## 8.6 p-Wert (Nullhypotesen, Alternativhypothese)

### p-Wert

Der *P-Wert* ist die Wahrscheinlichkeit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten wie das aktuell beobachtete.

- Testentscheid auch mit Hilfe des *p*-Wertes durchführen

### p-Wert und Statistischer Test

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

- ▶ Verwerfe  $H_0$  falls  $p\text{-Wert} \leq \alpha$
- ▶ Belasse  $H_0$  falls  $p\text{-Wert} > \alpha$

- Viele Computer-Pakete liefern den Testentscheid nur mit *p*-Wert

- Wie signifikant?

|                               |                              |
|-------------------------------|------------------------------|
| $p\text{-Wert} \approx 0.05$  | : schwach signifikant, “.”   |
| $p\text{-Wert} \approx 0.01$  | : signifikant, “*”           |
| $p\text{-Wert} \approx 0.001$ | : stark signifikant, “**”    |
| $p\text{-Wert} \leq 10^{-4}$  | : äußerst signifikant, “***” |

- *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

Der *p*-Wert (*p-value*) gibt die Wahrscheinlichkeit an, dass die beobachteten Daten oder ein noch extremes Ergebnis auftreten, wenn die Nullhypothese wahr ist. Die Nullhypothese ist die Annahme, dass es keinen Zusammenhang oder keinen Effekt gibt. Es ist wichtig zu beachten, dass der *p*-Wert allein keine Aussage über die Stärke oder Größe des beobachteten Effekts macht. Es gibt lediglich an, ob die beobachteten Daten mit der Nullhypothese vereinbar sind oder nicht.

## 8.7 Testentscheid

- In Beispiel oben

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

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

$$\mu \neq 80$$

## 8.8 z-Test vs t-Test

z-Test = Normalverteilung (qnorm) (Vergleichen von p-Wert mit Signifikanzniveau)

t-Test = t-Verteilung (t.test); Normalverteilt

- Wenn **in der Aufgabe die Varianz oder die Standardabweichung gegeben** wird, dann arbeiten wir mit dem **z-Test**. Mit pnorm() die p-Value ausrechnen und anschliessend die Entscheidung treffen.
  - Wenn **diese Angabe fehlt, oder wir die Angabe anzweifeln sollen**, dann arbeiten wir dem **t-Test**
- Bisher: Verfahren heisst **z-Test**
  - Stillschweigend vorausgesetzt: Standardabweichung *bekannt*
  - Praxis: Praktisch nie der Fall
  - Folgender **t-Test**: Setzt keine Standardabweichung voraus
  - Darum: **t-Test** viel wichtiger als **z-Test**
  - Vorgehen sehr ähnlich **z-Test** → Nur andere Verteilung
  - Wie vorher Annahme: Daten Realisierungen von

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

- Praxis: Annahme, dass  $\sigma_X$  bekannt ist, meist unrealistisch

Peter Büchel (HSLU I)

Hypothesentest

ASTAT: Block 09

## 8.9 t-Verteilung

Wenn die Varianz geschätzt ist, dann müssen wir einen t-Test machen.

### t-Verteilung

Die Verteilung der Teststatistik beim t-Test unter der Nullhypothese

$$H_0 : \mu = \mu_0$$

ist gegeben durch

$$T = \bar{X}_n \sim t_{n-1} \left( \mu, \frac{\sigma_X}{\sqrt{n}} \right)$$

*grösse Stichprobe  
Varianz geschätzt*

wobei  $t_{n-1}$  eine t-Verteilung mit  $n - 1$  Freiheitsgraden ist

## 9 Vertrauensintervall, Zweistichprobentest, Wilcoxon-Test

### 9.1 Wilcoxon Test

Wenn die Daten nicht normalverteilt sind.

#### 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)

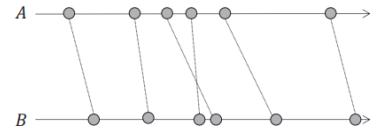
Selbst in den ungünstigsten Fällen ist er nie viel schlechter

### 9.2 Vertrauensintervall

Intervall, das angibt, wo, grob gesagt, der wahre Mittelwert mit einer bestimmten vorgegebenen W'keit liegt. Falls Daten nicht Normalverteilt sind.

### 9.3 Gepaarte Stichproben

- **Vorher-Nachher-Messungen**
- Pro Versuchseinheit (z.B. Prüfkörper) zwei Messungen (z.B. einmal Gerät A und einmal Gerät B) an dieselben Objekten
- Veränderung/Effekt innerhalb derselben Gruppe untersuchen
- Jede Beobachtung einer Gruppe kann eindeutig einer Beobachtung der anderen Gruppe zugeordnet werden
- Stichprobengrösse ist in beiden Gruppen zwangsläufig gleich
- Beobachtungen sind nicht unabhängig, da an gleicher Versuchseinheit zweimal gemessen wird
- Jede Versuchseinheit (z.B. Prüfkörper) wird an beiden Messgeräten gemessen
- Differenz zwischen den einzelnen Messungen kann beurteilt werden (Paare bilden und Differenz anschauen)



### 9.4 Ungepaarte (unabhängige) Stichproben

- 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“
- Gruppen von Objekten/Personen, bei denen jeweils eine Messung durchgeführt wird
- Unterschied zwischen zwei unabhängigen Gruppen/Bedingungen untersuchen
- keine Möglichkeit, Messungen an dieselben Objekten/Personen durchzuführen
- Keine Zuordnung von Beobachtungen möglich
- Stichprobengrösse können verschieden sein (müssen aber nicht!)
- Man kann die eine Gruppe vergrössern, ohne dass man die andere vergrössert
- Differenz nur über **Mittelwert/Durchschnitt** der aller Daten 1 und aller Daten 2

Beispielsweise könnten in einer Studie die Gewichtsveränderungen von zwei verschiedenen Diätgruppen verglichen werden, wobei jede Gruppe ihre eigenen Messungen hat.



# 10 Lineare Regression

- Allgemein: Quantitative Zielgröße  $Y$  und  $p$  verschiedene Prädiktoren  $X_1, X_2, \dots, X_p$
  - Annahme: Es besteht *irgendein* Zusammenhang zwischen  $Y$  und  $X_1, X_2, \dots, X_p$
  - Allgemeine Form:
- $$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$
- $f$  irgendeine feste, aber unbekannte Funktion von  $X_1, X_2, \dots, X_p$
  - Größe  $\varepsilon$ : Zufälliger Fehlerterm unabhängig von  $X_1, X_2, \dots, X_p$  mit Mittelwert 0

**Reduzible Fehler:** Schätzung mit statistischen Methoden verbessern.

**Irreduzible Fehler:** Fehler kann nicht beeinflusst werden, wie gut auch die Schätzung ist.

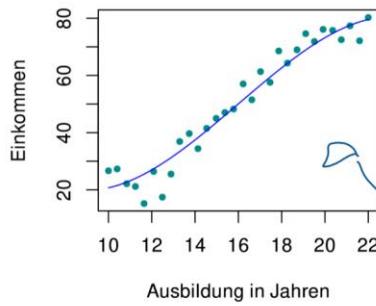
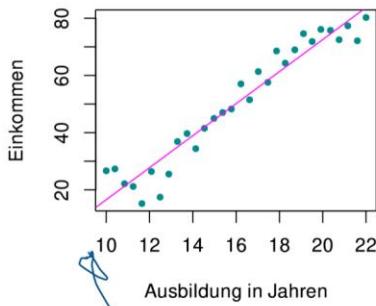
## Lineare Regression mit R

```
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  $b = 0.0467$

## 10.1 Lineares und kubisches Modell



Aus Daten: **Lineares Modell** (oben links):

$$f(X) = \beta_0 + \beta_1 X$$

Auch **kubisches Modell** (Polynom 3. Grades) möglich (oben rechts):

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

$$f(x) = m \cdot x + q$$

$\Rightarrow$  hängt von Datensatz ab

## 10.2 Multiple und einfache Regression

Zielvariable, Prädiktor

- Beispiel **Werbung**: Lineares Modell:

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

multiple  
Regression

- Beispiel **Einkommen**: Lineares Modell:

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

↓  
Zielvariable      ↓  
Prädiktor

einfache  
Regression

## 10.3 Methode der kleinsten Quadrate

Residuum:

Vorhergesagter Wert für  $Y$  abhängend vom  $i$ -ten Wert von  $X$ , also  $x_i$ :

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

*i*-tes Residuum:  
 $r_i = y_i - \hat{y}_i$

*verdorben*

*Messener Wert (von der Datenbank)*

Differenz zwischen dem  $i$ -ten *beobachteten* Wert der Zielgröße und dem  $i$ -ten von unserem linearen Modell *vorhergesagten* Wert der Zielgröße

## 10.4 Summe der Quadrate der Residuen (RSS)

**Summe der quadrierten Residuen oder Fehler** im Regressionsmodell (*falls alle auf einer Geraden liegen, gibt es keine Fehler*). Der RSS gibt an, wie viel der Variabilität der abhängigen Variablen das Modell nicht erklären konnte. Wenn der RSS klein ist, dann erklärt das Modell die Variabilität der abhängigen Variablen gut. Wenn der RSS groß ist, bedeutet dies, dass das Modell einen großen Teil der Variabilität in der abhängigen Variablen nicht erklären kann und dass es möglicherweise eine bessere Modellierung der Daten erfordert.

Summe der Quadrate der Residuen (RSS genannt)

Es gilt dann

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

Oder äquivalent:

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

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

## 10.5 Differentialrechnung

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

## 10.6 R2 Modell

Abschätzung der Genauigkeit des Modells messen. Kann für jede Regression angewendet werden (quadratisch und linear). Kein guter Wert, um verschiedene Modelle zu vergleichen (je mehr Variablen, desto grösser der Wert). Erhöht sich immer, je mehr Variablen berücksichtigt werden

- Qualität einer linearen Regression abgeschätzt durch den *residual standard error (RSE)* und die *R<sup>2</sup>-Statistik*
- R<sup>2</sup>* wichtiger
- R<sup>2</sup>-Statistik:* Wert zwischen 0 und 1
- Sie gibt an, welcher Anteil der Variabilität in *Y* mit Hilfe des Modells durch *X* erklärt werden
- Wert nahe bei 1: ein grosser Anteil der Variabilität wird durch die Regression erklärt. Das Modell beschreibt also die Daten sehr gut.
- Wert nahe bei 0: Regression erklärt die Variabilität der Zielvariablen nicht

- Definition *R<sup>2</sup>*:

$$R^2 = \frac{\text{Varianz Modell}}{\text{Varianz Sample}}$$

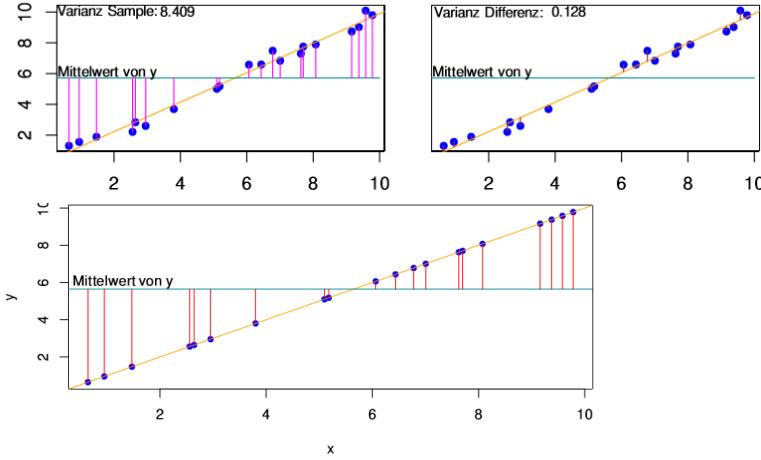
- Beispiel:

$$R^2 = \frac{8.281}{8.409} = 0.985$$

- Code:

```
summary(lm(y ~ x))$r.squared
[1] 0.9848312
```

- Nochmals Abbildung:



- Varianz: „Mittelwert“ der quadrierten Unterschiede der *y*-Werte der Datenpunkte zu  $\bar{y}$

- Varianz:

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

- Output:

- Korrelation:

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

- R<sup>2</sup>*:

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

- Varianz:

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

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

- Alternative Definition von *R<sup>2</sup>*:

$$R^2 = 1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$$

- Bedeutung:

- Varianz Differenz: Varianz des Samples, dass *nicht* durch das Modell erklärt wird
- $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$ : Anteil der Varianz vom Sample, der *nicht* vom Modell erklärt wird
- $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$ : Anteil der Varianz vom Sample, der vom Modell erklärt wird
- *R<sup>2</sup>*: Anteil der Varianz vom Sample, der vom Modell erklärt wird

## 10.7 Hypothesentest von $\beta_1$

Hypothesentest: Statistische Signifikanz von  $\beta_1$

- Standardfehler: Hypothesentest für die Regressionsparameter durchführen

- Häufigste Hypothesentest: Testen der *Nullhypothese*

$H_0$ : Es gibt *keinen* Zusammenhang zwischen  $X$  und  $Y$

- *Alternativhypothese*

$H_A$ : Es gibt *einen* Zusammenhang zwischen  $X$  und  $Y$

- Mathematisch:

$$H_0: \beta_1 = 0$$

- Gegen:

$$H_A: \beta_1 \neq 0$$

$$Y \approx \beta_0 + \beta_1 \cdot X$$

- $\beta_1 = 0$ , dann:



$$Y = \beta_0 + \varepsilon$$

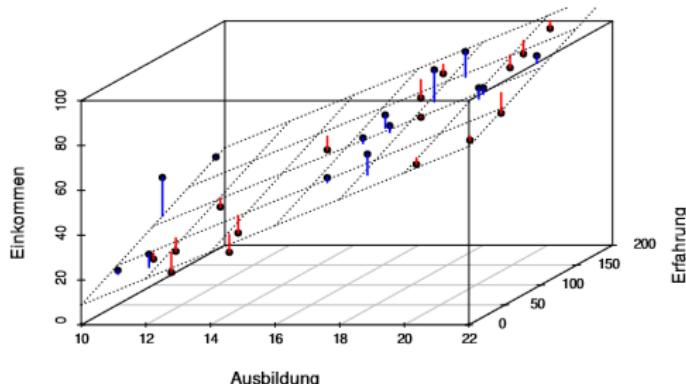
- $Y$  hängt *nicht* von  $X$  ab

- Nullhypothese testen:  $\hat{\beta}_1$  genügend weit von 0 weg, damit  $\beta_1$  nicht 0

- Mit *t*-Statistik

# 11 Multiple lineare Regression

- Analog einfaches lineares Regressionsmodell: Suchen Ebene, die am „besten“ zu den Datenpunkten passt



## 11.1 Zusammenhang zwischen Variablen und Zielgröße

- Hypothesentest:
- Multiple lineare Regression mit  $p$  erklärenden Variablen: Alle Regressionskoeffizienten ausser  $\beta_0$  Null sind (keine Variable hat Einfluss):

$$\beta_1 = \beta_2 = \dots = \beta_p = 0$$

- Nullhypothese

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

- Alternativhypothese

$$H_A : \text{mindestens ein } \beta_i \text{ ist ungleich 0}$$

- Berechnung der F-Statistik mit p-Wert

## 11.2 Bestimmung der erklärenden Variablen

- F-Statistik beurteilen (somit erkennen: haben die erklärenden Variablen überhaupt einen Einfluss auf die Zielgröße?) (F-Statistik; über 0.05 kein Einfluss)
  - P-Werte von Prädiktoren analysieren (welche Werte haben einen Einfluss, falls unter 0.05?)
- 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

### 11.3 Lineare vs Multiple Regression

- 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

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_3 + 1)$$

Multiple Regression interpretiert:

- 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   |

## 12 Qualitative Variablen

Synonym = Faktoren



Indikatorvariable (Dummy-Variable), die nur zwei mögliche numerische Werte annehmen kann.

### 12.1 Qualitativ vs. Quantitativ

```
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"
```

## 12.2 Koeffizientenschätzung

**Wichtig, + oder - sagen nicht ob positiver oder negativer Wert; muss beim Koeffizienten schauen z.B. B\_1.**

- Wichtig: Vorhersagen für die Zielgröße hängen *nicht* von Kodierung ab

- Einziger Unterschied: Interpretation der Koeffizienten

  
 muss und  
 Gewicht  
 nicht auf  
 Zielgrößen

### Beispiel

- Für Gender:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ 0 & \text{falls } i\text{-te Person männlich} \end{cases}$$

- Verwenden diese Variable als erklärende Variable im Regressionsmodell

- Modell:

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

- $\beta_0$ : durchschn. Kreditkartenrechnungen der Männern
- $\beta_0 + \beta_1$ : durchschn. Kreditkartenrechnungen der Frauen
- $\beta_1$ : durchschn. *Unterschied* der Rechnungen Männern/Frauen

- Tabelle: Koeffizientenschätzungen für unser Modell:

|                 | Koeffizient | Std.fehler | t-Statistik | P-Wert   |
|-----------------|-------------|------------|-------------|----------|
| Intercept       | 509.80      | 33.13      | 15.389      | < 0.0001 |
| gender [female] | 19.73       | 46.05      | 0.429       | 0.6690   |

```

balance <- Credit[, "Balance"]
gender <- Credit[, "Gender"] == "Female"
round(summary(lm(balance ~ gender))$coef, digits = 5)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 509.80311 33.12808 15.38885 0.00000
genderTRUE 19.73312 46.05121 0.42850 0.66852

```

TRUE = 1  
 FALSE = 0  
 Stellen runden  
 nur Koeffizienten mitbringen

- Geschätzte durchschnittliche Rechnungen für Männer: \$ 509.80
- Geschätzter Unterschied zu Frauen: \$ 19.73
- Frauen: \$ 509.80 + \$ 19.73 = \$ 529.53
- p-Wert für Indikatorvariable  $\beta_1$  mit 0.6690 sehr hoch
- Kein statistisch signifikanter Unterschied der **balance** von Frauen und Männern

Zusammenhang zwischen Geschlecht und Kreditkartenrechnung

Immer zu Baseline betrachten, hier Baseline (Männer) 509.8 und Female +19.7 (da Female True). Die die nicht aufgeführt = Baseline.

- 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:

$$x_{i2} = \begin{cases} 1 & \text{falls } i\text{-te Person kaukasisch} \\ 0 & \text{falls } i\text{-te Person nicht kaukasisch} \end{cases}$$

- Level ohne Indikatorvariable (hier Afroamerikaner): Baseline  $\beta_0$

13 15 / 55

- Beide Variablen in Regressionsgleichung aufnehmen:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch } \beta_2 > 0 \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch } \beta_1 > 0 \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch } \beta_1 \beta_2 = 0 \end{cases}$$

- $\beta_0$ : Durchschn. Kreditkartenrechnungen von Afroamerikanern

- $\beta_1$ : Differenz der durchschn. Rechnungen von Afroamerikanern und Asiaten

- $\beta_2$ : Differenz der durchschn. Rechnungen von Afroamerikanern und Kaukasien

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:
(Intercept) β_0 531.00 46.32 11.464 <2e-16 ***
ethnicityAsian β_1 -18.69 65.02 -0.287 0.774
ethnicityCaucasian β_2 -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
```

$\beta_0, \beta_1, \beta_2$  Nullhypothese  
angenommen, kein Zusammenhang

## Mit qualitativer Variable (hier price):

Modell: Für **Urban** wählen wir die Dummy-Variable:

$$x_{2i} = \begin{cases} 1 & \text{falls } i\text{-te Person lebt in der Stadt} \\ 0 & \text{falls } i\text{-te Person lebt auf dem Land} \end{cases}$$

Für **US** wählen wir die Dummy-Variable

$$x_{3i} = \begin{cases} 1 & \text{falls } i\text{-te Person lebt in den USA} \\ 0 & \text{falls } i\text{-te Person lebt nicht in den USA} \end{cases}$$

Das Modell lautet dann

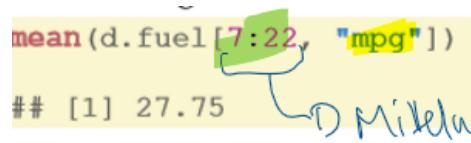
$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \text{Price} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \text{Price} + \begin{cases} \beta_2 + \beta_3 + \varepsilon_i & \text{falls } i\text{-te Person urban in den USA lebt} \\ \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person urban nicht in den USA lebt} \\ \beta_3 + \varepsilon_i & \text{falls } i\text{-te Person ländlich in den USA lebt} \\ \varepsilon_i & \text{falls } i\text{-te Person ländlich nicht in den USA lebt} \end{cases} \end{aligned}$$

