

Artificial Intelligence - Data Science
Ergänzendes Skriptum zur Statistik

Alexandra Posekany

WS 2021

Inhaltsverzeichnis

Einleitung	5
Grundbegriffe der Wahrscheinlichkeitsrechnung	6
Zufallsexperimente, Ereignisse und Wahrscheinlichkeit	6
Der frequentistische Wahrscheinlichkeitsbegriff, Laplace-Wahrscheinlichkeit	9
Elementare Eigenschaften von Wahrscheinlichkeiten	9
Zufallsvariablen und deren Verteilungen	10
Diskrete Verteilungen	10
Kontinuierliche Verteilungen	13
Momente einer Verteilung - Erwartungswert, Varianz, Schiefe	15
Bedingte Wahrscheinlichkeiten und Unabhängigkeit von Ereignissen	16
Kombinatorik	21
Wichtige Wahrscheinlichkeitsverteilungen und ihre Anwendungen	23
Diskrete Verteilungen	23
Gleichverteilung	23
Binomialverteilung und Bernoulli-Verteilung	25
Hypergeometrische Verteilung	27
Negativbinomialverteilung, geometrische Verteilung	29
Poissonverteilung	30
Übersicht über alle diskreten Verteilungen	32
Stetige Verteilungen	34
Gleichverteilung	36
Gauß'sche Normalverteilung und Standardnormalverteilung	38
Log-Normalverteilung	46
Student-t Verteilung	47
Gamma-Verteilung, Exponentialverteilung, χ^2 -Verteilung	48
Beta-Verteilung	50
Pareto Verteilung:	51
Übersicht über Stetige Verteilungen	52
Umsetzung mithilfe von Softwarepaketen im Überblick	54
Umsetzung in Maxima:	54
Umsetzung in R:	55
Beschreibende Statistik	56
Merkmale und Daten	56
Datenmatrix	57
Schätzung und Darstellung qualitativer Messungen	57
Eigenschaften von quantitativen Messungen	59
Erweiterte Inhalte - Weit aus mehr Lage- und Streuungsschätzer	66
Schätzung und Darstellung quantitativer Messungen	69
Zählvariablen	69
Stetige numerische Variablen	70
Quantil-Quantil-Plot (QQ-Plot)	73
Fallbeispiele	74
Inferenzstatistik	80
Konzept der Konfidenzintervalle	80
Konfidenzintervall und Prädiktionsintervall für den Mittelwert	82
Prädiktionsintervall und Konfidenzintervall für Proportionen (relative Häufigkeiten) = Vertrauensbereiche für Kategorienhäufigkeiten	88
Additions to Confidence versus prediction regions	92
Konzept der Hypothesentests	93

Was bei Hypothesentest zu beachten ist!	94
Das Konzept der Signifikanz und p-Werte	95
Fehler beim Hypothesentesten	96
Richtung der Hypothesentests - Formulierung der Fragestellung	100
Erweiterte Inhalte zu Verteilungen von Teststatistiken	103
Sampling distributions	104
Bayesian Hypothesis tests	105
Resampling Methods: Bootstrap distribution	105
Resampling Methods: Permutation Distribution	105
Wichtige Tests im Überblick	106
Parametrische und nicht-parametrische Tests im Vergleich	106
Überblick über die wichtigsten Tests nach Zielsetzung (Eigenschaft, die getestet wird)	107
Proportionentest	110
Proportionentest mit 1 Stichprobe	110
Estimating the Required sample size	113
Proportionentest zwischen 2 Stichproben	115
Mittelwertschätzung und Tests auf Lage	118
1 Stichproben t-Test	118
Choosing the appropriate distribution	119
2 Stichproben t-Test und Welch's t-Test	128
Nichtparametrischer Test zum Vergleich von Lage	129
Abhängige Stichproben t-Test	130
Schätzung und Test von Varianzen	142
1 Stichproben Varianz-Test	142
F-Test = 2 Stichproben Varianz-Test	145
Ergänzende Inhalte zu Hypothesentesten	146
Inference about two means: Independent samples	147
Inference about two means: Independent samples	148
Abhängigkeit von 2 oder mehr Variablen	151
Zusammenhänge zwischen 2 kategorialen Variablen	151
Zusammenfassung und Visualisierung	151
Visualisierungen	153
Inferenz durch χ^2 -Test für Homogenität und Unabhängigkeit	156
Zusammenhang zwischen einer metrischen Variable und kategorialen Variablen: Varianzanalyse	
(ANOVA = ANalysis Of VAriance)	160
Einfache Varianzanalyse (ANOVA)	160
Zweiweg Varianzanalyse (ANOVA)	163
ANOVA als Methode zur Modellselektion	170
ANOVA Models in R	172
Zusammenhänge zwischen 2 metrischen Variablen	174
Inferenz durch lineare Regression	174
Visualisierung und Korrelation als Maß für linearen Zusammenhang	177
Kleinste Quadrate Methode	183
Multiple lineare Regression	186
Modellselektion	187
Erweiterte Konzepte: Voraussetzungen für die Anpassung eines Regressionmodells	188
Schätzung der Gerade und Konfidenzbereiche	192
Modellzusammenfassung	194
Erweiterte Aspekte: Modellselektion	197
Beispiele zur Regression in R	199
Erweiterte Konzepte: Transformationen	202
Multiples Regressionsmodell - Beispiel Schweiz	213
Zusammenhänge zwischen einer kategorialen und metrischen Variablen - logistische Regression . .	217

Inferenz durch logistische Regression	217
Erweiterte Aspekte der Regression	222
Lineare Regression mit Nebenbedingungen: LASSO, Ridge-Regression und Elastic nets	222
Regression Trees und Forests	226

Einleitung

Sehr oft, besonders in deutschsprachige Ingenieursliteratur, wird Statistik in zwei Hauptbereiche nach ihrer Zielsetzung unterschieden, und zwar in

1. *deskriptive*, oder auch *beschreibende* Statistik, *Datenexploration*, als auch
2. *deduktive*, oder *schließende* Statistik, *Inferenzstatistik*.

Deskriptive Statistik beschäftigt sich mit dem Beschreiben, Zusammenfassen und Darstellen empirisch erhobener Datensätze, also dem Visualisieren von Messungen, dem Ermitteln von Schätzwerten und Maßzahlen für Lage und Streuung von Daten. Diese Aufgabe wird heutzutage ausschließlich von Software wie zum Beispiel *SPSS*, *Matlab*, *R*, *Mathematica* oder *Maxima* und zu großem Teil auch von Tabellenkalkulationsprogrammen wie *Excel* oder *OpenOffice* erledigt.

Das Erlernen dieser oft umfangreichen Softwarepakete ist nicht Ziel dieser Lehrveranstaltung, sondern wird von darauf aufbauenden vermittelt, während im Rahmen dieser Vorlesung und Übung grundlegende Begriffe der Wahrscheinlichkeitsrechnung und Statistik, welche unabdingbar für das Verständnis im Umgang mit Daten sind. Dementsprechend werden wir uns in Kapitel 1 nur mit den wichtigsten Charakteristiken begnügen, und nur so viele Beispiele explizit von Hand rechnen, um ihre Funktionsweise zu verstehen.

Deduktive Statistik hingegen beschäftigt sich mit der eigentlichen Aufgabe der Statistik. Mittels Modellen der Wahrscheinlichkeitstheorie wird der Prozess statistischen Messens, beispielsweise dem Ziehen unabhängiger Stichproben, in einen klaren Rahmen gefasst und analysiert: wie wahrscheinlich ist es, dass eine Stichprobe repräsentativ für ihre Grundgesamtheit ist? Wie extrem muss eine Stichprobe ausfallen, um Annahmen über die zugrunde liegende Verteilung in Frage zu stellen? Dies führt zu den Themen

- Schätzen von Verteilungen und deren Parametern (Bereichsschätzer, Konfidenzintervalle), und
- Testen von Hypothesen, d.h. Annahmen über zugrundeliegende Verteilungen eines Modells verifizieren/falsifizieren.

Beide Themen, Konfidenzintervalle und Tests, sind Kernfragen in Qualitätskontrolle und statistischer Prozess-Steuerung und Hauptziel dieses Skriptums.

Kapitel 1 erläutert die Grundbegriffe der Wahrscheinlichkeitsrechnung Zufallsexperiment, Zufallsvariablen, Wahrscheinlichkeit, Verteilungen und ihre wichtigsten Kenngrößen. An dieser Stelle werden wir nicht in größter Allgemeinheit arbeiten, sondern uns darauf beschränken, die im weiteren verwendete Notation zu erläutern.

Anschließend, in Kapitel 2, werden dann die wichtigsten Wahrscheinlichkeitsverteilungen und deren zugrundeliegenden Modelle erklärt. Die besprochenen Verteilungen sind nur ein kleiner Ausschnitt häufig verwendeteter Verteilungen, sollen aber die Basis für die darauf aufbauenden Hypothesentests und Konfidenzintervalle sein.

In Kapitel 3 besprechen wir graphische Darstellungsarten von empirisch erhobenen Daten und die populärsten Maßzahlen für eindimensionale, d.h. skalarwertige Daten.

Das Kapitel 4 beschäftigt sich mit Intervallschätzung und Hypothesentests, welche zwei Seiten der Betrachtung und Beantwortung von Fragen aus Daten sind. Die Möglichkeiten, bei Entscheidungen Fehler zu machen und diese zu quantifizieren, durch Signifikanz und Testschärfe werden dabei erläutert.

Das folgende Kapitel 5 beschäftigt sich mit Modellierung von Daten durch Regression und die Schätzung, Interpretation und Visualisierung solcher Modelle.

Caveat: dieses Skriptum ist als Begleitung und nicht Ersatz der Vorlesung gemeint. Es ist somit kurz gefasst und an vielen Stellen nicht ausreichend für ein Selbststudium.

Grundbegriffe der Wahrscheinlichkeitsrechnung

Zufallsexperimente, Ereignisse und Wahrscheinlichkeit

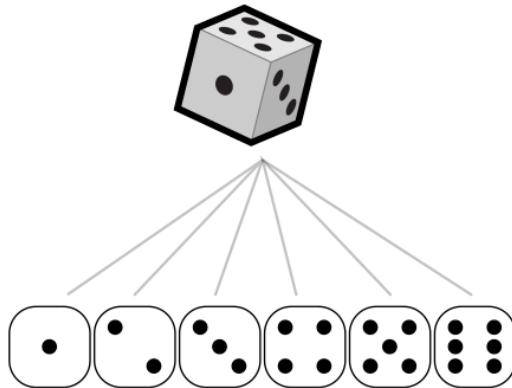
Wir bezeichnen ein nicht-deterministisches Experiment, dh. ein solches, bei dem wir nicht mit Gewissheit den Ausgang vorhersehen können als Zufallsexperiment Ω . Ein solches Experiment kann "echte" Zufälligkeit in sich tragen, wie z.B. nach gegenwärtigem Stand des Wissens der Zustand eines quantenmechanischen Systems, aber auch derart sein, dass es einfach nur zu schwierig (oder gar technisch nicht möglich) ist, alle Parameter hinreichend genau zu bestimmen, um die zukünftige Entwicklung eines Systems befriedigend vorherzusagen: klassische Beispiele dieses Typs sind

- mechanisch sehr empfindliche Ein- oder Wenigteilchensysteme, wie z.B. der Wurf eines Würfels oder das Werfen einer Roulette-Kugel,
- Dynamiken von Vielteilchensystemen, z.B. Thermodynamik von Gasen, thermisches Rauschen in einem elektrischen Schaltkreis, das Wechselspiel lokaler Wetterbedingungen,
- Systeme, deren zugrundelegenden Mechanismen nicht im Detail verstanden wird, wie z.B. die Handelsentscheidungen, die den Kurs eines an der Börse gehandelten Derivats bestimmen.

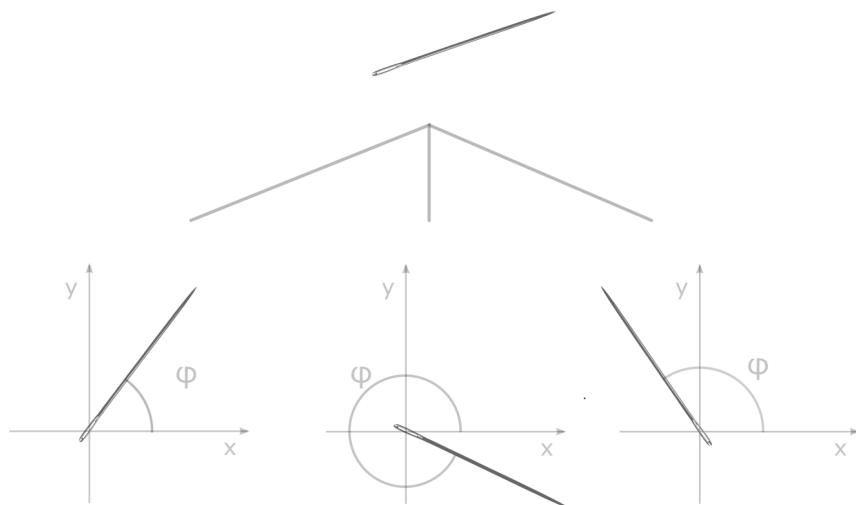
Der mathematische Formalismus von Zufallsexperimenten unterscheidet das Zufallsexperiment Ω , dessen mögliche Resultate ω , und die tatsächliche auftretenden Ereignisse A .

Ein **Zufallsexperiment** Ω besitzt im allgemeinen mehrere möglichen zukünftige Realitäten bzw. **Resultate**, die wir mit Kleinbuchstaben $\omega_1, \omega_2, \omega_3, \dots$ usw. bezeichnen, Ω selbst werden wir formal als logische Zusammenfassung bzw. Menge all dieser möglichen Realitäten $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$ schreiben. Je nach Beispiel kann Ω aus endlich oder unendlich, diskret oder kontinuierlich vielen Realitäten ω bestehen, welche auch **Elementareignisse** genannt werden. Zusammenfassungen von Realitäten $A \subseteq \Omega$ nennen wir **Ereignisse**, und bezeichnen wir im weiteren mit Großbuchstaben A, B, C, usw. Wenn das Ergebnis ω eines Zufallsexperiments dem Ereignis A zugehörig ist, dh. $\omega \in A$, dann sagen wir, dass *das Ereignis A ein- bzw. auftritt*.