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \text{Price} + \beta_2 x_{2i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \text{Price} + \begin{cases} \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person in den USA lebt} \\ \varepsilon_i & \text{falls } i\text{-te Person nicht in den USA lebt} \end{cases} \\ &= 13.03 - 0.055 \cdot \text{Price} + \begin{cases} 1.2 + \varepsilon_i & \text{falls } i\text{-te Person in den USA lebt} \\ \varepsilon_i & \text{falls } i\text{-te Person nicht in den USA lebt} \end{cases} \end{aligned}$$

## 13 Formelsammlung

Vektor: vektor1

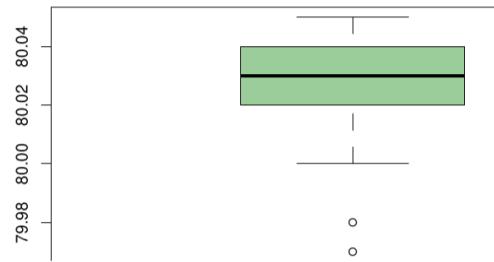
| Beschreibung                                                                                                                             | Befehl                                                                                                                                                                                       |
|------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Vektor bilden                                                                                                                            | <pre>vektor1 &lt;- c(183, 184, 154, 152)  winner &lt;- c(183, 191, 185, 185, 182, 182, 188, 1           182, 182, 193, 183, 179, 179, 175)</pre>                                             |
| Dokument zugreifen (CSV lesen)                                                                                                           | <pre>vektor1 &lt;- read.csv("C:\\\\Users\\\\No\u00ebl\\\\Documents\\\\ASTAT\\\\weather.csv") &gt; data &lt;- read.csv("C:\\\\users\\\\No\u00ebl\\\\Documents\\\\ASTAT\\\\weather.csv")</pre> |
| Dokument lesen (dat oder txt); mit sep= ` ` Kolonen bei Komma trennen                                                                    | <pre>vektor1 &lt;- read.table("C:\\\\Users\\\\No\u00ebl\\\\Documents\\\\ASTAT\\\\SW03\\\\geysir.dat", header = TRUE)</pre>                                                                   |
| Zahlen sortieren                                                                                                                         | <pre>sort(vektor1)</pre>                                                                                                                                                                     |
| Order<br>Macht eine Reihenfolge vom kleinsten zum gr\u00f6sstten Wert. Kleinster Wert hier 1, an Stelle 5; danach Wert 2 an Stelle 6,... | <pre>order(vektor1)  &gt; y ① ⑥ ④ ③ ⑤ ② ⑦ [1] 6 8 3 9 1 2 10 &gt; order(y) [1] 5 6 3 1 2 4 7</pre>                                                                                           |
| Wert von Zeile und Spalte abfragen                                                                                                       | <pre>vektor1[2,3]  [1] 11 <i>Zeile Spalte</i> &gt; data[2,3]</pre>                                                                                                                           |
| Daten einer Spalte ausgeben (Hier Luzern und Z\u00fcrich)                                                                                | <pre>vektor1[, c("Luzern", "Zurich")]  d) data[, c("Luzern", "Zurich")]  ##      Luzern Zurich ## Jan       2      4 ## Feb       5      0 ## Mar      10      8</pre>                       |

|                                                                                           |                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Daten untersuchen                                                                         | head(vektor1)<br>?vektor1                                                                                                                                                                                  |
| Bestimmte Spalte sortieren                                                                | order(vektor1[, 'Zurich'])<br><br><code>order(data[, 'Zurich'])</code><br><br><code>## [1] 2 1 3 4 5 6</code>                                                                                              |
| Zeilen 1-5 ausgeben.                                                                      | vektor1[1:5, ]<br><br><code>d.fuel[1:5, ]</code><br><br><code>##   X weight mpg type</code><br><code>## 1 1 2560 33 Small</code><br><code>## 2 2 2345 33 Small</code><br><code>## 3 3 1845 37 Small</code> |
| Mittelwert/Durchschnitt einer Spalte berechnen, hier Spalte 3                             | mean(vektor1[, 3])<br><br><code>mean(d.fuel[, 3])</code><br><br><code>## [1] 24.58333</code><br><br><code>mean(d.fuel[, "mpg"])</code><br><br><code>## [1] 24.58333</code>                                 |
| Mittelwert/Durchschnitt mehrerer Zeilen einer Spalte berechnen; Spalte «mpg», Zeilen 7-22 | mean(vektor1[7:22, 3])<br><br><code>mean(d.fuel[7:22, "mpg"])</code><br><br><code>## [1] 27.75</code><br><br>           |
| Länge der Vektoren bestimmen                                                              | length(vektor1)<br><br><code>length(winner)</code><br><br><code>## [1] 19</code>                                                                                                                           |
| Einträge 6 bis 10 des Vektors vektor1 ausgeben                                            | vektor1[6:10]<br><br><code>winner[6:10]</code><br><br><code>## [1] 182 188 188 188 185</code>                                                                                                              |

|                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 3., 5. und 10. bis 12. Eintrag auswählen.                                            | <pre>vektor1[c(3,5,10:12)]</pre> <pre>winner[c(3, 5, 10:12)]</pre> <pre>## [1] 185 182 185 185 177</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| 7. und 8. Eintrag ändern/ersetzen.                                                   | <pre>vektor1[7] &lt;- 189</pre> <pre>vektor1[8] &lt;- 189</pre> <pre>winner[7] &lt;- 189</pre> <pre>winner[8] &lt;- 189</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| (Empirische) Varianz                                                                 | <pre>var(vektor1)</pre> <pre>var(winner)</pre> <pre>## [1] 22.09942</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| ** (NICHT empirische)<br><i>idR diese verwenden.</i>                                 | <pre>mean((vektor11 - vektor12)^2)</pre> <pre>var.X &lt;- sum((werte - ew)^2 * 1/3)</pre> <pre>x &lt;- c(1, 2, 3, 4, 5, 6)</pre> <pre>ave &lt;- mean(x)</pre> <pre>var &lt;- mean((x - ave)^2)</pre> <pre>var</pre> <pre>## [1] 2.916667</pre> $\text{Var}(X) = \frac{1}{3} \cdot (0 - 7)^2 + \frac{1}{3} \cdot (10 - 7)^2 + \frac{1}{3} \cdot (11 - 7)^2 = 24.6667$ <p>Gesamtwert <math>\rightarrow</math> Durchschnittswert<br/> <math>\text{var.X} &lt;- \text{sum}((\text{werte} - \text{ew})^2 * 1/3)</math><br/> <math>\text{var.X}</math><br/> <math>\## [1] 24.66667</math> <math>\rightarrow</math> Wahrscheinlichkeit (<math>p</math>)</p> |
| Mittelwert/Durchschnitt; Achtung<br>Kommazahlen gehen nicht! Zuerst Vektor<br>machen | <pre>mean(vektor1)</pre> <pre>mean(opponent)</pre> <pre>## [1] 181.0526</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

|                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|--------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Standardabweichung                                                 | <pre>sd(vektor1)</pre> <pre>sd(winner)</pre> <pre>## [1] 4.701002</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Wurzel ziehen                                                      | <pre>sqrt(4)</pre> <pre>winner.sd &lt;- sqrt(winner.var)</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Summe                                                              | <pre>sum(5+6)</pre> <pre>winner.var &lt;- sum((winner - mean(winner))^2) / (length(winner)-1)</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Graphen beschriften/zeichnen<br>(main, col, xlab, ylab, type, lty) | <pre>plot(vektor1,       type = "l",       col = "blue",       lty = 2,       main = "Haupttitel",       xlab = "Ein paar Zahlen",       ylab = "Andere Zahlen"     )</pre> <p>c) i) Die Optionen <u>main = "..."</u>, <u>col = "..."</u>, <u>xlab = "..."</u> und <u>ylab = "..."</u> dürften klar sein.</p> <p>Die Option <u>type = "..."</u> gibt den Linientyp an. Siehe auch<br/> <a href="https://www.dummies.com/programming/r/how-to-create-different-plot-types-in-r/">https://www.dummies.com/programming/r/how-to-create-different-plot-types-in-r/</a></p> <p>Die Option <u>lty = "..."</u> gibt den Linientyp für „durchgezogene“ Linien vor. Siehe auch</p> |

|                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Mit abline(...) weitere (drei) Linien zur Grafik hinzufügen.</p> <p>x-Achse (v)</p> <p>y-Achse (h)</p> <p>Steigung</p> <p>a = y-Achsenabschnitt<br/>b = Steigung</p> <p>a, b : Einzelwerte, die den Achsenabschnitt und die Steigung der Linie angeben</p> <p>Unterstes: <math>y=2x+1</math></p> | <pre>abline(v = 3, col = "green", lty = 1) //durchgezogen, grün abline(h = 4, col = "red", lty = 3) // gepunktet, rot abline(a = 1, b = 2, col = "brown", lty = 5) // gestrichelt lang, braun ) <math>\nearrow x\text{-Achse} = v</math> <math>\nearrow y\text{-Achse} = h</math> </pre> <pre><b>abline(v = 3, col = "green", lty = 1) abline(h = 4, col = "red", lty = 3) abline(a = 1, b = 2, col = "brown", lty = 5)</b></pre> |
| <p>Median (R sortiert Daten automatisch)</p>                                                                                                                                                                                                                                                        | <pre>median(vektor1)  <b>median(waageA)</b> ## [1] 80.03</pre>                                                                                                                                                                                                                                                                                                                                                                    |
| <p>Quantile/Quartile</p> <p>p = 0.25 = Prozent 25% (unteres Quartal)</p> <p>p = 0.75 = 75% (oberes Quartal)</p> <p>type=2 = Standardrundung</p>                                                                                                                                                     | <pre>quantile(vektor1, p = 0.25, type = 2) quantile(vektor1, p = 0.75, type = 2) <b>quantile(waageB, p = 0.25, type = 2)</b> ## 25% ## 79.96 # Syntax für das obere Quartil: p=0.75  <b>quantile(waageA, p = 0.75, type = 2)</b> ## 75% ## 80.04</pre>                                                                                                                                                                            |
| <p>Quartilsdifferenz: Die Hälfte der Messwerte liegt im Bereich von 0.02</p>                                                                                                                                                                                                                        | <pre>IQR(vektor1, type=2)  <b>IQR(waageA, type=2)</b> ## [1] 0.02</pre>                                                                                                                                                                                                                                                                                                                                                           |

|                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quartalsdifferenz & Quartile zusammen                                                                                                                                                              | <pre>quantile(vektor1, p = c(0.25, 0.75), type = 2) quantile(noten, p = c(0.25, 0.75), type = 2) ## 25% 75% ## 3.80 5.35</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| In Schritten aufwärts mehreres miteinander berechnen (Abstand von 0.2, von 0.2 - 1)                                                                                                                | <pre>quantile(vektor1, p = seq(from = 0.2, to = 1, by = 0.2), type = 2) 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</pre>                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Boxplot</b><br><br>xlab steht für x-Label, die Beschriftung der x-Achse<br><br>ylab steht für y-Label, die Beschriftung der y-Achse<br><br>col steht für color<br><br>main steht für Haupttitel | <pre>boxplot(vektor1, col = "green")</pre><br><pre>boxplot(waageA,         col = "darkseagreen3"       )</pre><br> <p>The boxplot displays the distribution of the 'waageA' dataset. The y-axis ranges from 79.98 to 80.04. The box represents the interquartile range (IQR) from approximately 80.00 to 80.03. The median is at 80.02. Whiskers extend from 79.99 to 80.04. Two outliers are shown as small circles below the whiskers at approximately 79.99 and 80.01.</p> |

## Boxplot mit verschiedenen Gruppen

Mit axis(...) lässt sich die Beschriftung der Achsen ändern:

Mit side = ... wird die Seite gewählt: 1 unten (x), 2 links (y), 3 oben, 4 rechts

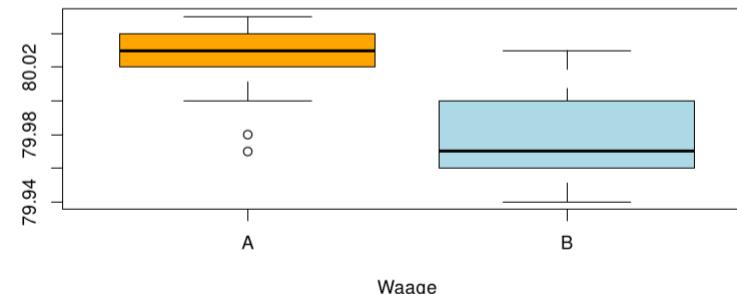
Mit at = ... werden die Stellen auf der jeweiligen Seite gewählt. In diesem Fall unten an der Stelle 1 und 2 (hier A und B).

Mit labels = ... wird angegeben, was an den entsprechenden Stellen geschrieben werden soll.

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

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

)
```



## Boxplot nach Art der Spalte

Nach «spray» wird geordnet

Werte von «count» werden genommen  
(Boxplot)

```
> InsectSprays
 count spray
1 10 A
2 7 A
3 20 A
4 14 A
5 14 A
6 12 A
7 10 A
8 23 A
9 17 A
10 20 A
11 14 A
12 13 A
13 11 B
14 17 B
```

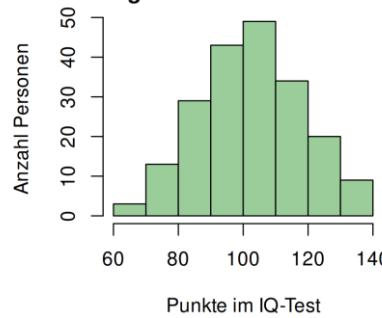
```
boxplot(count ~ spray,
```

```
 data = InsectSprays,
 col=c("orange", "blue", "darkseagreen", "deeppink", "brown",
 "aquamarine")
```