Wenn zum Beispiel das Zufallsexperiment Ω nur das Resultat eines Würfelwurfs beschreiben soll, dann sind die Elementareignisse ω stellvertretend für die resultierende Augenzahl. Die Grundgesamtheit Ω besteht somit nur aus sechs verschiedenen Realitäten, die wir formal durch $\Omega = \{1, 2, 3, 4, 5, 6\}$ parametrisieren.



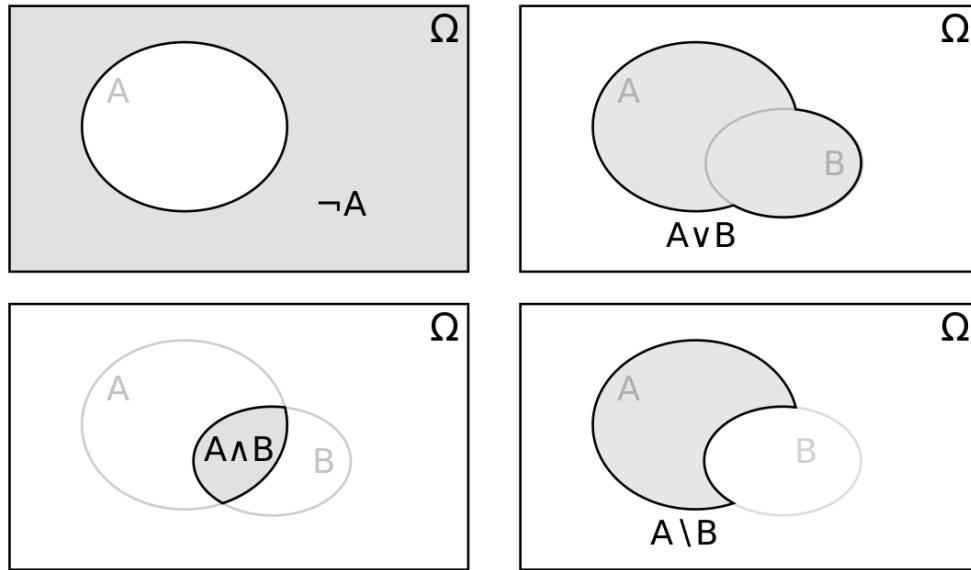
Wenn wir andererseits beispielsweise die Orientierung φ einer auf den Boden geworfenen Nadel bezüglich eines von uns festgelegten Bezugssystems beschreiben wollen, dann kann potentiell jeder Winkel zwischen 0 und $2 \cdot \pi$ Radian auftreten, also $\varphi \in [0, 2\pi)$.



Logische Terme von Aussagen, die Ereignisse beschreiben Im Zusammenhang mit Zufallsexperimenten werden wir zwischen logischen Aussagen über dessen Ausgang und den von ihnen beschriebenen Ereignissen nicht unterscheiden. Gelegentlich, wenn angebracht, werden wir solch logische Aussagen als verschachtelten Term von zwei oder mehr Ereignissen anschreiben. Hierbei halten wir uns an die übliche Notation. Zur graphischen Veranschaulichung nutzen wir Venn-Diagramme und zur sprachlichen Veranschaulichung ein konkretes Beispiel aus der Medizin.

Das Ereignis A beschreibt, dass ein Medikament Wirkung zeigt. Das Ereignis B, dass durch die Einnahme der Medikamente keine Nebenwirkung auftritt. Dann unterscheidet man 4 konkrete Zusammenhänge mit und zwischen diesen Ereignissen.

- Die **Negation = Verneinung = Umkehrung** ist die zu A komplementäre Aussage bzw. dessen komplementäres Ereignis, die wir mit $\neg A$ bezeichnen. In unserem Beispiel ist $\neg A$ “das Medikament zeigt keine Wirkung” und $\neg B$ “durch Einnahme des Medikaments tritt mindestens 1 Nebenwirkung auf”.
- Die **Disjunktion = “Oder”-Relation** $A \vee B$ beschreibt, dass Ereignis A, Ereignis B oder beide zugleich eintreten können. In unserem Beispiel wäre $A \vee B$ “Das Medikament zeigt Wirkung oder durch die Einnahme der Medikamente tritt keine Nebenwirkung ein”.
- Die **Konjunktion = “Und”-Relation** $A \wedge B$ beschreibt, dass Ereignis A und Ereignis B beide zugleich eintreten müssen. In unserem Beispiel wäre $A \wedge B$ “Das Medikament zeigt Wirkung und durch die Einnahme der Medikamente tritt keine Nebenwirkung ein”.
- Die **Differenz** $A \setminus B$ beschreibt, dass Ereignis A eintritt, während Ereignis B nicht gleichzeitig eintreten darf. In unserem Beispiel wäre $A \setminus B$ “Das Medikament zeigt Wirkung, aber durch die Einnahme der Medikamente tritt mindestens Nebenwirkung ein”.



Der frequentistische Wahrscheinlichkeitsbegriff, Laplace-Wahrscheinlichkeit

{Laplace Wahrscheinlichkeit}

Wir bestimmen die Menge aller günstigen Ereignisse A und vergleichen diese mit der Menge aller möglichen Ereignisse Ω , indem wir die Größen dieser Mengen durcheinander dividieren

$$\mathbb{P}[G] = \frac{|G|}{|\Omega|}$$

Diese Wahrscheinlichkeit für das Eintreten des günstigen Ereignisses G heißt **Laplace Wahrscheinlichkeit**.

Dabei ist $|G|$ die Anzahl der günstigen Ereignisse, wenn wir konkrete von einander getrennte Ereigniskategorien haben, wie etwa beim Würfelwurf, oder die Länge des Intervalls, wenn wir ganze Bereiche von möglichen Messausgängen haben, zwischen denen sich keine Grenzen ziehen lassen.

In unserem obigen Beispiel ist die Wahrscheinlichkeit, einen 4er zu würfeln

$$\mathbb{P}[\text{'4er würfeln'}] = \frac{1}{6} = 0.17$$

und die Wahrscheinlichkeit, eine gerade Zahl zu würfeln

$$\mathbb{P}[\text{'eine gerade Zahl würfeln'}] = \frac{3}{6} = 0.5$$

Elementare Eigenschaften von Wahrscheinlichkeiten

Für Wahrscheinlichkeiten von Ereignissen gelten grundsätzlich folgende Eigenschaften:

- die Wahrscheinlichkeit eines Ereignisses A ist eine Zahl zwischen 0 und 1:

$$0 \leq \mathbb{P}[A] \leq 1,$$

- die Wahrscheinlichkeit für das komplementäre Ereignis $\neg A$ ist

$$\mathbb{P}[\neg A] = 1 - \mathbb{P}[A],$$

welche wir sinngemäß komplementäre Wahrscheinlichkeit von A nennen,

- wenn zwei Ereignisse A, B disjunkt sind, dh. keinen gemeinsamen Schnitt besitzen, kurz $A \wedge B = \emptyset$, dann gilt

$$\mathbb{P}[A \vee B] = \mathbb{P}[A] + \mathbb{P}[B],$$

- wenn die Ereignisse A und B nicht disjunkt sind, dann gilt die Formel

$$\mathbb{P}[A \vee B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \wedge B].$$

Zufallsvariablen und deren Verteilungen

Eine **Zufallsvariable** bzw. **Zufallsgröße** ist ein numerischer Wert, der vom Ausgang ω eines Zufallsexperiments abhängt. Gegebenfalls schreiben wir eine Zufallsgröße auch als Funktion $X : \Omega \rightarrow \mathbb{R}$, d.h. je nach Ausgang Ω des Experiments nimmt die Zufallsgröße X einen (möglicherweise anderen) Wert $x = X(\omega)$ an.

Wenn X nur Werte einer aus einer endlichen oder abzählbaren Liste $\{a_1, a_2, a_3, \dots\} \subset \mathbb{R}$ annimmt, spricht man von einer **diskreten Zufallsgröße**.

Anderenfalls, wenn der Wertebereich von X einen gesamten kontinuierlichen Bereich der reellen Zahlen $[a, b] \subseteq \mathbb{R}$ ausschöpft, von einer **kontinuierlichen Zufallsgröße**.

Selbstverständlich können auch mehrdimensionale Objekte wie z.B. Vektoren vom Ausgang eines Zufallsexperiments abhängen, wir werden allerdings im Rahmen dieser Vorlesung nicht darauf eingehen.

Diskrete Verteilungen

Mathematisch entspricht die **diskrete Wahrscheinlichkeitsverteilung** den **relativen Häufigkeiten** jeder Teilkategorie, wenn man unendlich oft beobachten würde, also $\mathbb{P}[X = x] = \lim_{n \rightarrow \infty} \frac{h_n(x)}{n}$.

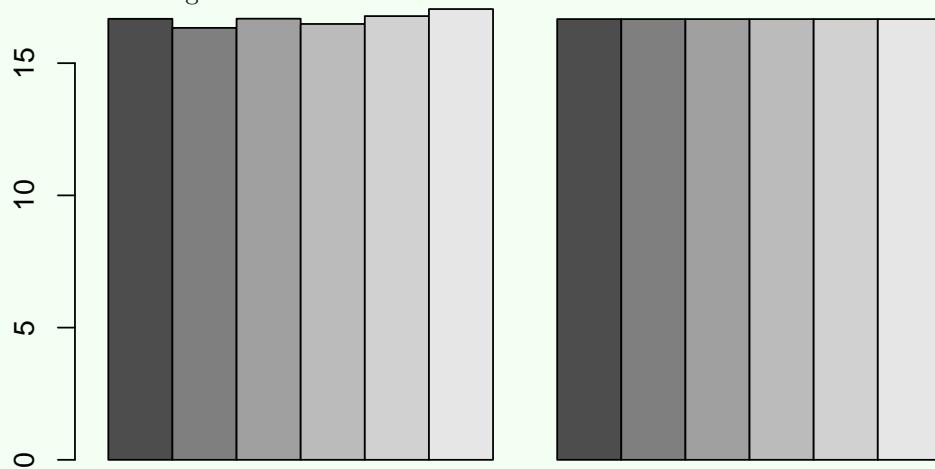
Im konkreten Beispiel unseres obigen Würfelwurfs mit einem 6-seitigen Würfel bedeutet das Folgendes: Aus der Grundgesamtheit $\Omega = \{1, 2, 3, 4, 5, 6\}$ können die Augenzahlen 1 bis 6 jeweils als Einzelereignis ermittelt werden. Wir führen mehrere virtuelle Würfelwürfe durch:

absolute H.	relative H. (%)	Wahrscheinlichkeit (%)
1	9	18
2	8	16
3	7	14
4	9	18
5	9	18
6	8	16

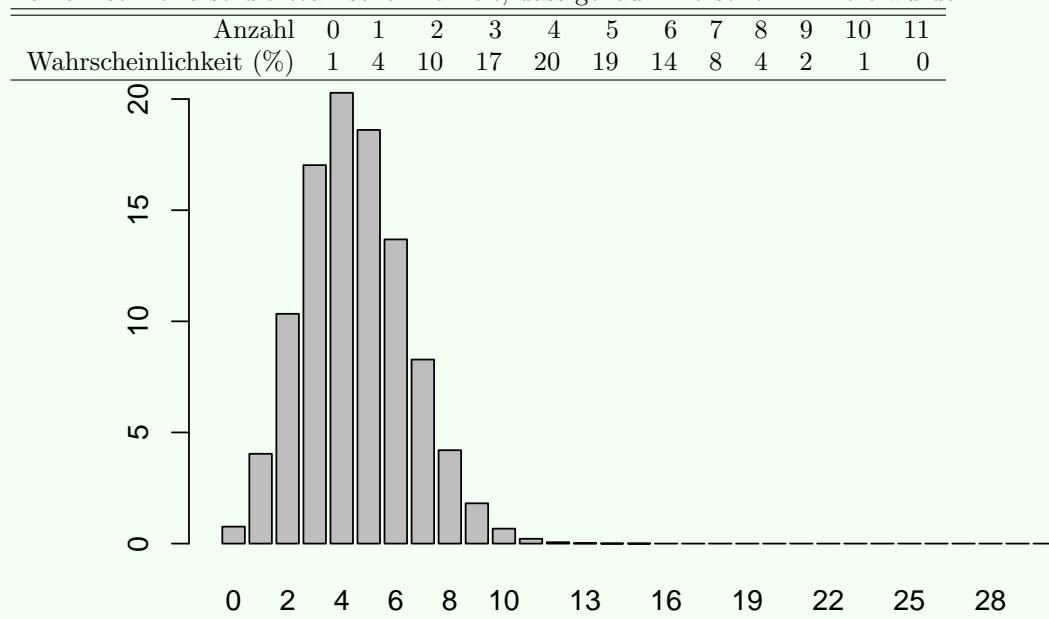
Bei nur wenigen Beobachtungen sind die relativen Häufigkeiten sehr weit weg von den theoretischen Wahrscheinlichkeiten. Wenn wir mehr Würfelwürfe durchführen, verändert sich dieses Bild.

absolute H.	relative H. (%)	Wahrscheinlichkeit (%)
1	8338	17
2	8167	16
3	8341	17
4	8241	16
5	8390	17
6	8523	17

Graphisch dargestellt bedeutet das Folgendes, wobei rechts die relativen Häufigkeiten und links die Wahrscheinlichkeiten dargestellt sind:



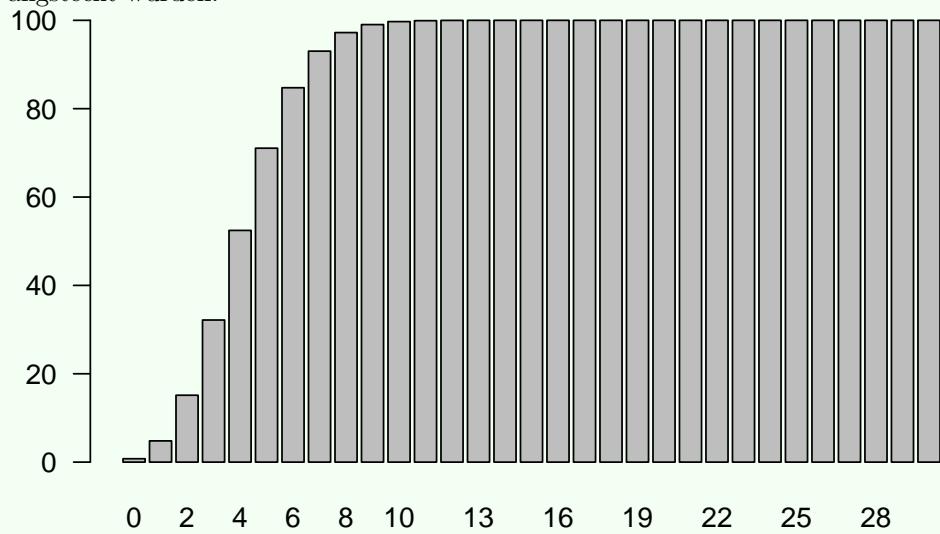
Für ein anderes Beispiel, das die Wahrscheinlichkeit an Covid-19 zu erkranken in einem Cluster mit 15% Ansteckungswahrscheinlichkeit in einem Studienjahrgang mit 30 Studierenden darstellt. Jeder Balken ist hier also die Wahrscheinlichkeit, dass genau k Personen infiziert wurden.