```
)
```

```
boxplot(count ~ spray,
 data = InsectSprays,
 col=c("orange", "blue", "darkseagreen", "deeppink",
 "brown", "aquamarine"))
```

|                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Folgen von Zahlen abbilden                                                                                                                                                                                                                                                                                                                                                                                                                 | seq(from = 2, to = 10, by =2)                                                                                                                                                                                                                                                                                                                                                                                       |
| Differenz zwischen zwei Spalten (Spalte 1 und Spalte 3 der Datei «mannfrau»).                                                                                                                                                                                                                                                                                                                                                              | <pre>alter.mann &lt;- mannfrau[, 1] alter.frau &lt;- mannfrau[, 3]  boxplot(alter.mann - alter.frau,         col = "orange")</pre>                                                                                                                                                                                                                                                                                  |
| <b>Mittelwert für verschiedene Zeilen</b> (hier gibt es mehrere Zeilen mit A, B, ....),<br>Spalte 1(Count) Mittelwert berechnen<br>(FUN=Funktion)<br>Spalte 2 (Spray) wird sortiert.<br><br><pre>&gt; InsectSprays   count spray  1    10    A  2     7    A  3    20    A  4    14    A  5    14    A  6    12    A  7    10    A  8    23    A  9    17    A 10    20    A 11    14    A 12    13    A 13    11    B 14    17    B</pre> | <pre>tapply(Dokumentname[, "Spalte1"], Dokumentname [, "Spalte2"], FUN = mean)  tapply(InsectSprays[, "count"], InsectSprays[, "spray"], FUN = mean) ##          A          B          C          D          E          F ## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667 ``` tapply(diet\$weight.loss, diet\$Diet, mean) ##          1          2          3 ## -3.300000 -3.025926 -5.148148</pre> |

|                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Spalte in Tabelle hinzufügen                                                                                                                         | dokumentname\$neueSpalte <- datenInSpalte<br><br><pre>diet\$weight.loss &lt;- diet\$weight6weeks - diet\$pre.weight<br/><br/>head(diet)<br/><br/>##   Person gender Age Height pre.weight Diet weight6weeks weight.loss<br/>## 1     25      0    41    171       60     2      60.0      0.0<br/>## 2     26      0    32    174      103     2     103.0      0.0<br/>## 3      1      0    22    159       58     1      54.2     -3.8<br/>## 4      2      0    46    192       60     1      54.0     -6.0<br/>## 5      3      0    55    170       64     1      63.3     -0.7<br/>## 6      4      0    33    171       64     1      61.1     -2.9</pre> |
| Spalte zugreifen                                                                                                                                     | dokumentname\$spaltenname                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Spaltennamen anzeigen                                                                                                                                | colnames(vektor1)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Spalte löschen / entfernen                                                                                                                           | Tabelle.1 <- within(Tabelle, rm(Spalte))<br><br><pre>&gt; Auto.2 &lt;- within(Auto, rm(name))</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Spaltennamen anhängen/erkennen                                                                                                                       | attach(vektor1)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <b>rnorm</b><br><br><u>Zufällige Werte auswählen für Normalverteilung</u><br><br>Daten = 200<br>Mittelwert = 100<br>Standardabweichung (Breite) = 15 | rnorm(n = 200, mean = 100, sd = 15)<br><br><pre>iq &lt;- rnorm(n = 200, mean = 100, sd = 15)<br/><br/>hist(iq,<br/>      col = "darkseagreen3",<br/>      xlab = "Punkte im IQ-Test",<br/>      ylab = "Anzahl Personen",<br/>      main = "Verteilung der Punkte in einem IQ-Test"<br/>)</pre> <p style="text-align: center;"><b>Verteilung der Punkte in einem IQ-Test</b></p>                                                                                                                                                                                              |

|                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Histogramm</b> für Daten vektor1                                                    | hist(vektor1, ...)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| 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                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Histogramm</b> für eine Zeile                                                       | hist(vektor1[, "Zeilenname"])                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| <b>Mehrere Histogramme zeichnen:</b>                                                   | par(mfrow = c(2, 4))                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <b>Par(mfrow:</b>                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Mfrow (Multi File Row (                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| (2) Zeilen                                                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| (4) Spalten                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Dass Bilder nebeneinander bzw. Zeilen dargestellt werden                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Breaks:</b>                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Anzahl Klassen festlegen (jedoch nur ein Vorschlag von R, R nimmt evtl. einen anderen) | <p>The figure displays four histograms of the 'iq' variable side-by-side. Each histogram has 'Histogram of iq' as its title. The x-axis for all is labeled 'iq' and ranges from 60 to 140. The y-axis is labeled 'Häufigkeit' (Frequency).</p> <ul style="list-style-type: none"> <li>The first histogram, titled '100 Klassen', shows 100 bins with a maximum frequency of approximately 6.</li> <li>The second histogram, titled '20 Klassen', shows 20 bins with a maximum frequency of approximately 15.</li> <li>The third histogram, titled 'Sturges-Regel von R', shows approximately 13 bins with a maximum frequency of approximately 30.</li> <li>The fourth histogram, titled '4 Klassen', shows 4 bins with a maximum frequency of approximately 60.</li> </ul> |

|                                      |                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Histogramm</b>                    |                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| breaks = Anzahl der Balken           | par(mfrow = c(2, 2))<br><br>hist(iq,<br>breaks = 100,<br>xlab = "100 Klassen",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>)<br><br>hist(iq,<br>breaks = 20,<br>xlab = "20 Klassen",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>)<br><br>hist(iq,<br>breaks = 3,<br>xlab = "4 Klassen",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>) | par(mfrow = c(2, 2))<br><br>hist(iq,<br>breaks = 100,<br>xlab = "100 Klassen",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>)<br><br>hist(iq,<br>breaks = 20,<br>xlab = "20 Klassen",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>)<br><br>hist(iq,<br>breaks = "sturges", # default R<br>xlab = "Sturges-Regel von R",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>)<br><br>hist(iq,<br>breaks = 3,<br>xlab = "4 Klassen",<br>ylab = "Häufigkeit",<br>col = "darkseagreen3"<br>)                                                                                                                                                                                                                                                                                                         |
| breaks = seq(41, 96, by = 11)        |                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| von 41-96 mit 11 Abstand             |                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Histogramm: normiert zeichnen</b> | hist(vektor1,<br>freq = F, main = "Histogramm von Waage A",<br>col = "darkseagreen3",<br>ylim = c(0, 25)<br>)                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| Die Höhe des Histogrammes.           | ylim = c(0, 25)                                                                                                                                                                                                                                                                                                                                                | <p>The figure shows a histogram titled 'Histogramm von Waage A'. The x-axis is labeled with values 79.96, 79.98, 80.00, 80.02, 80.04, and 80.0. The y-axis is labeled 'Density' and ranges from 0 to 25. The histogram consists of several bars. The first bar (79.96-79.98) has a height of approximately 7. The second bar (79.98-80.00) has a height of approximately 3. The third bar (80.00-80.02) has a height of approximately 11. The fourth bar (80.02-80.04) has a height of approximately 25. The fifth bar (80.04-80.06) has a height of approximately 3. The sixth bar (80.06-80.08) has a height of approximately 4. A vertical black line is drawn at x=80.0, extending from the bottom to the top of the plot area. A pink bracket on the left side indicates the y-axis scale from 0 to 25.</p> |

|                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Rechteck zeichnen:</p> <p>Zuerst unterer Punkt links, dann oberer Punkt rechts.</p>                                          | <pre>rect(80.02, 0, 80.04, 23.1, col = "darkseagreen4")</pre> <ul style="list-style-type: none"> <li>Code:</li> </ul> <pre>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")</pre>                                                                                                                                                                                                            |
| <p>Dichte ausrechnen eines Rechtecks von einem Histogramm</p>                                                                   | <ul style="list-style-type: none"> <li>Dichte der Klasse von 80.02 – 80.04 ist etwa 23</li> <li>Fläche dieses Balkens (dunkelgrüne Fläche in Abbildung):</li> </ul> $23 = \frac{0,46}{0,02} \quad (80.04 - 80.02) \cdot 23 = 0.46$ <p style="text-align: center;">grundfläche <math>\cdot</math> Höhe</p> <ul style="list-style-type: none"> <li>Fläche mit 100 multipliziert: Prozentzahl der Daten, die in diesem Balken liegen</li> <li>Also etwa 46 % der Daten befinden sich zwischen 80.02 und 80.04</li> </ul> |
| <p>Lineare Regression (Linear Model)</p>                                                                                        | <p>lm()</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| <p>Lineare Regression wird von der Form <math>y = a + bx</math> angepasst</p> <p>mit <code>summary(lm...)</code> mehr infos</p> | <p>lm(y~x)<br/>lm(preis ~ seiten)</p> <h3>Lineare Regression mit R</h3> <p>stat endet Schritte</p> <pre>seiten &lt;- seq(50, 500, 50)  preis &lt;- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5,          26.1, 29.1)  lm(preis ~ seiten) ##</pre>                                                                                                                                                                                                                                                                  |
| <p>Alle Daten eines Datensatzes interpretieren.</p>                                                                             | <p>lm(formula = medv ~ ., data = Boston)</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |

| <p><b>Bestimmen von B0 und B1 und die Regressionsgerade.</b></p> <p>Zielvariable: spalteY/Verkauf/B0</p> <p>Prädiktor: spalteX/TV/B1</p> <p>Für Nullhypothese B1 vergleichen, falls dieser p-Wert unter 0.05, dann 0 Hypothese verwerfen.</p>           | <p><code>lm(tabelle\$spalteY ~ tabelle\$spalteX)</code></p> <p>Beispiel <code>Werbung</code>: <math>\hat{\beta}_0</math> und <math>\hat{\beta}_1</math> und die Regressionsgerade bestimmen</p> <pre>lm(Verkauf ~ TV) ##  ## Call: ## lm(formula = Verkauf ~ TV) ##  ## Coefficients: ## (Intercept)          TV ## 7.03259        0.04754</pre> <p>Wert unter <code>Intercept</code>: <math>\hat{\beta}_0 \rightarrow y</math>-Achsenabschnitt</p> <p>Wert unter <code>TV</code>: <math>\hat{\beta}_1 \rightarrow</math> Steigung der Geraden</p> <p>Lineares Modell:</p> $Y \approx 7.03 + 0.0475X$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|--------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| <p><b>Regressionsgerade zeichnen mit Streudiagramm</b></p> <p><code>plot(x,y)</code></p> <p><b>Bei abline sind die Spalten immer verdreht!</b></p> <p>Zuerst y, dann x</p> <p><b>Regression der kleinsten Quadrate (lineares Regressionsmodell)</b></p> | <pre>plot(seiten, preis,       col = "orange",       pch = 19,       xlab = "Seitenzahl",       ylab = "Buchpreis" ) abline(lm(preis ~ seiten), col = "deepskyblue")</pre> <pre>seiten &lt;- seq(50, 500, 50) preis &lt;- 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")</pre> <table border="1"> <caption>Data points from the scatter plot</caption> <thead> <tr> <th>Größe des Vaters (in cm)</th> <th>Größe des Sohnes (in cm)</th> </tr> </thead> <tbody> <tr><td>155</td><td>164</td></tr> <tr><td>158</td><td>166</td></tr> <tr><td>162</td><td>165</td></tr> <tr><td>165</td><td>166</td></tr> <tr><td>168</td><td>169</td></tr> <tr><td>170</td><td>170</td></tr> <tr><td>172</td><td>171</td></tr> <tr><td>174</td><td>172</td></tr> <tr><td>176</td><td>173</td></tr> <tr><td>178</td><td>174</td></tr> <tr><td>182</td><td>176</td></tr> <tr><td>184</td><td>177</td></tr> </tbody> </table> | Größe des Vaters (in cm) | Größe des Sohnes (in cm) | 155 | 164 | 158 | 166 | 162 | 165 | 165 | 166 | 168 | 169 | 170 | 170 | 172 | 171 | 174 | 172 | 176 | 173 | 178 | 174 | 182 | 176 | 184 | 177 |
| Größe des Vaters (in cm)                                                                                                                                                                                                                                | Größe des Sohnes (in cm)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 155                                                                                                                                                                                                                                                     | 164                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 158                                                                                                                                                                                                                                                     | 166                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 162                                                                                                                                                                                                                                                     | 165                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 165                                                                                                                                                                                                                                                     | 166                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 168                                                                                                                                                                                                                                                     | 169                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 170                                                                                                                                                                                                                                                     | 170                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 172                                                                                                                                                                                                                                                     | 171                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 174                                                                                                                                                                                                                                                     | 172                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 176                                                                                                                                                                                                                                                     | 173                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 178                                                                                                                                                                                                                                                     | 174                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 182                                                                                                                                                                                                                                                     | 176                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 184                                                                                                                                                                                                                                                     | 177                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                          |                          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

|                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Linie/Gerade zeichnen anhand der Steigung<br/>Plot muss zuerst erstellt worden sein.</p>                                                                                                                   | <pre>abline(lm(y ~x), col = &lt;&lt; deepskyblue &gt;&gt;) abline(lm(preis ~ seiten), col = "deepskyblue")</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <p>Streudiagramme für den gesamten Datensatz</p>                                                                                                                                                              | <pre>pairs(vektor1)</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <p><b>Streudiagramm (Scatterplot)</b></p> <p>Oben:</p> <p>X-Achse: Seite/vektor11<br/>y-Achse: Preis/vektor12</p> <p>Unten:</p> <p>y-Achse: Verkauf<br/>x-Achse: Zeitung</p> <p>Beide machen das gleiche.</p> | <pre>plot(vektor11, vektor12)  seiten &lt;- seq(50, 500, 50) preis &lt;- 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"     )  plot(vektor1 ~ vektor1, col = "darkcyan", xlab = "vektor11", ylab = "vektor1")  TV &lt;- Werbung[, 1] Radio &lt;- Werbung[, 2] Zeitung &lt;- Werbung[, 3] Verkauf &lt;- Werbung[, 4] <del>Spalte A</del>  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")</pre> |

|                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                        |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Koeffizenz der Regressionsgeraden<br/>(Gleichung herausfinden)</b></p> <p><math>y = a + bx</math></p> <p>Der Intercept (y-Achsenabschnitt) ist <math>a = 110.44</math> und die Steigung ist <math>b = 0.29</math>, also</p> <p><math>y = 110.44 + 0.29x</math></p> <p>Nimmt auf 1, <b>0.29</b> zu.</p> | <pre>lm(vektor1\$spalte1 ~vektor1\$spalte2)  lm(df\$groesse.frau~df\$groesse.mann)  ## ## Call: ## lm(formula = df\$groesse.frau ~ df\$groesse.mann) ## ## Coefficients: ## (Intercept) df\$groesse.mann ## 110.4440      0.2884</pre> |
| <p>Korrelation berechnen<br/>(Korrelationskoeffizienten)</p>                                                                                                                                                                                                                                                 | <p><code>cor(seiten, preis)</code> ODER <code>cor(vektor1)</code></p> <pre>cor(seiten, preis) ## [1] 0.9681122</pre>                                                                                                                   |
| <p>Paket für die Wahrscheinlichkeit</p>                                                                                                                                                                                                                                                                      | <p><code>library(MASS)</code></p>                                                                                                                                                                                                      |
| <p>Brüche berechnen</p>                                                                                                                                                                                                                                                                                      | <pre>A &lt;- 1/2 B &lt;- 3/2 fractions(A*B)</pre>                                                                                                                                                                                      |
| <p>Bestimmte Werte nicht mitzählen/ausschliessen. Hier werden die "Abwesenden" aus der Spalte "Version" nicht mitgezählt.</p>                                                                                                                                                                                | <pre>vektor11 &lt;- subset(vektor1, SpalteX != "abwesend") # abwesende herausstreichen punkte_anwesend &lt;- subset(punkte_roh, version != "abwesend") dim(punkte_anwesend)</pre>                                                      |
| <p>Anzahl Zeilen und Spalten herausfinden</p>                                                                                                                                                                                                                                                                | <pre>dim(vektor1) 20 (zeilen) 11(spalten)</pre>                                                                                                                                                                                        |
| <p>Daten aufspalten</p>                                                                                                                                                                                                                                                                                      | <pre>vektor11 &lt;- subset(vektor1,SpalteX == "Spaltenwert") # Daten spalten in Version A und B punkteA &lt;- subset(punkte_anwesend,version == "A") punkteB &lt;- subset(punkte_anwesend,version == "B")</pre>                        |
| <p>Spalte abschneiden (hier erste und letzte)</p>                                                                                                                                                                                                                                                            | <pre>vektor11 &lt;- vektor1[Zeilenvon:Zeilenbis] # letzte Spalten (version) in A abschneiden punkteA &lt;- punkteA[2:10]</pre>                                                                                                         |

|                                                                                                                                                                                                                |                                                                                                                                                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Spalten vertauschen (hier wird die zweite Spalte an die erste Stelle gesetzt, die erste an die zweite, etc.)                                                                                                   | <pre>vektor11 &lt;- vektor1 [c("A2","A1","A4","A3","A6","A5","A8","A7","A9")] # Spalten der Version B tauschen und Namen der Aufgaben anpassen punkteB &lt;- punkteB [c("A2","A1","A4","A3","A6","A5","A8","A7","A9")] ...</pre> |
| Titel/Name der Spalten ändern                                                                                                                                                                                  | <pre>names(vektor1) &lt;- c("Titel1","Titel2","Titel3") names(punkteB) &lt;- c("A1","A2","A3","A4","A5","A6","A7","A8","A9")</pre>                                                                                               |
| Mehrere Listen/Teillisten zusammenführen.                                                                                                                                                                      | <pre>vektor13 &lt;- rbind(vektor1, vektor11) # Teillisten wieder zusammenführen punkte_bereinigt &lt;- rbind(punkteA,punkteB)</pre>                                                                                              |
| Spalte hinzufügen (hier werden die Werte einer Spalte zusammengezählt und in der Spalte Summe hinzugefügt)                                                                                                     | <pre>vektor1\$neueSpalte &lt;- Daten # Punktetotal als weitere Spalte anfügen (Summe Zeilensumme Spalte hinzufügen) punkte_bereinigt\$summe &lt;- rowSums(punkte_bereinigt[,c(1:9)])</pre>                                       |
| <b>Erwartungswert</b> berechnen sum                                                                                                                                                                            | <p>Wir verwenden zur Berechnung R:</p> <pre>x &lt;- 2:12 p &lt;- c(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1) / 36</pre><br><pre>E &lt;- sum(x*p) E ## [1] 7</pre>                                                                         |
| <b>Experiment simulieren mit sample</b><br><br><b>Size:</b> wie oft wir, dass wir das Experiment ausführen möchten.<br><br><b>Replace:</b> T(rue): Werte dürfen sich wiederholen<br><br>Zahlen 1-6 simulieren. | <pre>vektor1 &lt;- sample(1:6, size = 10, replace = T) x &lt;- sample(1:6, size = 10, replace = T) x ## [1] 3 3 3 4 3 6 5 2 3 6 mean(x) ## [1] 3.8</pre>                                                                         |

|                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Experiment wiederholen.</b></p> <p><b>Cat:</b> gibt Resultat aus.</p>                                                                                                                                                                                                                                                                                                              | <pre>For(i in 1:10) {   x &lt;- sample(1:6, size = 10, replace = T)   cat(mean(x), " ") }</pre> <p>• Simulieren 10mal 10 Würfe: Berechnen entsprechende Durchschnitte</p> <p>for (i in 1:10) {<br/>  x &lt;- sample(1:6, size = 10, replace = T)<br/>  cat(mean(x), " ")<br/>}<br/>## 3.3 3.5 4.2 3.4 2.9 3 3.9 2.9 4.1 2.7</p>                                                |
| <p><b>Normalverteilung (Fläche der Wahrscheinlichkeit)</b></p> <p>kumulative Verteilungsfunktion einer Normalverteilung zu berechnen.</p> <p>Q= Quantil = 130 (Messwert)</p> <p>Standardabweichung = 15</p> <p>Messwert Eingabe, Ergibt Wahrscheinlichkeit.</p> <p>pnorm (q = 130... → 0-130 (kleiner))</p> <p>1-pnorm (q=130,...) → über 130</p> <p>Wie viele zwischen 98 &amp; 104</p> | <p>pnorm(q = 130, mean = 100, sd = 15)</p> <p>pnorm(q = 130, mean = 100, sd = 15)<br/>## [1] 0.9772499</p> <p>• Berechnung mit R:</p> <p>1 - pnorm(q = 130, mean = 100, sd = 15)<br/>## [1] 0.02275013</p> <p>• Gesucht:</p> $P(98 \leq \bar{X}_{16} \leq 104) = 0.444$ <p>pnorm(q = 104, mean = 100, sd = 20/sqrt(16)) - pnorm(98, 100, 20/sqrt(16))<br/>## [1] 0.4435663</p> |

## Normalverteilung: Quartil berechnen (%)

Quartil 1 = 0.025 (Prozent)

Quartil 2 = 0.975 (Prozent)

- unten 2.5% der Werte und oben 2.5% der Werte ergibt 95% der Werte
- 95% der Werte zwischen 70-129

Wahrscheinlichkeit für Normalverteilung eines Quartils berechnen.

Wahrscheinlichkeit Eingabe, Messwert Ausgabe

Diese Punkte sind die Grenze zum Verwerfungsbereich.

Bei zentraler Grenzwertsatz/durchschnittliche Wahrscheinlichkeit immer nur Sigma bei std und nie quadrieren wie in Formel. std/sqrt(n)

Durchschnitt\_Xi

Standardabweichung: 20

Mittelwert = 100

Anzahl (n) = 16

ergibt P-Wert!

`qnorm (p= 0.025, mean = 100, sd= 15)`

`qnorm (p= c(0.025, 0.975), mean = 180, sd= 7.4)`

```
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
```

`pnorm(q= 10, mean=100, sd = (20/sqrt(16)))`

- Es gilt  $\mu = 100$  und  $\sigma_x = 20 \rightarrow \text{std}$
- Annahme i.i.d.  $n=16$
- Betrachten durchschnittliche Lebensdauer  $\bar{X}_{16}$
- Annähernd verteilt wie:

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

Peter Büchel (HSLU I) Zentraler Grenzwertsatz ASTAT: Block 08

- Gesucht:

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

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

Zentraler Grenzwertsatz (siehe [Kapitel](#)), dort gut erklärt!

Falls Summe verlangt ist (Summe\_Xi)

Meistens ist der durchschnittliche Wert verlangt, dann nicht so!

Standardabweichung = 0.3

Mittelwert = 1.5

Anzahl = 50

ergibt P-Wert!

`pnorm(q = 80, mean = (n*1.5), sd = (50*(0.3/sqrt(n))))`

- Es gilt  $\mu = 1.5$  und  $\sigma_X = 0.3$

- Schneemenge (Summe)  $S_{50}$  der nächsten 50 Tage

- Soll 80 nicht übersteigen

- Es gilt annähernd:

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

- Gesucht:

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

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

$$\begin{aligned} \text{Std} &= 0.3 \\ \text{Var} &= 0.09 \end{aligned}$$

$\Rightarrow$  Möglich gesucht

Schleifen for

Mit cat die Werte ausgeben

```
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")
```

```
}
```

```
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
```

Erwartungswert / Durchschnittlicher Wert

```
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
```

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

Zahlen runden

Round( ...., 2)

|                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                      |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| t-test: qnorm () = qt ()                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                      |
| t-test: pnorm () = pt()                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                      |
| Freiheitsgrad                                                                                                                                                                                                                                                                                  | Df                                                                                                                                                                                                                   |
| <b>z-test: Zufallszahl</b><br><br>Immer gleiche Zufallszahlen: set.seed<br>Mit Zufallszahl 1 starten, 2 Kommastellen                                                                                                                                                                           | set.seed(Anfangswert)<br><br><pre>set.seed(1) waageA.sim1 &lt;- 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</pre>                                           |
| <b>z-test: Grenze des Verwerfungsbereichs</b><br><br>Signifikanzniveau 5% = 0.05 = p; einseitiger Test. Bei MEAN angenommener Wert/Nullhypothese nehmen! (mu)<br>Mit Mittelwert (mean(171.48)) dann vergleichen (hier verworfen, da 171 in 0-174.)                                             | qnorm(p=0.05, mean=180, sd=10/sqrt(8))<br><br><pre>qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8)) ## [1] 174.1846</pre>                                                                                                |
| <b>z-Test: Verwerfungsbereich / Quantile / (Vertrauensintervall)</b><br><br>Ausserhalb dieses Bereichs (79.98-80.16) wird die Nullhypothese verworfen;<br>95% der Werte dazwischen                                                                                                             | qnorm(p = c(0.025, 0.975), mean = 80, sd = 0.02/sqrt(6))<br><br>• Grenzen entsprechen den 0.025- und 0.975-Quantilen<br><br><pre>qnorm(p = c(0.025, 0.975), mean = 80, sd = 0.02/sqrt(6)) ## [1] 79.984 80.016</pre> |
| <b>p-Wert: Wahrscheinlichkeit</b><br><br>Durchschnitt/Mittelwert(mean)=70.25<br>Nullhypothese =70 (mu), angenommen<br>Anzahl = 12<br>Standardabweichung = 1.5<br>Nullhypothese (falls Wert unter 0.05 verworfen)<br>Alternativhypothese<br>Signifikanzwert = 0.05 (immer)<br><i>Bei z-Test</i> | pnorm(q = 70.25, mean = 70, sd = 1.5/sqrt(12))<br><br><pre>pnorm(q = 70.25, mean = 70, sd = 1.5/sqrt(12)) ## [1] 0.7181486</pre>                                                                                     |

|                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Wilcox Test:</b></p> <p>Beidseitiger Test</p> <p>X: Mehrere Werte</p>                                                                                                                                                                                                                                                                                            | <pre>wilcox.test(mf\$frau, mf\$mann, alternative = "two.sided", paired=TRUE) x &lt;- 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</pre>                                                                                                                                                                                                                                                                                                                                                            |
| <p><b>t-test</b></p> <p>Hypothesentest</p> <p>Mindestwert<br/>(angenommen); mu; Nullhypothese = 180;<br/>Stichprobemittelwert</p> <p>Mit alternative=less <b>einseitiger</b>/greater Test</p> <p>x -&gt; Datenreihe</p> <p>p-Wert</p> <p>mean of x = mean(x)</p> <p><b>Immer schauen, dieser Wert mit kleinerem Median zuerst bei less! Bei greater umgekehrt!</b></p> | <p>t.test(x, mu = 180, alternative = "less")</p> <ul style="list-style-type: none"> <li>• t-Test nach unten:</li> </ul> <pre>groesse &lt;- 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")</pre> <ul style="list-style-type: none"> <li>• Befehl t.test(...):</li> </ul> <pre>t.test(x, mu = 5) ## ## One Sample t-test ## ## data: x ## t = 0.51041, df = 19, p-value = 0.6156 ## alternative hypothesis: true mean is not equal to 5 ## 95 percent confidence interval: ##  4.333353 6.096647 ## sample estimates: ## mean of x ##  5.215</pre> <p><i>politis wichtig sollte sein, wir SITZEN</i></p> <p><i>Vergissintervall, um zu schauen ob Nullhypothese angenommen wird</i></p> |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Statistischer t-Test</b> für gepaarte Stichproben<br/>(Normalverteilte Daten)</p> <p>Normalerweise Nullhypothese <math>\mu = 0</math></p> <p>Anpassen falls nur <b>einseitig!!!!</b></p> <p>Signifikanzniveau 5%</p> <p>Mean of the differences: durchschnittlichen Unterschied zwischen den vorherigen und nachherigen Werten an. Es wird Wert 1 – Wert 2 gerechnet, somit diese positiv oder negativ interpretieren. Ansonsten mit mean sich helfen. Wert1 – Wert2 = + (dann Wert2 kleiner bzw. Wert1 grösser).</p> <p>Falls 0 nicht in Confidence Intervall, Nullhyp verwerfen (hier verwerfen; Differenzen).</p> | <pre>t.test(ma, mb, alternative="two.sided", mu=0, paired=TRUE, conf.level=0.95)  vorher &lt;- c(25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28) nachher &lt;- 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</pre> <p>Nullhypothese wird auf Signifikanzniveau von 5 % verworfen, da p-Wert 0.001633 kleiner als 0.05</p> |
| <p><b>Statistischer t-Test</b> für ungepaarte Stichproben (Normalverteilte Daten)</p> <p>Normalerweise Nullhypothese <math>\mu = 0</math></p> <p>Anpassen falls nur <b>einseitig!!!!</b></p> <p>Confidence Intervall für Vertrauensintervall beachten.</p>                                                                                                                                                                                                                                                                                                                                                                 | <pre>t.test(ma, mb, alternative="two.sided", mu=0, paired=FALSE, conf.level=0.95)  x &lt;- 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 &lt;- 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</pre>                                                       |
| <p><b>Vertrauensintervall</b></p> <p>B0: Verkauf</p> <p>B1: TV</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | <pre>confint(lm(Tabelle\$SpalteY ~ Tabelle\$SpalteX), level = 0.95)  confint(lm(Verkauf ~ TV), level = 0.95)  ## ## 2.5 % 97.5 % ## (Intercept) 6.12971927 7.93546783 ## TV 0.04223072 0.05284256</pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |

## p-Wert von B1 (multiples Lineares Regressionsmodell)