Diese Darstellung ist die **Wahrscheinlichkeitsfunktion** der diskreten Verteilung, da durch die Höhe des Balkens die Wahrscheinlichkeit jeder Kategorie dargestellt wird.

Die aufsummierten Wahrscheinlichkeiten bis zu einer bestimmten Kategorie sind die Darstellung der Quantile, welche nur für ordinale Kategorien und Zählvariablen sinnvoll sind. Die Darstellung dieser aufsummierten Wahrscheinlichkeiten erfolgt über Balken oder eine Treppenfunktion und heißt **kumulative Verteilungsfunktion**.

Hier für die oben dargestellten Ansteckungszahlen die aufsummierte Wahrscheinlichkeit, dass bis k Personen angesteckt wurden.



Kontinuierliche Verteilungen

Bei kontinuierlichen Verteilungen hat ein einzelner Zahlenwert stets Wahrscheinlichkeit 0, da sonst ihre überabzählbare Summe keine endliche Wahrscheinlichkeit von höchstens 1 ergeben könnte.

Die Verteilung einer kontinuierlichen Zufallsgröße X kann mittels ihrer **Wahrscheinlichkeitsdichte**

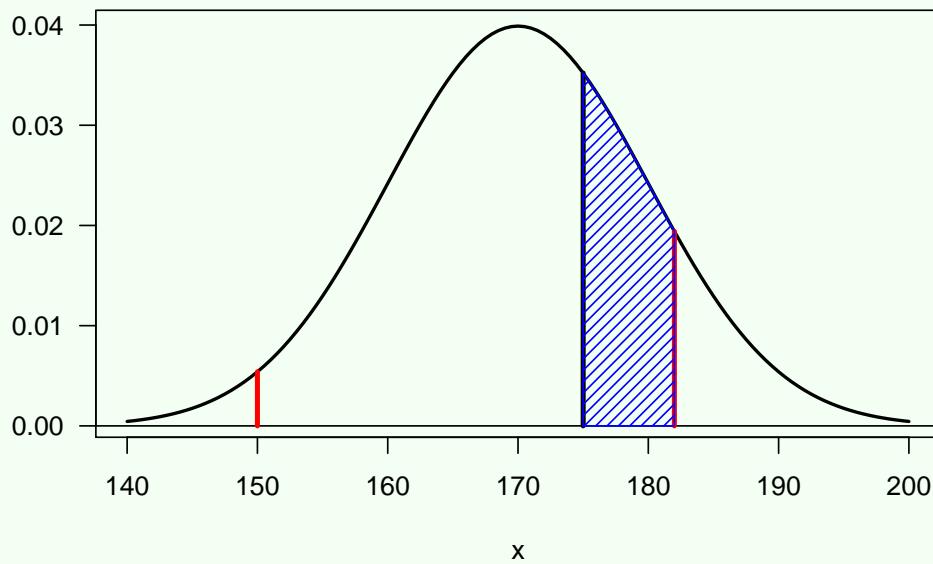
$$p(x) = \lim_{\Delta x \rightarrow 0} \frac{\mathbb{P}[x \leq X \leq x + \Delta x]}{\Delta x}$$

beschrieben werden. Eine **konkrete Wahrscheinlichkeit für einen Intervallbereich** errechnet sich als die **Fläche unter der Kurve der Dichtefunktion**, also das Integral der Dichtefunktion

$$\mathbb{P}[X \in [a, b]] = \int_a^b p(x)dx$$

Die Körpergrößen von Menschen sind erfahrungsgemäß normalverteilt. Als mittlere Körpergröße im westlichen Bereich wird ein Wert von etwa 170 cm angenommen.

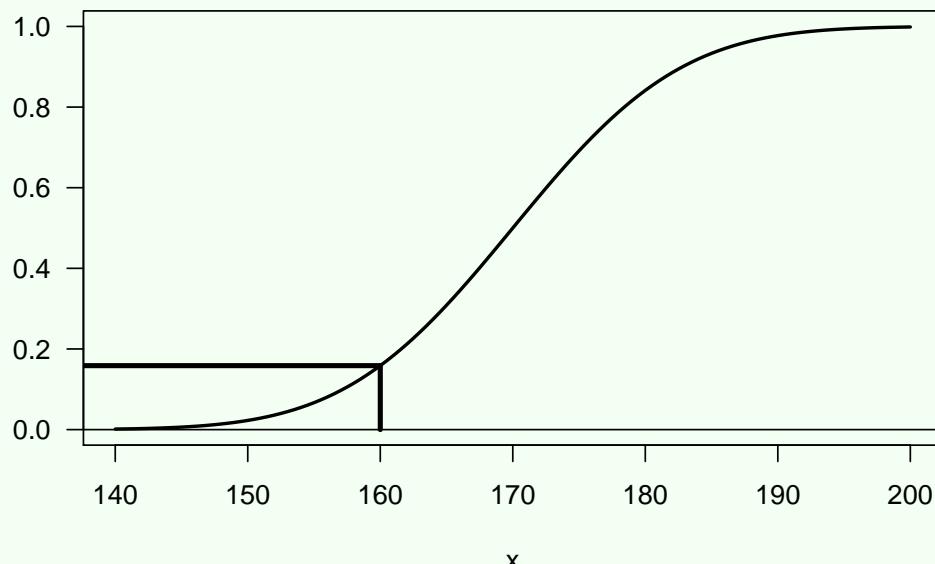
Dichtefunktion



Die konkrete Körpergröße von 150 cm hat eine Wahrscheinlichkeit von 0 als solche beobachtet zu werden. Aber eine Körpergröße zwischen 175 cm und 182 cm hat sehr wohl eine konkrete Wahrscheinlichkeit, nämlich $\mathbb{P}[175 \leq X \leq 182] = \int_{175}^{182} p(x)dx = 0.1935 = 19.35\%$.

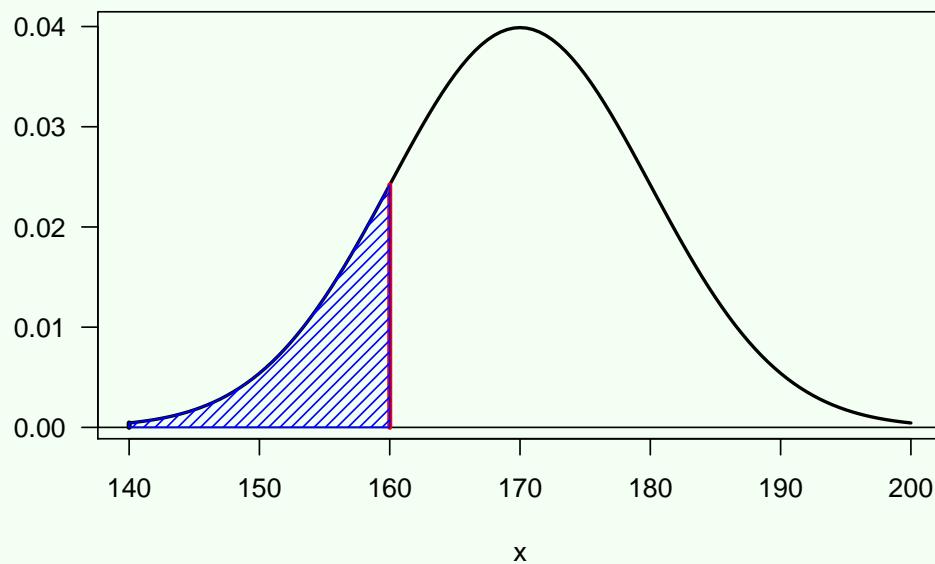
Da die Fläche unter der Kurve, also die Stammfunktion, der Dichtefunktion für die Ermittlung der Wahrscheinlichkeit so relevant ist, hat diese Funktion eine eigene Bezeichnung, die **kumulative Verteilungsfunktion**. In ihrer Darstellung wird an der y-Achse die Wahrscheinlichkeit und an der x-Achse die konkreten Werte, bis zu denen diese Wahrscheinlichkeit ermittelt wird, aufgetragen.

Verteilungsfunktion



Hier wird an der Stelle 160 cm sehr wohl eine Wahrscheinlichkeit dargestellt, nämlich die Wahrscheinlichkeit, eine Körpergröße von höchstens 160 cm zu haben, $\mathbb{P}[-\infty \leq X \leq 160] = \int_{-\infty}^{160} p(x)dx = 0.1587 = 15.87\%$. Diese Wahrscheinlichkeit entspricht in der obigen Darstellung jener Fläche unter der Kurve:

Dichtefunktion



Momente einer Verteilung - Erwartungswert, Varianz, Schiefe

Als **n-tes Moment** bezeichnet man in der Statistik und Wahrscheinlichkeitsrechnung die Summe der n-ten Potenz der Datenwerte gewichtet mit ihrer Wahrscheinlichkeit, mathematisch

$$\mathbb{E}[X^n] = \begin{cases} \int f(x) \cdot x^n dx & \text{mit der Dichtefunktion } f(x) \\ & \text{der stetigen Zufallsvariable } X \text{ bzw.} \\ \sum_{i=0}^{\infty} p(x_i) \cdot x_i^n & \text{mit der Wahrscheinlichkeitsverteilung } p(x_i) \\ & \text{der diskreten Zufallsvariable } X. \end{cases}$$

Das bekannteste Moment einer Verteilung ist ihr 1. Moment, der **Erwartungswert** $\mathbb{E}[X] = \mu$. Sein empirischer Schätzwert ist der *arithmetische Mittelwert*.

Der Erwartungswert bzw. der Mittelwertes einer Verteilung unterscheidet sich daher in seiner Berechnung ja nachdem, ob diese stetig oder diskret ist. Im Falle einer diskreten Wahrscheinlichkeit haben die einzelnen Ausgänge, also die Kategorien, tatsächliche Häufigkeiten und Wahrscheinlichkeiten, daher kann der Kategorienvwert mit dieser Häufigkeit multipliziert werden und die dadurch entstehenden gewichteten Teilwerte werden aufaddiert.

Ein typisches Beispiel dafür ist, wenn ein Messgerät nur ganzzahlige Ergebnisse zurückliefert, was häufig bei einer Hauswaage oder der Altersermittlung der Falle ist. Bei der folgenden Altersverteilung von Studierenden eines Bachelorstudiengangs

Alter	Häufigkeit
18	5
19	17
20	9
21	6
22	3
27	1

ermittelt man den arithmetischen Mittelwert durch Gewichtung der Alterswerte durch die entsprechenden Häufigkeiten ihres Auftretens, also $18 \cdot \frac{5}{41} + 19 \cdot \frac{17}{41} + 20 \cdot \frac{9}{41} + 21 \cdot \frac{6}{41} + 22 \cdot \frac{3}{41} + 27 \cdot \frac{1}{41} = 20$ Jahre.

Im Falle unterschiedlicher Messungen, welche nicht als Kategorien zusammengefasst bzw. aufgefasst werden, entspricht die Berechnung des Mittelwerts einer Stichprobe genau der Berechnung des 1. Moments, wenn man als Verteilung eine Gleichverteilung über alle Beobachtungen annehmen würde. Das bedeutet, dass als Laplace-Wahrscheinlichkeit jedes beobachteten Wertes einer Stichprobe der Größe n die relative Häufigkeit $\frac{1}{n}$ verwendet wird. Dadurch entsteht die Formel des **arithmetischen Mittelwerts**

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

Der arithmetische Mittelwert von Messungen von Einwaagegewichten einer Probe mit folgenden Messungen in Milligramm:
4.96, 4.98, 4.92, 5.03, 4.74, 4.86, 5.07, 5.1, 4.89, 5.17, 4.95, 5.02
ergibt sich als 59.71 dividiert durch 12 = 4.98 mg.