```
summary(lm(y ~ x))
summary(lm(formula = medv ~ ., data =
Boston)) (Alles)
summary(lm(Boston$medv ~ Boston$Istat +
Boston$age))
```

Zielvariable/Output: spalteY/Verkauf/B0

Prädiktor: spalteX/TV/B1

R-Squared = R2

Für Nullhypothese B1 vergleichen, falls dieser p-Wert unter 0.05, dann H0 Hypothese verwerfen.

Falls Residuen klein, passt das Modell zur den vorhergesagten Werten. (Hier)

Gleichung hier: Verkauf = 7.032594 - 0.0475 \* TV

→ 7.03 ist der Wert für Verkauf bei TV = 0 && 0.0475 ist die Steigung (pro 1000CHF mehr, werden 47.5 TV mehr verkauft); immer auf Einheit des Prädiktors beziehen, nicht auf B0.

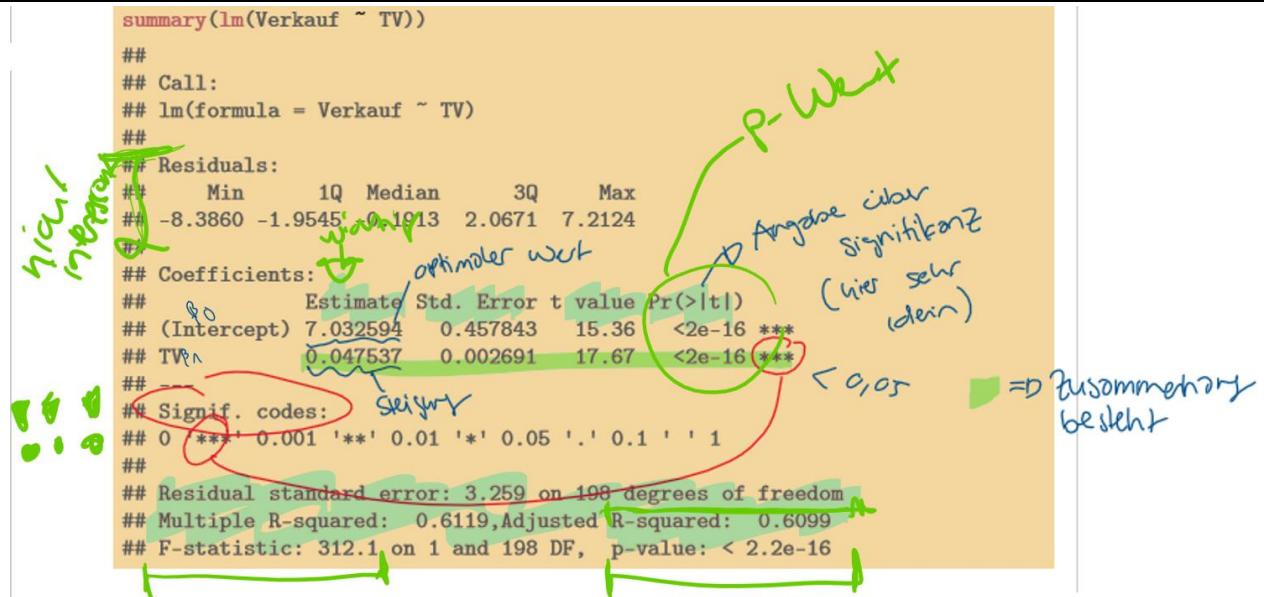
**Estimate:** Dies ist die geschätzte Größe des Regressionskoeffizienten für die unabhängige Variable. Es zeigt an, wie sich die abhängige Variable verändert, wenn sich die unabhängige Variable um eine Einheit ändert (Steigung), während alle anderen Variablen konstant gehalten werden.

**Std. Error:** Dies ist die Standardabweichung des Schätzfehlers für den geschätzten Regressionskoeffizienten. Es gibt an, wie genau die Schätzung des Koeffizienten ist. Wenn der Standardfehler hoch ist, ist die Schätzung unsicherer als wenn der Standardfehler niedrig ist.

**t value:** Dies ist der t-Wert für den Hypothesentest des Nullwertes für den Regressionskoeffizienten. Es zeigt an, wie groß die Abweichung des geschätzten Koeffizienten vom Nullwert ist und ob der geschätzte Koeffizient signifikant von Null verschieden ist oder nicht. Ein hoher t-Wert (positiv oder negativ) zeigt an, dass der Koeffizient signifikant von Null verschieden ist und dass der Zusammenhang zwischen der unabhängigen und abhängigen Variable statistisch signifikant ist. Ein t-Wert nahe Null zeigt an, dass der Koeffizient nicht signifikant von Null verschieden ist und dass der Zusammenhang zwischen der unabhängigen und abhängigen Variable nicht statistisch signifikant ist.

**RSE:** Je kleiner der RSE, desto besser passt das Modell zu den Daten.

Die Zielvariable Y nimmt zum B i Einheiten zu bzw. ab (kommt auf das Vorzeichen an), wenn X i um eine Einheit zunimmt.



**Bestimmtheitsmaß (Multiple R-squared)** gibt an, welcher Anteil der Varianz der abhängigen Variablen durch die unabhängigen Variablen erklärt wird. Ein Wert von 1 bedeutet, dass alle Varianz in den abhängigen Variablen durch die unabhängigen Variablen erklärt wird, während ein Wert von 0 bedeutet, dass die unabhängigen Variablen überhaupt keine Varianz in den abhängigen Variablen erklären können. Hier werden 61.119% der Variabilität in Verkauf durch TV mit linearer Regression erklärt. Hat nichts mit Signifikanz zu tun, sondern wie gut das Modell zu den Variablen passt.

**F-Statistik:** p-Wert für multiples lineares Modell. Einfluss der Variablen auf die Zielgröße (p-Wert). Zusammenhang zwischen Zielvariable und den Prädiktoren (unter 0.05 mindestens einer der Prädiktoren hat Einfluss auf die zu erklärende Variable ...). Ist höher anzusehen auf das ganze Modell als die einzelnen P-Werte.

Immer von unterem Wert ausgehen zu Intercept (siehe Rechts).

```
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 >0.05

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Wie gut passt Modell zu Output = 0.8956. > sehr gut
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
```

In gesamthaft Betrachtung (gesamtes Modell)

Was berechnet?  
Nicht wichtig

$\hat{\beta}$

Koeffizienten

B0 = Verkauf = 2.938  
B1 = TV  
B2 = Radio  
B3 = Zeitung

```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469 0.054459 237.285 <2e-16 ***
Price -0.054459 0.054459 -1.000 0.3164
UrbanYes -0.021916 0.021916 -0.999 0.3214
USYes 1.200573 0.054459 21.893 <2e-16 ***
Fals US, dann 1.2 mehr Verkauf
```

## Multiples lineares Regressionsmodell: Koeffizienten der Regressionsgleichung (mit Summary dasselbe)

Einkommen = B0 (Zielvariable)

Ausbildung = B1 (Prädiktor 1)

Erfahrung = B2 (Prädiktor 2)

Keine Ausbildung und keine Erfahrung = -50.1

Pro zusätzliches Jahr Ausbildung 5.89 CHF mehr

Pro Monat Erfahrung zusätzlich, 0.173 CHF mehr

`coef(lm(Verkauf ~ TV + Radio + Zeitung))`

- Schätzung von  $\beta_0, \beta_1$  und  $\beta_2$  mit R:

$$\hat{\beta}_0 = -50.086; \quad \hat{\beta}_1 = 5.896; \quad \hat{\beta}_2 = 0.173$$

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

- Multiples lineares Modell:

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

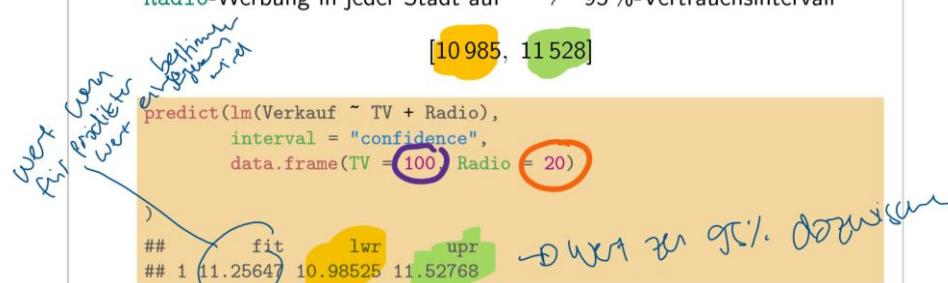
## Vertrauensintervall für Durchschnitt

**Predict** = Vorhersage, für Verkauf, wenn TV 100 und Radio 20. Verkauf wird in dem Intervall 10.98 – 11.5 liegen (zu 95%).

**Fit** = Vorhersage für den Wert von Verkauf bei den angegebenen Werten von TV und Radio.

- Wenden CHF 100 000 für TV-Werbung und CHF 20 000 für Radio-Werbung in jeder Stadt auf → 95 %-Vertrauensintervall

[10 985, 11 528]