Als **n-tes zentrales Moment** bezeichnet man in der Statistik und Wahrscheinlichkeitsrechnung die Summe der n-ten Potenz der Differenz der Datenwerte vom Erwartungswert $\mathbb{E}[X] = \mu$ gewichtet mit ihrer Wahrscheinlichkeit, mathematisch

$$\mathbb{E}[(X - \mu)^n] = \begin{cases} \int f(x) \cdot (x - \mu)^n dx & \text{mit der Dichtefunktion } f(x) \\ & \text{der stetigen Zufallsvariable } X \text{ bzw.} \\ \sum_{i=0}^{\infty} p(x_i) \cdot (x_i - \mu)^n & \text{mit der Wahrscheinlichkeitsverteilung } p(x_i) \\ & \text{der diskreten Zufallsvariable } X. \end{cases}$$

Das bekannteste zentrale Moment einer Verteilung ist das 2. zentrale Moment, die **Varianz** $\mathbb{E}[(X - \mu)^2]$. Ihr Schätzwert ist die **Stichprobenvarianz**. Die Varianz quadriert die Abweichungen vom Erwartungswert, wodurch das Resultat nicht mehr in seiner Größenordnung mit dem Erwartungswert vergleichbar ist. Um diesen Mangel zu beheben, zieht man die Wurzel aus der Varianz, was die **Standardabweichung** ergibt, welche wieder dieselbe Einheit und Skala wie der Erwartungswert hat.

Die wichtigsten empirischen Momente

1. Moment	Lage	Erwartungswert	\bar{x}	$= \frac{1}{n} \sum_{i=1}^n x_i$
2. zentrales Moment	Streuung	Varianz	s_n^2	$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
3. zentrales Moment	Symmetrie	Schiefe	$skew$	$= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}}$
4. zentrales Moment	Rändergewicht	(Exzess) Kurtosis	$kurt$	$= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$

Bedingte Wahrscheinlichkeiten und Unabhängigkeit von Ereignissen

Bedingte Wahrscheinlichkeiten

Die **Wahrscheinlichkeit eines Ereignisses B bedingt zu einem anderen Ereignis A** ist die Auftrittshäufigkeit von B unter jenen Messungen, in denen A aufgetreten ist:

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \wedge B]}{\mathbb{P}[A]}$$

In anderen Worten $P(B|A)$ ist der Prozentsatz an A-Ergebnissen, bei dem auch B eingetreten ist.

Statistische Unabhängigkeit

Man nennt B von A statistisch bzw. stochastisch unabhängig, oder kurz unabhängig, wenn die Auftrittswahrscheinlichkeit von B nicht vom Eintreten von A beeinflusst wird, dh.

$$\mathbb{P}[B|A] = \mathbb{P}[B|\neg A] = \mathbb{P}[B]$$

Betrachten wir das mit einem konkreten Beispiel. Ereignis A ist "Die Straße ist naß.", des Gegenereignis $\neg A$ ist daher "Die Straße ist nicht naß.". Ereignis B ist "Es regnet.", des Gegenereignis $\neg B$ ist daher "Es regnet nicht.". Dann bedeutet das bedingte Ereignis

$$\overbrace{A}^{\text{Die Straße ist nass}} \quad | \quad \overbrace{B}^{\text{es regnet}}$$

"Die Straße ist nass, wenn es regnet." Die Wahrscheinlichkeit $\mathbb{P}[A|B]$, dass die Straße nass ist, wenn es regnet, ist außerhalb von überdachten oder überbrückten Arealen so gut wie 100%. In die umgekehrte Richtung hat das Ereignis

$$\overbrace{B}^{\text{es regnet}} \quad | \quad \overbrace{A}^{\text{die Straße nass ist}},$$

dass es regnet, wenn die Straße nass ist, eine Wahrscheinlichkeit $\mathbb{P}[B|A]$ von deutlich unter 100%, da ein mittlerweile vergangener Regen Ursache für eine nasse Straße sein kann, wie auch Straßenreinigung, Rohrbrüche, Überflutungen, wenn Wasser über die Ufer tritt etc.

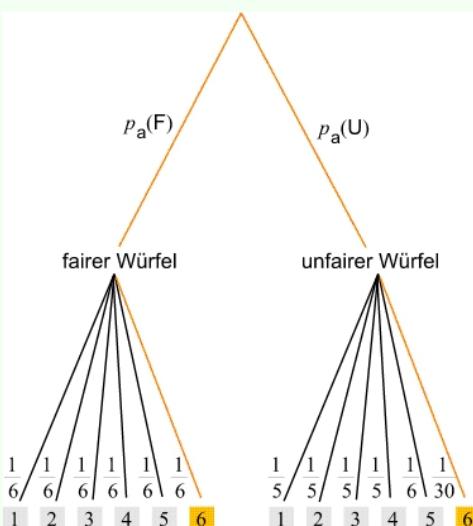
Eine äquivalente Formulierung für die Unabhängigkeit von B zu A ist

$$\mathbb{P}[A \wedge B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$$

Aus letzterer Charakterisierung folgt: Statistische Unabhängigkeit ist ein symmetrischer Begriff, dh. wenn B von A statistisch unabhängig ist, so ist ebenso A von B statistisch unabhängig. Dies bedeutet, dass Ereignisse unabhängig sind, wenn ihre Eintrittswahrscheinlichkeiten miteinander multipliziert ein Eintrittswahrscheinlichkeit des Geschehens beider Ereignisse zum selben Zeitpunkt ist.

Betrachten wir das mit einem konkreten Beispiel. Im Spiel "Die Siedler von Catan" wird grundsätzlich mit 2 sechsseitigen Würfel gewürfelt und dann das Ergebnis der beiden Würfelwürfe addiert. Dabei hat wie wir uns bereits bei den Laplacewahrscheinlichkeiten überlegt haben auf einem gleichmäßigen und nichtgezinkten fairen Würfel jeder Augenzahl dieselbe Wahrscheinlichkeit gewürfelt zu werden. Die Wahrscheinlichkeit mit dem 1. Würfel W1 einen 4er mit dem 2. Würfel W2 einen 3er zu würfeln, wäre $\mathbb{P}[W1 = "4" \wedge W2 = "3"]$. Das Werfen des 1. Würfels wirkt sich weder auf die Augenzahl des zweiten Würfels aus, noch umgekehrt, also sind diese beiden Ereignisse stochastisch unabhängig. Daher ist die Wahrscheinlichkeit $\mathbb{P}[W1 = "4" \wedge W2 = "3"]$ dasselbe wie das Produkt der beiden einzelnen Wahrscheinlichkeiten $\mathbb{P}[W1 = "4"]$ und $\mathbb{P}[W2 = "3"]$, also $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Würde aus einem bestimmten Grund wie der Beschaffenheit des Untergrundes oder, dass die Würfel gezinkt sind und etwa durch Magnete ihre gegenseitige Lage beeinflussen, würde die Wahrscheinlichkeit für das Ereignis $W1 = "4" \wedge W2 = "3"$ einen anderen Wert als $\frac{1}{36}$ betragen.

Diese bedingten Wahrscheinlichkeiten werden mithilfe eines **Baumdiagramms** dargestellt in der zuerst das erste Ereignis ohne Vorbedingung in der obersten Hierarchieebene eingetragen wird. Dann folgt die Bedingung durch diese Ereignis bzw. die Einzelereignisse dieser Ebene und ihre Auswirkung auf die nächste abhängige Ebene.



Hier wird zunächst unterschieden, ob es sich um einen fairen Würfel oder einen unfairen Würfel handelt. Dies ist die unabhängige Hierarchieebene in dieser Darstellung, also kein "Fairer Würfel"/"Augenzahl" sondern nur "Fairer Würfel". In der nächsten Ebene wird dann bedingt darunter, ob der Würfel fair ist wie auf der linken Seite oder unfair wie auf der rechten Seite dargestellt, die Wahrscheinlichkeiten, die jeweilige Augenzahl zwischen 1 und 6 zu würfeln, illustriert. Beim fairen Würfel ist dies jeweils gleich $\frac{1}{6}$, beim unfairen Würfel ist die Wahrscheinlichkeit, einen 6er zu würfeln mit $\frac{1}{30}$ deutlich kleiner und dafür die Wahrscheinlichkeit 1er bis 4er zu würfeln mit $\frac{1}{5}$ etwas größer. Wer 6er würfeln möchte, wird also deutlich benachteiligt.

Bedingte Häufigkeiten bzw. Wahrscheinlichkeiten werden aus realen Daten durch das Erstellen von **Kontingenztafeln**, welche die Auftrittshäufigkeiten oder Wahrscheinlichkeiten aller logischer Kombinationen von A und B beschreiben, bestimmt. Dabei ist zu beachten, dass die außenstehenden Häufigkeiten, die sogenannten **Randhäufigkeiten** die Zeilen- bzw. Spaltensumme der jeweiligen gemeinsamen Häufigkeiten sind. Letztendlich summieren sich diese selbst spalten- bzw. zeilenweise auf die Gesamtanzahl n, was bei händischem Berechnen öfters als Rechenprobe eingesetzt wird.

$Y \setminus X$	a_1	\dots	a_J	Zeilenhäufigkeiten
b_1	$n_{1,1}$	\dots	$n_{1,J}$	$n_{1,..}$
\vdots	\vdots	\ddots	\vdots	\vdots
b_I	$n_{I,1}$	\dots	$n_{I,J}$	$n_{I,..}$
Spaltenhäufigkeiten	$n_{.,1}$	\dots	$n_{.,J}$	n

Dabei stehen die Wahrscheinlichkeiten für das Eintreten zweier Ereignisse gleichzeitig, also $\mathbb{P}[A \wedge B]$ in den jeweiligen Zellen der Tabelle, während an den Rändern die Randhäufigkeiten $\mathbb{P}[A]$ und $\mathbb{P}[B]$ zu finden sind. Ob Einzelereignisse also unabhängig sind oder nicht, erkennt man durch Bilden des Produkts der Randhäufigkeiten und Vergleich mit der Häufigkeit des Eintretens beider Ereignisse. Sind diese Werte ident (bei realen Beobachtungen annähernd gleich), spricht man von unabhängigen Ereignissen, anderenfalls von abhängigen Ereignissen. Dieses Prinzip ist die Basis des χ^2 -Tests, der genau diese Unabhängigkeit von in Kontingenztafeln erfassten Ereignissen überprüft.

Den Zusammenhang zwischen bedingten Ereignissen und Randhäufigkeiten bilden zwei wichtige Konzepte:

Pfadregeln "UND" und "ODER"

Bei unabhängigen Ereignissen A und B gilt grundsätzlich, dass das Ereignis "A und B tritt ein" das Produkt der beiden Einzelereignisse ist.

$$\mathbb{P}[A \wedge B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$$

Da Wahrscheinlichkeiten Zahlen zwischen 0 und 1 sind, ist das Produkt zweier solcher Zahlen stets kleiner (oder gleich) den beteiligten Faktoren. Die Wahrscheinlichkeit, dass zwei Ereignisse gleichzeitig eintreten, ist also kleiner (oder gleich) den Einzelwahrscheinlichkeiten der Ereignisse.

Bei unabhängigen Ereignissen A und B gilt grundsätzlich, dass das Ereignis "A oder B tritt ein" die Summe der beiden Einzelereignisse ist.

$$\mathbb{P}[A \vee B] = \mathbb{P}[A] + \mathbb{P}[B]$$

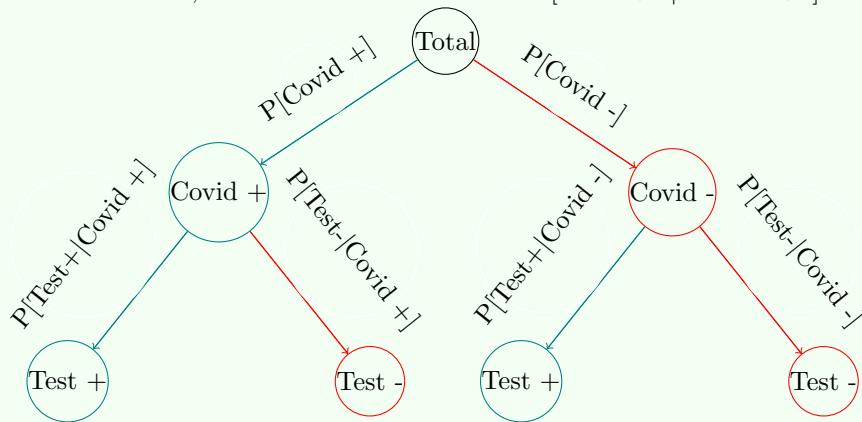
Im Unterschied zum UND " Fall, ist beim ODER " Fall die Wahrscheinlichkeit, dass entweder das eine oder das andere Ereignis eintritt, größer (oder gleich) den Einzelwahrscheinlichkeiten der Ereignisse.

Satz von Bayes

Grundsätzlich gibt es eine klare Struktur, welche das Ereignis ist, das die Vorbedingung vorgibt und welches Ereignis davon abhängig ist, etwa Vorbedingung B und abhängig davon A|B. Nicht immer ist die Richtung der Bedingung eindeutig, manchmal möchte man auch Vorbedingung und bedingtes Ereignis vertauschen, also A und B|A. Genau diese Umkehrung der bedingten Wahrscheinlichkeiten wird durch den **Satz von Bayes** beschrieben.

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A|B] \cdot \mathbb{P}[B]}{\mathbb{P}[A]}$$

Als Beispiel für diese Konzepte betrachten wir ein Beispiel zur Durchführung von PCR-Tests zur Feststellung von Antikörpern gegen SARS-CoV2 als Nachweis von Covid19. Aus internationalen Studien kann innerhalb der Bevölkerung im Mai 2020 eine Inzidenz von 0.17% angenommen werden, als $\mathbb{P}[\text{"Covid+"}] = 0.0017$. Der Test ist so konstruiert, dass dieser eine hohe Sensitivität von 92.8% hat, also von allen tatsächlich mit SARS-CoV2 Infizierten, wird bei 92.8% der Personen ein positives Testergebnis zurückgegeben. Das bedeutet, dass die Wahrscheinlichkeit $\mathbb{P}[\text{"Test +"} | \text{"Covid+"}] = 0.928$. Bei 12% aller Nichtinfizierten Personen wird allerdings ebenfalls ein positives Testergebnis zurückgeliefert. Das bedeutet, dass die Wahrscheinlichkeit $\mathbb{P}[\text{"Test +"} | \text{"\neg Covid+"}] = 0.12$.