```
predict(lm(Verkauf ~ TV + Radio),
 interval = "confidence",
 data.frame(TV = 100, Radio = 20))
```

```
predict(lm(Verkauf ~ TV + Radio),
 interval = "confidence",
 data.frame(TV = 100, Radio = 20))
```

)

```
fit lwr upr
1 11.25647 10.98525 11.52768
```

|                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Korrelationskoeffizienten</b>                           | <pre>cor(data.frame(TV, Radio, Zeitung, Verkauf)) cor(Auto)</pre> <div style="background-color: #ffffcc; padding: 10px;"> <pre>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</pre> <p style="color: blue; font-family: cursive;">je mehr in Radio investiert, umso mehr in Zeitung</p> </div> |
| <b>R2</b>                                                  | <pre>summary(lm(y ~ x))\$r.squared</pre> <ul style="list-style-type: none"> <li>▶ <math>R^2</math>:</li> </ul> <div style="background-color: #ffffcc; padding: 10px;"> <pre>summary(lm(y ~ x))\$r.squared ## [1] 0.9848312</pre> </div> <ul style="list-style-type: none"> <li>▶ Varianz:</li> </ul> <div style="background-color: #ffffcc; padding: 10px;"> <pre>var(y) ## [1] 8.40886</pre> </div>                                                                                                                                                                  |
| <b>R2 Quadratischer Zusammenhang</b>                       | <pre>summary(lm(y ~ I(x^2)))\$r.squared</pre> <div style="background-color: #ffffcc; padding: 10px;"> <pre>summary(lm(y ~ I(x^2)))\$r.squared ## [1] 0.9942619</pre> </div>                                                                                                                                                                                                                                                                                                                                                                                           |
| <b>Paare bilden und vergleichen in einem Streudiagramm</b> | <pre>pairs(~Balance + Age + Cards + Education + Income + Limit + Rating, data = Credit, pch = ".", col = "darkcyan")</pre> <div style="background-color: #ffffcc; padding: 10px;"> <pre>pairs(~Balance + Age + Cards + Education + Income + Limit + Rating,       data = Credit, pch = ".", col = "darkcyan")</pre> </div>                                                                                                                                                                                                                                            |

## Koeffizientenschätzung (qualitativ erklärende Variablen)

Female hier auf TRUE, somit 1

TRUE = 1

FALSE = 0

Darf nicht einfach auf Vorzeichen achten beim Resultat interpretieren. Muss auf die Codierung schauen.

Das Pluszeichen in  $B_0 + B_1 * X_1$  vor dem  $B_1 * X_1$  Term nicht per se eine Steigerung bedeutet. Es kommt dann auf das Vorzeichen des Koeffizienten  $B_1$  an.

Bei qualitativen Prädiktoren werden die Koeffizienten  $B_i$  als Abweichungen vom  $B_0$  betrachtet, wobei die Interpretation von der Kodierung des Faktors abhängt.

```
balance <- Credit$Balance
```

```
gender <- Credit$Gender == "Female"
```

```
round(summary(lm(balance ~ gender))$coef, digits = 5)
```

```
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

Female Unsurised
nur Koeffizienten interessant
```

```
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

Residuals interessieren an Namen
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 > 0.05
```

## Interaktionsterme lineares Modell

Istat:black ergibt Interaktionsterm zwischen Istat und black

Istat\*age (Prädiktoren&Interaktionsterm) ergibt Istat + age + Istat:age

Somit Wechselwirkung/Interaktion zwischen Istat und age (z.B. andere Steigung) (wegen \*). Wird Beziehung Istat und medv von age beeinflusst?

```
summary(lm(medv ~ lstat * age, data = Boston))
> summary(lm(medv ~ lstat * age, data = Boston))

call:
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
 Min 1Q Median 3Q Max
-15.806 -4.045 -1.333 2.085 27.552

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.0885359 1.4698355 24.553 < 2e-16 ***
lstat -1.3921168 0.1674555 -8.313 8.78e-16 ***
age -0.0007209 0.0198792 -0.036 0.9711
lstat:age 0.0041560 0.0018518 2.244 0.0252 *

signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared: 0.5557, Adjusted R-squared: 0.5531
F-statistic: 209.3 on 3 and 502 DF, p-value: < 2.2e-16
```

$$\hat{\beta}_0 = 36.10; \quad \hat{\beta}_1 = -1.39; \quad \hat{\beta}_2 = -0.0007; \quad \hat{\beta}_{12} = 0.004$$

Wir bekommen für das Modell

$$\text{medv} = 36.10 - 1.39 \cdot \text{lstat} - 0.00072 \cdot \text{age} + 0.0041 \cdot \text{lstat} \cdot \text{age}$$

## Unabhängig Variablen

Hier gibt es keinen Interaktionsterm zwischen Istat und age.  
Istat und age sind unabhängig von medv (gleiche Steigung).

lm(medv ~ lstat + age, data = Boston)

## 14 Glossar

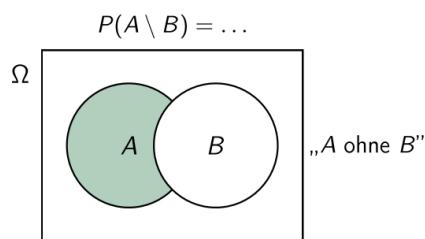
### Bezeichnungen

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

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

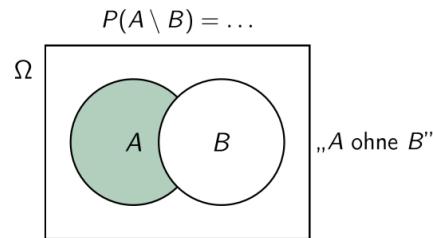
Prüfungsaufgabe:

Knobelaufgabe



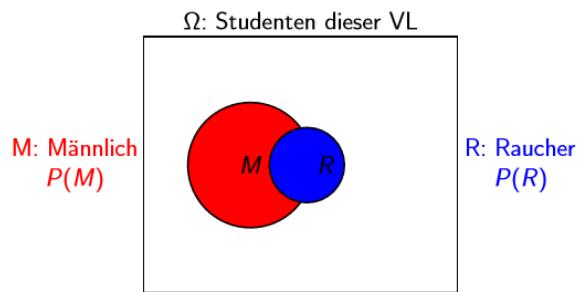
- 1 P(A) – P(B)
- 2 P(A) + P(B)
- 3 P(A) – P(A ∩ B)
- 4 P(A) + P(B) – P(A ∩ B)

Knobelaufgabe



- 1 P(A) – P(B)
- 2 P(A) + P(B)
- 3 P(A) – P(A ∩ B)
- 4 P(A) + P(B) – P(A ∩ B)

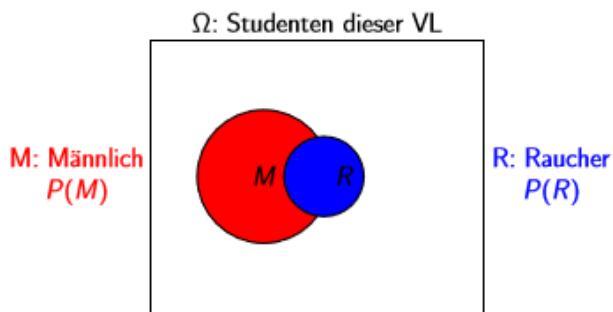
## Bedingte Wahrscheinlichkeit



Welche Aussagen sind korrekt?

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

## Bedingte Wahrscheinlichkeit



Welche Aussagen sind korrekt?

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

## 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:

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

- Gesucht:

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

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

## Beispiel

- 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  $\bar{X}_{16}$
- Annähernd verteilt wie:

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

Wahrscheinlichkeit (4 Punkte)

Sie haben die folgende Antwort gegeben:

Bei einem Zufallsexperiment werden ein roter (r) und ein blauer (b) Würfel gleichzeitig geworfen. Wir nehmen an, dass sie „fair“ sind, d. h. die Augenzahlen 1 bis 6 eines Würfels treten mit gleicher Wahrscheinlichkeit auf.

Beachten Sie: Falsche Antworten ergeben Punkteabzug.

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

| richtig                          | falsch                           |
|----------------------------------|----------------------------------|
| <input type="radio"/>            | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/>            |
| <input type="radio"/>            | <input type="radio"/>            |
| <input checked="" type="radio"/> | <input type="radio"/>            |
| <input type="radio"/>            | <input type="radio"/>            |
| <input checked="" type="radio"/> | <input type="radio"/>            |
| <input type="radio"/>            | <input type="radio"/>            |

"br" ist beispielsweise ein Elementarereignis

25

Die Wahrscheinlichkeit, dass das Produkt der Augenzahlen 7 ist, ist 0

Die Wahrscheinlichkeit, dass die Augensumme größer 11 ist, ist 35/36

Die Wahrscheinlichkeit, dass der blaue Würfel 5 ist, ist 1/6

| . | 1 | 2  | 3 | 4 | 5 | 6 |
|---|---|----|---|---|---|---|
| 1 | 1 | 2  | 3 | 4 | 5 | 6 |
| 2 | 2 | 4  | 6 | 8 |   |   |
| 3 | 3 | 6  | 9 |   |   |   |
| 4 | 4 | 8  |   |   |   |   |
| 5 | 5 | 10 |   |   |   |   |
| 6 | 6 | 12 |   |   |   |   |

$$= \overline{P(Summe > 11)} = \overline{P(Summe = 12)} = \frac{1}{36}$$

$$\overline{P(b=5)} = \frac{1/6}{36}$$