Die „UND“-Regel hier umgesetzt, bedeutet, dass wir die Wahrscheinlichkeit, an Covid erkrankt zu sein UND ein positives Testergebnis zu erhalten, durch Multiplikation der beiden Wahrscheinlichkeiten der Hierarchieebene erhalten, also

$$\mathbb{P}[\text{"Test +"} | \text{"Covid+"}] \cdot \mathbb{P}[\text{"Covid+"}] = \mathbb{P}[\text{"Test +"} \wedge \text{"Covid+"}] = 0.0017 \cdot 0.928 = 0.0015776.$$

Wir wenden die „ODER“-Regel an, um die Wahrscheinlichkeit zu bestimmen, dass überhaupt ein positives Testergebnis ermittelt wird. Hier folgen wir beiden Zweigen des Baumes, die zu einem positiven Testergebnis führen und addieren jeweils diese Teilwahrscheinlichkeiten.

Schließlich setzen wir den Satz von Bayes ein, um anstatt der Wahrscheinlichkeit für ein positives Testergebnis, wenn man an Covid19 erkrankt ist, die Wahrscheinlichkeit zu erhalten, dass man tatsächlich an Covid19 erkrankt, wenn man ein positives Testergebnis erhält. Die erste Wahrscheinlichkeit interessiert Pharmafirmen und Gesundheitsbehörden für die allgemeine Treffsicherheit des Tests. Die zweite Wahrscheinlichkeit interessiert jede Person, welche einen solchen Test machen wurde und ein positives Ergebnis bekommt. Für dieses Umdrehen der Bedingungen benutzen wir den Satz von Bayes:

$$\mathbb{P}[\text{"Covid +"} | \text{"Test +"}] = \frac{\mathbb{P}[\text{"Test +"} | \text{"Covid +"}] \cdot \mathbb{P}[\text{"Covid +"}]}{\mathbb{P}[\text{"Test +"}]} = 0.0015776 \setminus \mathbb{P}[\text{"Test +"}]$$

Hier benötigen wir nun die Wahrscheinlichkeit, einen positiven Test zu bekommen. Wir wissen von allen Covid19-Erkrankten und von allen Nicht-Erkrankten wie wahrscheinlich sie ein positives Testergebnis erhalten. Wir kombinieren nun die UND' und die ODER'-Regel, um die Wahrscheinlichkeiten der beiden Pfade zusammenzurechnen:

$$\mathbb{P}[\text{"Test +"}] = \mathbb{P}[\text{"Test +"} | \text{"Covid+"}] \cdot \mathbb{P}[\text{"Covid+"}] + \mathbb{P}[\text{"Test +"} | \text{"Covid-"}] \cdot \mathbb{P}[\text{"Covid-"}] = 0.0017 \cdot 0.928 + 0.9983 \cdot 0.12 = 0.1213736 = 12.14 \%$$

Daraus ergibt sich

$$\mathbb{P}[\text{"Covid +"} | \text{"Test +"}] = \frac{\mathbb{P}[\text{"Test +"} | \text{"Covid +"}] \cdot \mathbb{P}[\text{"Covid +"}]}{\mathbb{P}[\text{"Test +"}]} = 0.0015776 \setminus 0.1213736 = 0.0129979 = 1.3 \%$$

Dass diese Wahrscheinlichkeit extrem niedrig ist, liegt einerseits an der geringen Anzahl an Sars-COV2 Infizierten und andererseits an der hohen Rate von Falsch Positiven mit 12 %. Dieses Beispiel soll auch als Warnung dienen, dass positive Tests eine Überprüfung erfordern und man andererseits vorsichtig mit dem Umgang mit falsch positiven Ergebnissen sein soll.

Für eine ausführliche Erklärung sei dieses Video empfohlen: <https://www.youtube.com/watch?v=lG4Vkp0G3ko>.

Kombinatorik

Bei Ereignissen und deren Wahrscheinlichkeiten ist in manchen Situationen die Reihenfolge des Eintretens relevant, etwa bei Nukleinsäurenketten im Erbgut oder Peptidketten in Proteinen, und in anderen Situationen irrelevant, etwa wenn die Anzahl der Patienten, die durch ein Medikament erfolgreich behandelt werden, gemessen wird.

Betrachten wir Nucleinsäuretriplets, die aus den 3 verschiedenen Nucleinsäuren Adenin, Cytosin und Guanin gebildet werden. Hier gibt es stets mehr Möglichkeiten, da etwa bei einem DNA-Strang das Triplet "ACG" eine andere Bedeutung als "AGC", "CAG", "CGA", "GAC" oder "GCA" hat. Alleine durch Umordnen ergeben sich hier 6 Möglichkeiten mit den 3 Nucleinsäuren Triplets zu bilden. Diese 6 Möglichkeiten resultieren aus 3 verschiedenen Optionen für die 3. Stelle, danach nur noch je 2 Optionen für die 2. Stelle und die 3. Stelle ist dann jeweils bereits belegt, also $3 \cdot 2 \cdot 1$. Die Kurzschreibweise dafür ist $3!$, gesprochen "3 Faktorielle" oder "3 Fakultät".

Grundsätzlich betrachten wir mögliche Arten, etwas umzuordnen, also die Reihenfolge von Objekten zu verändern. Allgemein heißt das Ändern der Reihenfolge **Permutation** und hat als Grundformel

$$n!$$

Situationen, bei denen die **Reihenfolge relevant** ist, heißen **Variationen**.

Beispielsweise betrachten wir uns 5 Laborgeräte, welche in einem Laborraum platziert werden sollen, wobei nur Plätze für 3 Geräte zur Verfügung stehen. Hier ist sehr wohl eine unterschiedliche Anordnung vorliegend, ob die Zentrifuge vor dem Spektrometer oder umgekehrt platziert ist und es muss nun selektiert werden, welche 3 Geräte gleichzeitig in Verwendung sind. Wir können also aus 5 Geräten für den 1. Platz wählen, danach verbleiben noch 4 für den 2. Platz und 3 für den 3. Platz. Für die übriggebliebenen 2 Geräte ist kein Platz mehr vorhanden. Wir haben also $5 \cdot 4 \cdot 3$ Möglichkeiten dafür, oder anders gesagt $\frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = \frac{5!}{2!}$.

Daher lautet die allgemeine Formel für die **Variation** von k Elementen aus einer Grundmenge von n möglichen Elementen, die zur Auswahl stehen, **ohne Zurücklegen**, bei denen also einmal "verbrauchte" Beobachtungen nicht mehr beobachtet werden können,

$$\frac{n!}{k!}$$

Eine andere Option, wie die Reihenfolge relevant ist, aber Beobachtungen mehrfach vorkommen können, bietet das Toto-Spiel der österreichischen Lotterien, dessen Zufallszahlen gemeinsam mit den Lotto-Zahlen gezogen werden. 7 Zahlen, jeweils aus 0, 1, ..., 9 werden gezogen, wobei jede Stelle der 7-stelligen resultierenden Ziffer unabhängig von den anderen Stellen gezogen wird. Es kann also 9999999 genauso ein Ergebnis sein wie 1234567. An jeder Stelle gibt es 10 Möglichkeiten, welche Ziffer gezogen wird, daher gibt es für 7 solche Stellen 10^7 Möglichkeiten.

Allgemein hat eine **Variation** von k Elementen aus n möglichen Elementen **mit Zurücklegen**, also mit der Option dieselbe Beobachtung mehrfach zu beobachten die Formel

$$n^k$$

Situationen, bei denen die Reihenfolge nicht relevant ist, heißen **Kombinationen**.

Wenn von 30 behandelten Patienten 2 auf die Behandlung ansprechen, ist grundsätzlich irrelevant, ob dies der 1., 2., 3. etc. Patient ist. Allerdings ist wichtig, dass für den ersten erfolgreich behandelten Patienten 30 Personen zur Verfügung stehen, für den 2. erfolgreich behandelten Patienten nur noch 29 Personen, da der 1. bereits ausgeschieden ist. Da die Reihenfolge, wer Patient Nr. 1 und wer Patient Nr. 2 ist, grundsätzlich egal ist, wird durch die möglichen Anordnungsoptionen für 2 Personen $2!$ dividiert, also ergibt sich aus der Formel für die Variation, bei der für die Anordnungsmöglichkeiten korrigiert wird, $\frac{30!}{28! \cdot 2!}$.

Allgemein hat eine Kombination von k Elementen aus n Elementen ohne Zurücklegen die Formel

$$\frac{n!}{k! \cdot (n - k)!},$$

wobei die Kurzschreibweise dieses Ausdrucks $\binom{n}{k}$ lautet und **Binomialkoeffizient** heißt.

Möglichkeiten zur Anordnung mit und ohne relevanter Reihenfolge, mit und ohne Zurücklegen
Variation (Reihenfolge relevant) Kombination (Reihenfolge irrelevant)

ohne Zurücklegen

$$\frac{n!}{(n - k)!}$$

$$\binom{n}{k}$$

mit Zurücklegen

$$n^k$$

$$\frac{(n + k - 1)!}{k!(n - 1)!}$$

Wichtige Wahrscheinlichkeitsverteilungen und ihre Anwendungen

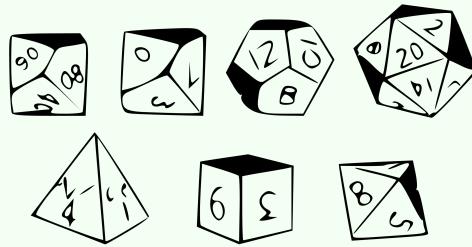
Diskrete Verteilungen

Gleichverteilung

Die diskrete Gleichverteilung ist konzeptuell die einfachste aller Verteilungen. Dabei hat jede der beteiligten Kategorien dieselbe Wahrscheinlichkeit, welche daher $\frac{1}{\text{Anzahl der Kategorien}}$ beträgt.

Wir kennen sie bereits von unserem Beispiel des Würfels des 6-seitigen Würfels sehr genau. Jede der Augenzahlen 1 bis 6 hat dieselbe Wahrscheinlichkeit $\frac{1}{6}$.

Pen & Paper Rollenspiel ist ein nicht mehr allzu weit verbreitetes Hobby, das die Basis für die moderne Computerspielindustrie gelegt hat. Hierbei wird nicht nur bei 6-seitigen Würfeln, sondern auch mit 4, 8, 10, 12, 20, 100-seitigen Würfeln gewürfelt, um gleichverteilte Zufallszahlen zu erzeugen, welche sich auf das virtuelle Spielgeschehen auswirken.



Allgemein ist so ein Würfelwurf mit einem n-seitigen Würfel genau das Zufallsexperiment, das die diskrete Gleichverteilung beschreibt

$$\mathbb{P}[X = x] = \frac{1}{n}, \quad x \in \{1, 2, \dots, n\}$$

Bei einem vollkommen gleich gewichteten 20-seitigen Rollenspielwürfel, hat also jede geworfene Zahl zwischen '1' und '20' die gleiche Wahrscheinlichkeit $\frac{1}{20}$.

Die diskrete Wahrscheinlichkeitsverteilung der diskreten Gleichverteilung lautet

$$\mathbb{P}[X = x] = \frac{1}{n}, \quad x \in \{1, 2, \dots, n\}$$

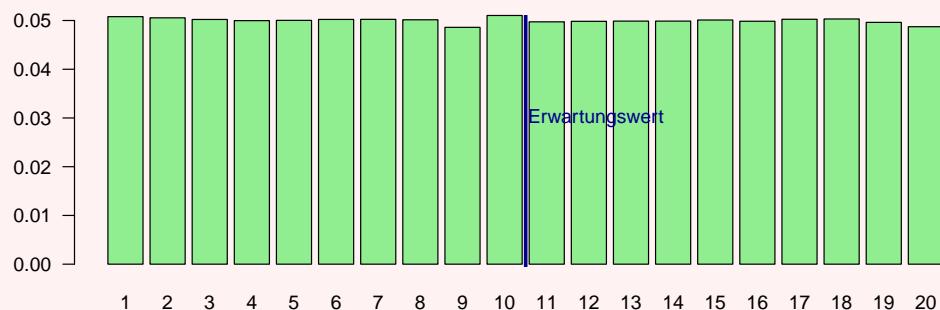
Der Erwartungswert dieser Verteilung lautet

$$\mathbb{E}[X] = \frac{n+1}{2}$$

die Varianz

$$\mathbb{V}[X] = \frac{n^2 - 1}{12}$$

Gleichverteilung für 20-seitigen Würfel



Binomialverteilung und Bernoulli-Verteilung

Wir beobachten ein Zufallsexperiment, bei dem ein bestimmtes Ergebnis, nennen wir es "Erfolg", mit Wahrscheinlichkeit p eintritt. Das Ergebnis einer solchen Messung ist 0 im Falle, dass kein Erfolg eintritt und 1 im Erfolgsfall, formal

$$\mathbb{P}[X = x] = \begin{cases} p & \text{für } x = 1 \text{ (Erfolg)} \\ 1 - p & \text{für } x = 0 \text{ (Misserfolg)} \end{cases}$$

Die Verteilung einer solchen Zufallsgröße X heißt **Bernoulli-Verteilung**, das Zufallsexperiment mit dieser Anordnung **Bernoulli-Experiment**.

Wiederholen wir ein solches **Bernoulli-Experiment n-mal unabhängig voneinander**, misst X die Anzahl der Erfolge bei n Versuchen, also ganze Zahlen zwischen 0 und n . Hier ist wichtig, dass die **einzelnen Bernoulli-Experimente einander nicht beeinflussen** und die **Wahrscheinlichkeit für einen Erfolg in allen Versuchen die gleiche** bleibt.

Eine solche Versuchsanordnung nennt man in der Wahrscheinlichkeitsrechnung **Bernoulli-Kette** wegen der *Verkettung von Bernoulli-Experimenten*. Die daraus resultierende Wahrscheinlichkeitsverteilung für die Anzahl der Erfolge heißt **Binomialverteilung**, formal

$$\mathbb{P}[X = x] = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

Eine Binomialverteilung zeichnen also folgende Eigenschaften aus:

- "Jedes Einzelexperiment kann nur 1 von genau 2 möglichen Ausgängen haben (Erfolg oder Misserfolg)."
- "Die Wahrscheinlichkeit für einen Erfolg verändert sich nicht bei mehrmaligem Experimentieren." Hier stelle man sich vor, dass beim Lottospiel die Wahrscheinlichkeit für das Ziehen einer Kugel in jeder Runde verändert wird, da Kugeln aus dem Spiel herausgezogen werden und dadurch die Anzahl verändert wird.
- "Es wird eine Reihe von Experimenten mit je zwei Versuchsausgängen hintereinander ausgeführt". Das ist die weniger formelle Beschreibung für eine Bernoullikette von Experimenten, welche die Grundlage bildet.

Bei einer Binomialverteilung ist der Erwartungswert einfach zu ermitteln und interpretieren, denn er ist die Anzahl der Versuch multipliziert mit der Wahrscheinlichkeit für einen Erfolg und ergibt die mittlere erwartete Anzahl an Erfolgen in der Experimentenanordnung.

$$\mathbb{E}[X] = n \cdot p$$

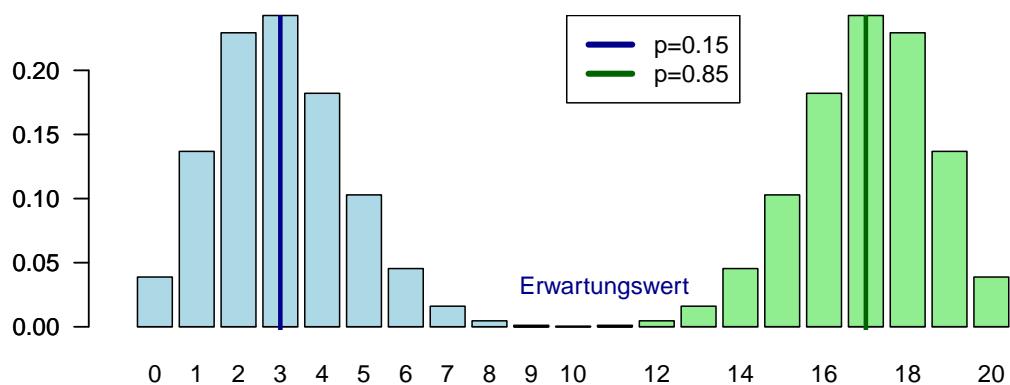
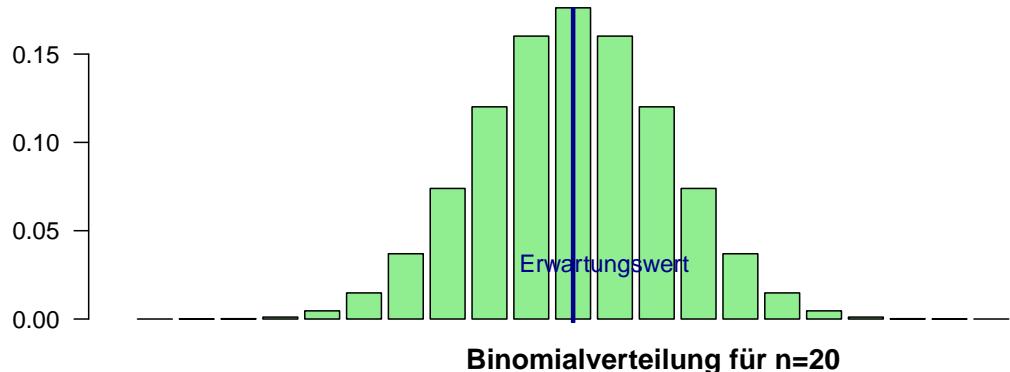
Die Varianz bezieht auch noch die Gegenwahrscheinlichkeit, dass ein Misserfolg resultiert, in Betracht,

$$\mathbb{V}[X] = n \cdot p \cdot (1 - p)$$

und die Standardabweichung lautet dann

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

Binomialverteilung für n=20, p=0.5



Ein Anwendungsbeispiel für Binomialverteilungen ist ein Szenario, in dem man bekanntermaßen die Wahrscheinlichkeit kennt, mit der bei einem Patienten die intendierte Wirkung eines Medikaments eintritt. In einer Überprüfungsstudie der Krankenkasse sollen 120 Patienten, die ein Medikament gegen Bluthochdruck verschrieben bekommen, betrachtet werden. Man weiß, dass das Medikament erwartungsgemäß in 78 % der Fälle eine positive Wirkung entfaltet.

Daher wird erwartet, dass im Mittel $120 \cdot 0.78 = 93.6$ Personen, also zwischen 93 und 94 Personen, positiv auf das Medikament ansprechen. Dieser Wert schwankt im Mittel mit einer Standardabweichung 4.5378409.

Die Wahrscheinlichkeit, dass mindestens ein Patient eine positive Wirkung zeigt, $\mathbb{P}[X \geq 1] = 1 - \mathbb{P}[x = 0] = 1 - 0.78^0 \cdot 0.22^{120} = 1 - 1.2323149 \times 10^{-79} = 1 - \text{pb}inom(0, \text{size}=120, \text{prob}=0.78) = 1$

Die Wahrscheinlichkeit, dass wenigstens die Hälfte der Patienten auf das Medikament anspricht beträgt $\mathbb{P}[X \geq 60] = 1 - \text{cd}f_binom(60, n = 120, p = 0.78)$ (Maxima) $= 1 - \text{pb}inom(60, \text{size} = 120, \text{prob} = 0.78)$ (R) $= 1$

Die Wahrscheinlichkeit, dass höchstens 90 Patienten auf das Medikament ansprechen, beträgt $\mathbb{P}[X \leq 90] = \text{cd}f_binom(90, n = 120, p = 0.78)$ (Maxima) $= \text{pb}inom(90, \text{size} = 120, \text{prob} = 0.78)$ (R) $= 0.2438408 = 24.3840826 \%$.

Hypergeometrische Verteilung

Die hypergeometrische Verteilung beschreibt klassische Experimente, bei denen Einheiten aus dem Prozess zur Beobachtung gezogen werden, wodurch es zu einer merkbaren Veränderung der Verteilung der Experimentereignisse kommt.

Ein allseits bekannter solcher Prozess ist die Lotterziehung, welche wöchentlich in viele Staaten durchgeführt wird. In Österreich ist das System "6 aus 45" gängig.

Dies bedeutet, dass 6 Kugeln aus 45 möglichen Kugeln gezogen werden. Vor der ersten Ziehung hat jede Kugel eine Wahrscheinlichkeit von $\frac{1}{45}$ gezogen zu werden. Nach dem Ziehen der ersten Kugel sind anstatt 45 nur noch 44 Kugeln in der Urne, wodurch jede Kugel eine Wahrscheinlichkeit von $\frac{1}{44}$ gezogen zu werden bekommt. Das bedeutet, dass sich die Wahrscheinlichkeit in jedem Schritt verändert und ist ein wichtiges Unterscheidungskriterium zur Binomialverteilung.

Auch hier sind "Erfolge" von Interesse. Im Falle des Lottospiels wären das wie viele der getippten Zahlen mit den gezogenen Zahlen übereinstimmen. Fangen wir mit einem "6er" an.

$$\mathbb{P}[\text{"6er"}] = \frac{1}{\binom{45}{6}} = 1 / 8.14506 \times 10^6 = 1.23 \times 10^{-5}\%$$

Bei einem "5er" gibt es schon mehrere Möglichkeiten, in welcher Reihenfolge die Kugeln gezogen werden können und welche 5 richtigen und 1 falsche Zahlen getippt werden.

$$\mathbb{P}[\text{"5er"}] = \frac{\binom{6}{5} \cdot \binom{39}{1}}{\binom{45}{6}} = 0.0029\%$$

Bei einem "4er" gibt es schon mehrere Möglichkeiten, in welcher Reihenfolge die Kugeln gezogen werden können und welche 4 richtigen und 2 falsche Zahlen getippt werden.

$$\mathbb{P}[\text{"4er"}] = \frac{\binom{6}{4} \cdot \binom{38}{2}}{\binom{45}{6}} = \text{dhyper}(x=4, m=6, n=39, k=6) \text{ (R)} = 0.14\%$$

Der 3er im Lotto wird entsprechend analog ermittelt.

Im allgemeinen Fall beträgt also die Wahrscheinlichkeitsverteilung der Hypergeometrischen Verteilung mit n Elementen, die aus N möglichen Elementen gezogen werden, von denen k von K interessanten Elementen „erwischt“ werden:

$$\mathbb{P}[X = k] = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

Bei einer hypergeometrischen Verteilung ist der Erwartungswert einfach zu ermitteln und interpretieren, denn er ist die Anzahl der Versuche multipliziert mit dem Anteil der interessanten Einheiten K an allen Elementen N und ergibt die mittlere erwartete Anzahl an Erfolgen in der Experimentenanordnung.

$$\mathbb{E}[X] = n \cdot \frac{K}{N}$$

Die Varianz lautet,

$$\mathbb{V}[X] = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}$$

und die Standardabweichung lautet dann

$$\sigma = \sqrt{n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}}$$

Diese Verteilung findet ihre Anwendung in der *Gen- und Genontologieanalyse* unter der Bezeichnung **“Fisher’s Exact Test”**. Dabei werden Gene, welche für einem metabolischen Pfad zugeordnet sind (Genontologie) k, mit den in einer Analyse (NGS, Microarray, SNP) als signifikant exprimiert gefunden Genen K verglichen, und den insgesamt im Organismus bekannten N bzw. im Organismus als aktiv gefundenen Genen n gegenübergestellt.

Negativbinomialverteilung, geometrische Verteilung

Die Negativbinomialverteilung wie die Binomialverteilung beschäftigt sich mit wiederholten Bernoulli-experimenten, deren Wahrscheinlichkeit sich nicht verändert. Im Unterschied zur Binomialverteilung zählt sie X die Anzahl der Erfolge 0 bis ∞ , bis eine bestimmte vorgegebene Anzahl an Misserfolgen r aufgetreten ist oder je nach Interpretation umgekehrt, die Anzahl der Misserfolge, bis eine bestimmte Anzahl an Erfolgen aufgetreten ist.

$$\mathbb{P}[X = k] = \binom{k + r - 1}{k} (1 - p)^r p^k$$

Das bedeutet, dass sie Szenarien beschreibt, in denen die Ethikkommission bei der Testung einer innovativen Krebsbehandlung vorschreibt, dass diese nur erfolgen darf, bis 15 PatientInnen verstorben sind. Dadurch wird r auf 15 fixiert und k zählt die Anzahl der erfolgreich bis dahin behandelten PatientInnen.

Ein Spezialfall der Negativbinomialverteilung ist die **geometrische Verteilung** (engl. **geometric distribution**), bei der $r=1$ gesetzt wird. Es werden also die *Erfolge bzw. Misserfolge gezählt, bis einmal das Gegenteil eintritt, gezählt*.

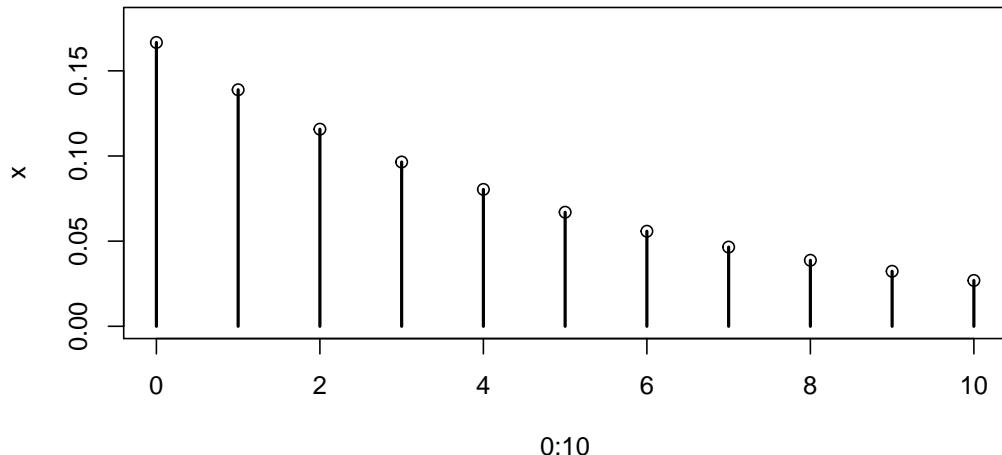
Stellen wir uns vor, ein(e) Studierende(r) kommt betrunken nach einer Studentenfeier nach Hause und kann sich nicht mehr erinnern, welcher der 6 Schlüssel an seinem/ihrem Schlüsselbund der passende Wohnungsschlüssel ist. Jeweils ein Schlüssel wird ausprobiert und danach wieder fallen gelassen. Da die Person so betrunken ist, hat sie sofort vergessen, welche(n) Schlüssel sie bereits probiert hat und kann jeden der Schlüssel mit gleicher Wahrscheinlichkeit $\frac{1}{6}$ wieder ausprobieren. Das Experiment ist natürlich vorbei, sobald der/die Betrunkene in die Wohnung gelangt, also den richtigen Schlüssel erwischt hat.

Diese Wahrscheinlichkeitsverteilung wird beschrieben durch die geometrische Verteilung

$$\mathbb{P}[X = k] = (1 - p)^1 p^k$$

beschrieben. Hier kann es nur eine Reihenfolge geben, nämlich zuerst alle Misserfolge (falscher Schlüssel) bis zum Erfolg (Finden des richtigen Schlüssels).

Geometrische Verteilung für Schlüssel



Poissonverteilung

Die **Poissonverteilung** wird auch die “**Verteilung der seltenen Ereignisse**” genannt. Im Unterschied zur Binomialverteilung zählt sie Ereignisse, deren Höchstanzahl nicht a priori bekannt ist, etwa die Anzahl von Wartenden bei einer Supermarktkassa oder Produktionsfehler wie Bläschen einschlüsse in Glasplatten. Daher kann die Zufallsvariable X grundsätzlich Werte von 0 bis ∞ annehmen.

Die Wahrscheinlichkeitsfunktion der Poisson-Verteilung lautet

$$\mathbb{P}[X = x] = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0, 1, 2, \dots$$

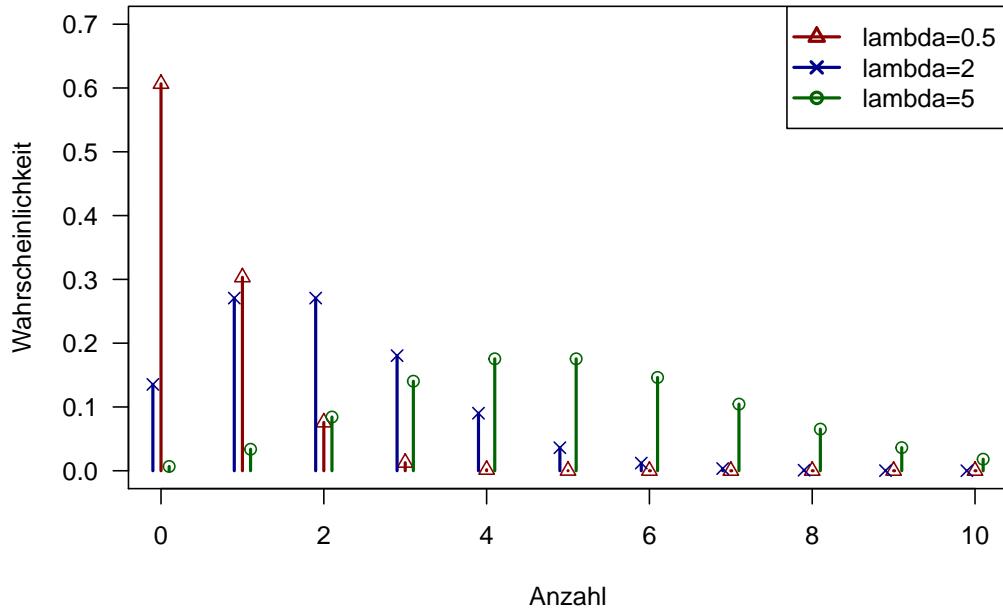
Bei der Poissonverteilung gilt, dass λ gleichzeitig auch Erwartungswert und Varianz der Poissonverteilung ist:

$$\mathbb{E}[X] = \lambda,$$

$$\mathbb{V}[X] = \lambda$$

Der Parameter λ wird dabei die mittlere **Auftrittsraten** des gesuchten Ereignisses interpretiert und kann jede positive reelle Zahl sein.

Poisson-Verteilung



Ein Beispiel hierfür wäre die Verteilung für die Anzahl von Verunreinigungen bei der Herstellung biologischer Nährmedien und Petrischalen. Ein unerfahrener Labortechniker zu Beginn seiner Karriere macht um Mittel 1 Fehler, der zur Verunreinigung auf der Petrischale führt, bei der Herstellung von Petrischalen. Hier wird also der Parameter $\lambda = 1$ gesetzt.

Die Wahrscheinlichkeit, dass er eine fehlerfreie Petrischale produziert ist

$$\mathbb{P}[X = 0] = \frac{1^0}{0!} e^{-1} = \text{pdf_poisson}(0, 1) \text{ (Maxima)} = \text{dpois}(0, \lambda = 1) \text{ (R)} = 0.3679 = 36.79\%.$$

Die Wahrscheinlichkeit, dass er eine Petrischale mit mindestens 1 Fehler produziert ist daher die Gegenwahrscheinlichkeit

$$\mathbb{P}[X \geq 1] = 1 - \mathbb{P}[X = 0] = 1 - \frac{1^0}{0!} e^{-1} = 1 - \text{pdf_poisson}(0, 1) \text{ (Maxima)} = 1 - \text{dpois}(0, \lambda = 1) \text{ (R)} = 0.6321 = 63.21\%.$$

Ein anderes Beispiel wäre die Warteschlange bei der Supermarktkassa. Aus Erfahrung wissen wir, dass die mittlere Länge der Schlange 2 Personen beträgt. Daher ist $\lambda = 2$. Wir wollen für dieses Szenario einige Fragen beantworten:

- Wie wahrscheinlich ist es, hin zu kommen, und der Erste zu sein?

Erster ist man, wenn 0 Personen in der Schlange vor einem warten, also suchen wir $\mathbb{P}[X = 0] = \frac{2^0}{0!} e^{-2} = \text{pdf_poisson}(0, 2) \text{ (Maxima)} = \text{dpois}(0, \lambda = 2) \text{ (R)} = 0.1353 = 13.53\%$.

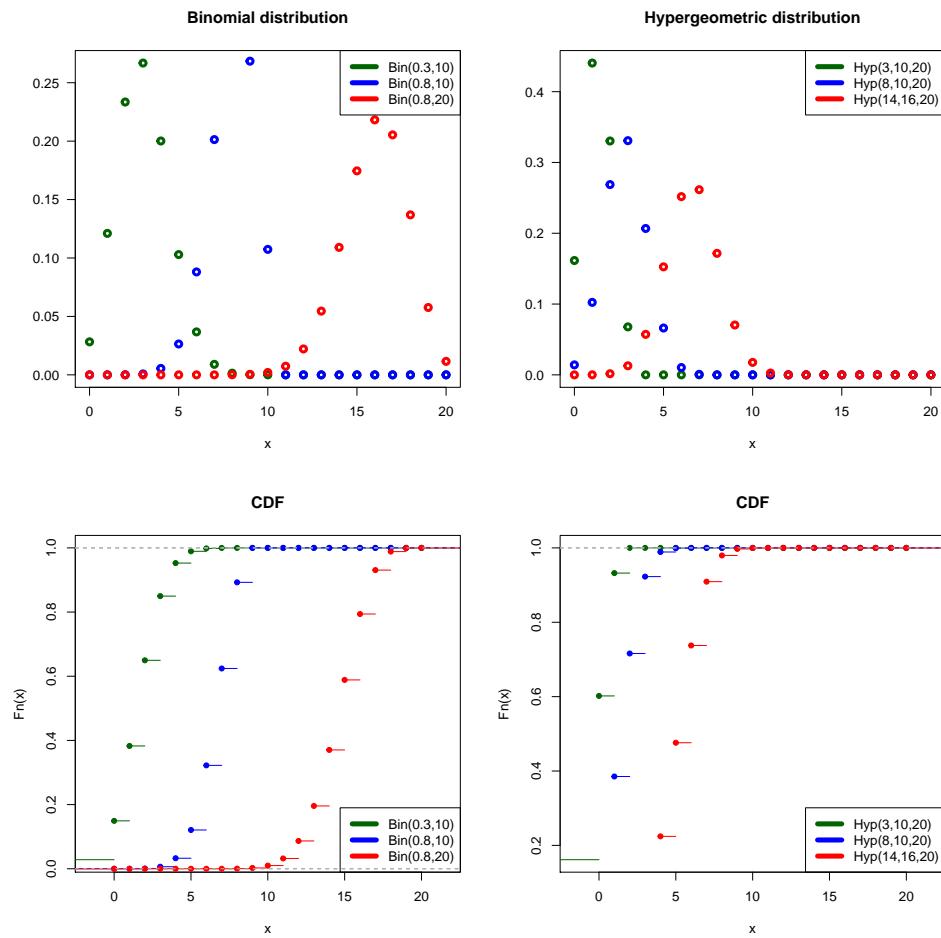
- Wie wahrscheinlich ist es, höchstens 2 Personen vor sich in der Schlange zu haben?

Hier suchen wir die Wahrscheinlichkeit $\mathbb{P}[X \leq 2] = \text{cdf_poisson}(2, 2) \text{ (Maxima)} = \text{ppois}(2, \lambda = 2) \text{ (R)} = 0.6767 = 67.67\%$.

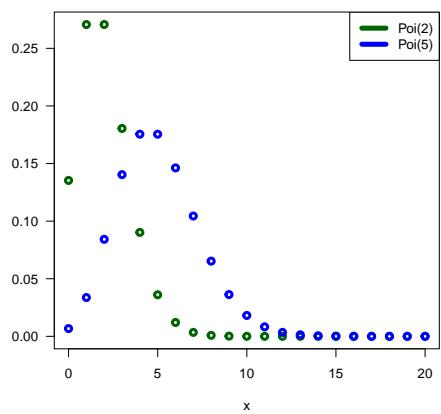
- Wie wahrscheinlich ist es, mindestens 2 Personen vor sich in der Schlange zu haben?

Hier suchen wir die Wahrscheinlichkeit $\mathbb{P}[X \geq 2] = 1 - \mathbb{P}[X = 1] = 1 - \left(\frac{2^0}{0!} e^{-2} + \frac{2^1}{1!} e^{-2} \right) = 1 - \text{cdf_poisson}(1, 2) \text{ (Maxima)} = 1 - \text{ppois}(1, \lambda = 2) \text{ (R)} = 0.594 = 59.4\%$.

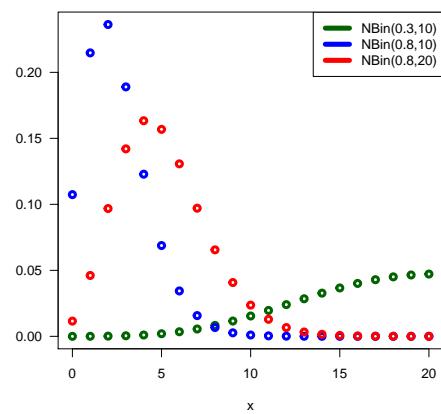
Übersicht über alle diskreten Verteilungen



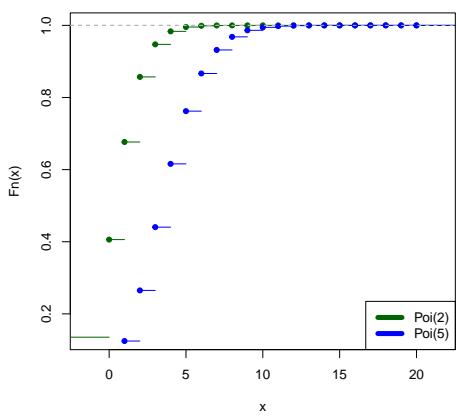
Poisson distribution



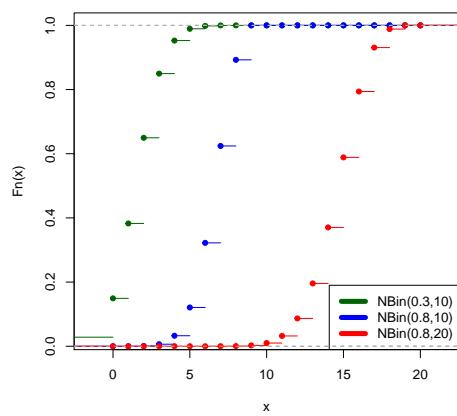
Negative Binomial distribution



CDF



CDF



Stetige Verteilungen

Bei kontinuierlichen Verteilungen hat ein einzelner Zahlenwert stets Wahrscheinlichkeit 0, da sonst ihre überabzählbare Summe keine endliche Wahrscheinlichkeit von höchstens 1 ergeben könnte. Die Verteilung einer kontinuierlichen Zufallsgröße X kann mittels ihrer **Wahrscheinlichkeitsdichte**

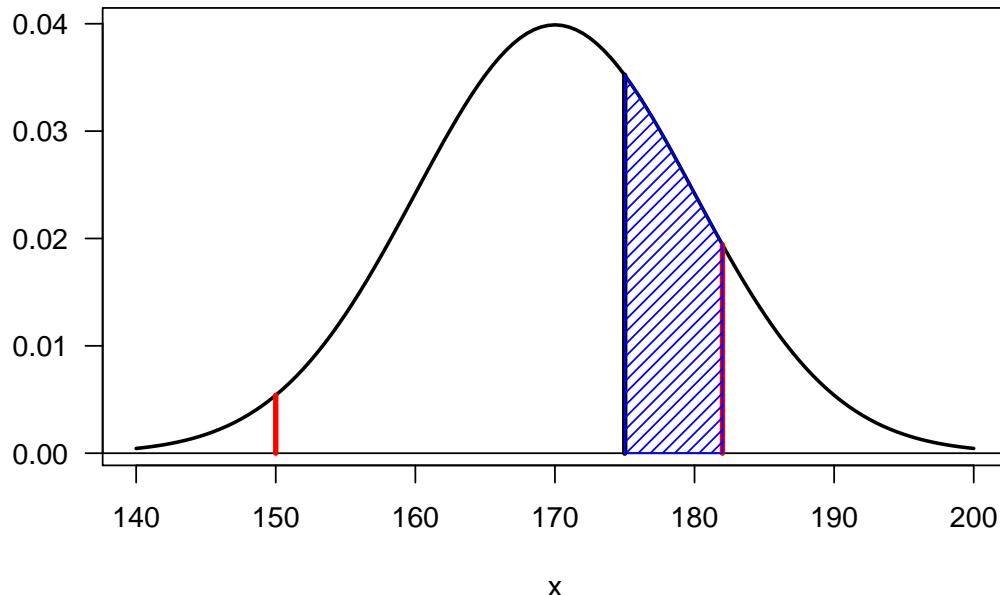
$$p(x) = \lim_{\Delta x \rightarrow 0} \frac{\mathbb{P}[x \leq X \leq x + \Delta x]}{\Delta x}$$

beschrieben werden. Eine konkrete Wahrscheinlichkeit für einen Intervallbereich errechnet sich als die Fläche unter der Kurve der Dichtefunktion, also das Integral der Dichtefunktion

$$\mathbb{P}[X \in [a, b]] = \int_a^b p(x) dx$$

Die Körpergrößen von Menschen sind erfahrungsgemäß normalverteilt. Als mittlere Körpergröße im westlichen Bereich wird ein Wert von etwa 170 cm angenommen.

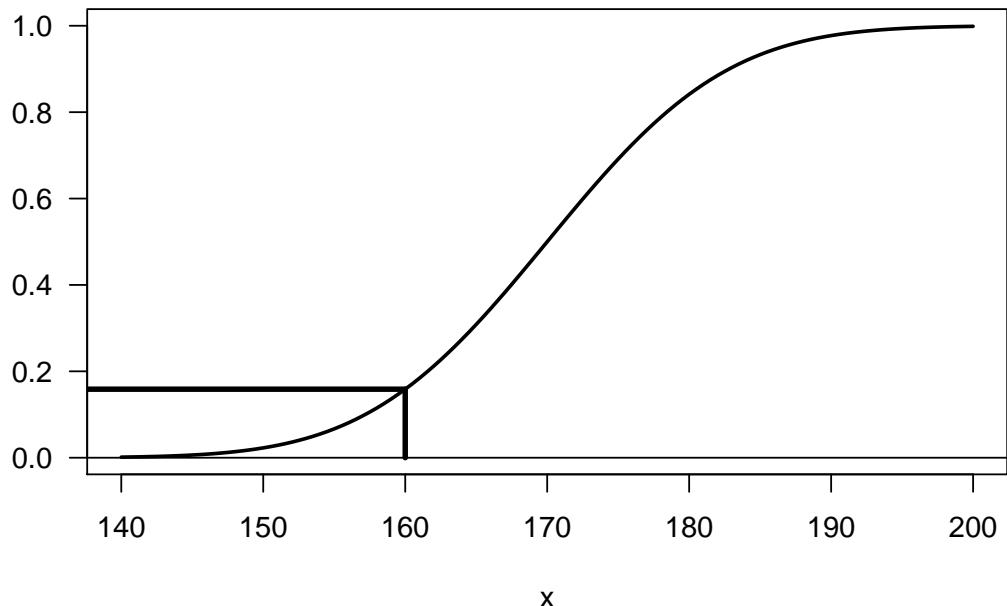
Dichtefunktion



Die konkrete Körpergröße von 150 cm hat eine Wahrscheinlichkeit von 0 als solche beobachtet zu werden. Aber eine Körpergröße zwischen 175 cm und 182 cm hat sehr wohl eine konkrete Wahrscheinlichkeit, nämlich $\mathbb{P}[175 \leq X \leq 182] = \int_{175}^{182} p(x) dx = 0.1935 = 19.35\%$.

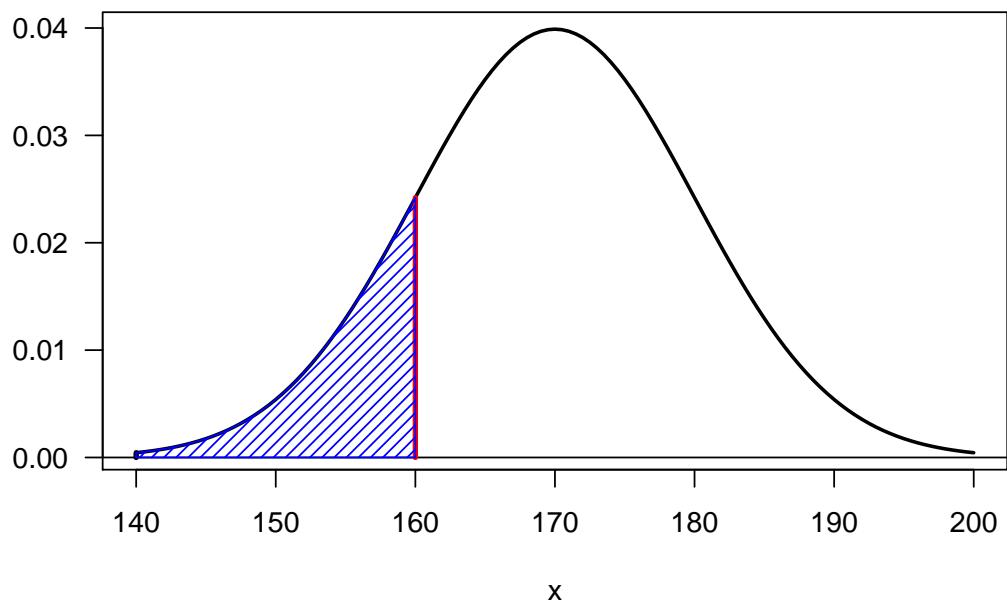
Da die Fläche unter der Kurve, also die Stammfunktion, der Dichtefunktion für die Ermittlung der Wahrscheinlichkeit so relevant ist, hat diese Funktion eine eigene Bezeichnung, die **kumulative Verteilungsfunktion**. In ihrer Darstellung wird an der y-Achse die Wahrscheinlichkeit und an der x-Achse die konkreten Werte, bis zu denen diese Wahrscheinlichkeit ermittelt wird, aufgetragen.

Verteilungsfunktion



Hier wird an der Stelle 160 cm sehr wohl eine Wahrscheinlichkeit dargestellt, nämlich die Wahrscheinlichkeit, eine Körpergröße von höchstens 160 cm zu haben, $\mathbb{P}[-\infty \leq X \leq 160] = \int_{-\infty}^{160} p(x)dx = 0.1587 = 15.87\%$. Diese Wahrscheinlichkeit entspricht in der obigen Darstellung jener Fläche unter der Kurve:

Dichtefunktion



Gleichverteilung

Die stetige Gleichverteilung haben wir wie die diskrete Gleichverteilung gleich zu Beginn des Kapitels kennengelernt. Das Werfen der Nadeln und Überprüfungen, in welchem Winkel sie aufkommen, entspricht einer stetigen Gleichverteilung, wenn man Winkel beliebig genau nachmisst und präzisiert.

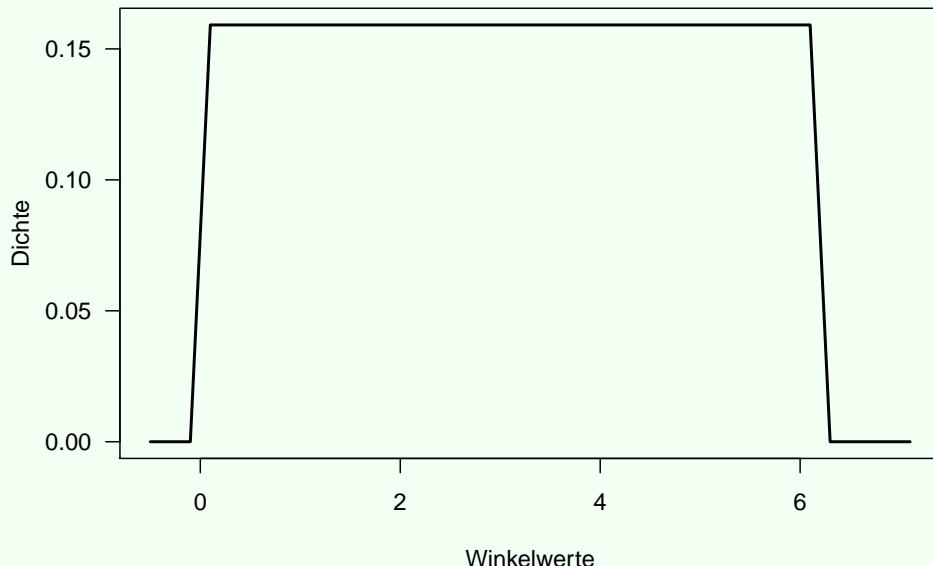
Hier wird also ein Teilbereich in einem ganzen Intervall als der Bereich von Interesse betrachtet. Wie wir bereits zuvor besprochen haben, hat ein einzelner Wert bei stetigen Verteilungen keine Wahrscheinlichkeit, sondern die Fläche unter der Dichtekurve entspricht der Wahrscheinlichkeit.

Die Wahrscheinlichkeitsdichte der Gleichverteilung im Intervall $[a, b]$ ist sehr einfach angeschrieben:

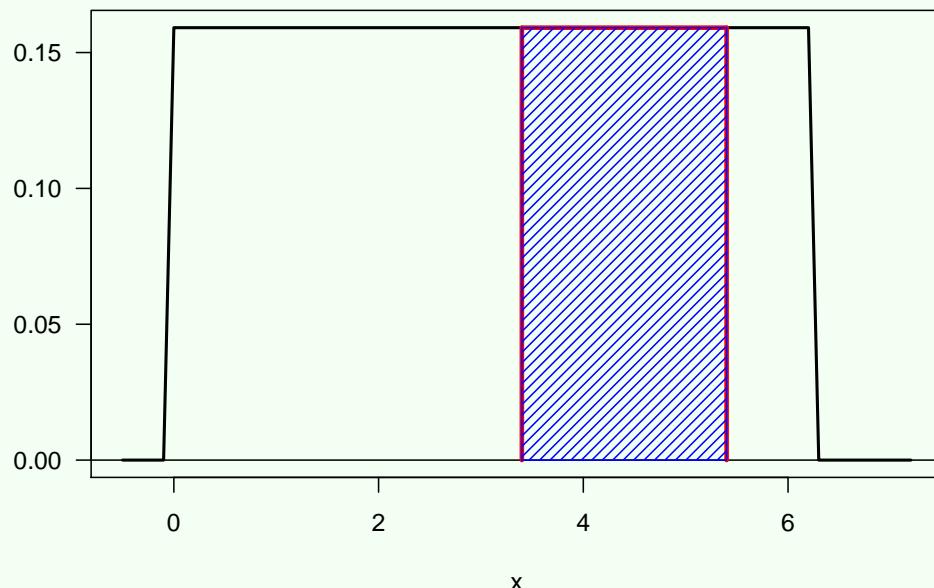
$$f(x) = \frac{1}{b-a}, x \in [a, b]$$

Für unsere Nadeln, die in einem Winkel von $[0, 2\pi]$ fallen können, bedeutet das, dass die Dichtefunktion folgendermaßen aussieht.

Dichtefunktion – Stetige Gleichverteilung



Dichtefunktion – Stetige Gleichverteilung



Die Wahrscheinlichkeit, dass die Nadel in einem Winkel zwischen 194.8° und 309.4° fällt beträgt
$$\int_{194.8}^{309.4} p(x)dx = \frac{309.4 - 194.8}{360 - 0} = 0.318 = 31.8\%$$

Gauß'sche Normalverteilung und Standardnormalverteilung

Die **Gauß'sche Normalverteilung**, auch Gauß-Verteilung oder nur kurz Normalverteilung genannt, ist eine der wichtigsten Verteilungen, da sie in vielen in Natur, Wirtschaft und Industrie auftretenden Prozessen auf natürliche Weise als Datenverteilung stetiger Messungen auftritt. Zusätzlich hat sie einige Eigenschaften, welche sie zur mathematisch wichtigen Funktion im Zusammenhang mit Modellen, Inferenz und Konfidenzbereichen macht.

Die **Normalverteilung** hat die grundsätzliche Formel

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

für ihre Dichtefunktion.

Die Normalverteilung besitzt die Parameter Mittelwert μ und Standardabweichung σ , die auch die Bedeutung haben, die wir von diesen Werten kennen:

- μ ist der Mittelwert, aber auch der Median und Modus der Normalverteilung, als die Stelle, an der das Zentrum der Daten liegt, daher im eigentlichen Sinne der Erwartungswert der Verteilung.

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

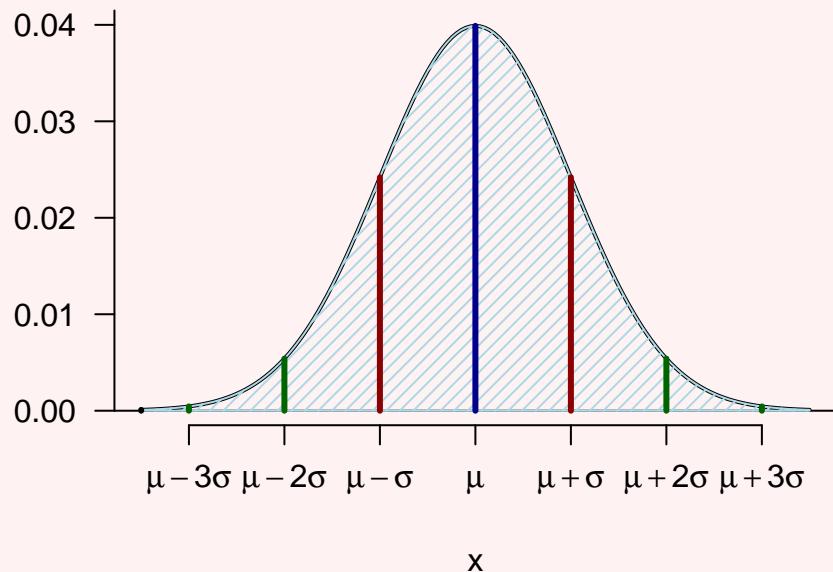
- σ ist die Standardabweichung und gibt die Schwankung der Daten wieder. Daher ist die Varianz der Daten $Var(X) = \sigma^2$.

Als Zusammenhang mit den Schätzwerten arithmetischer Mittelwert und Varianz bzw. Standardabweichung gilt, dass diese Schätzer dann am besten funktionieren und geeignet sind, wenn Daten annähernd wie eine Normalverteilung verteilt sind und keine Ausreißer enthalten, da sie sensibel gegenüber solchen Ausreißern und Abweichung von der Normalverteilung sind.

Dichtefunktion (Gauß'sche Glockenkurve) und Verteilungsfunktion

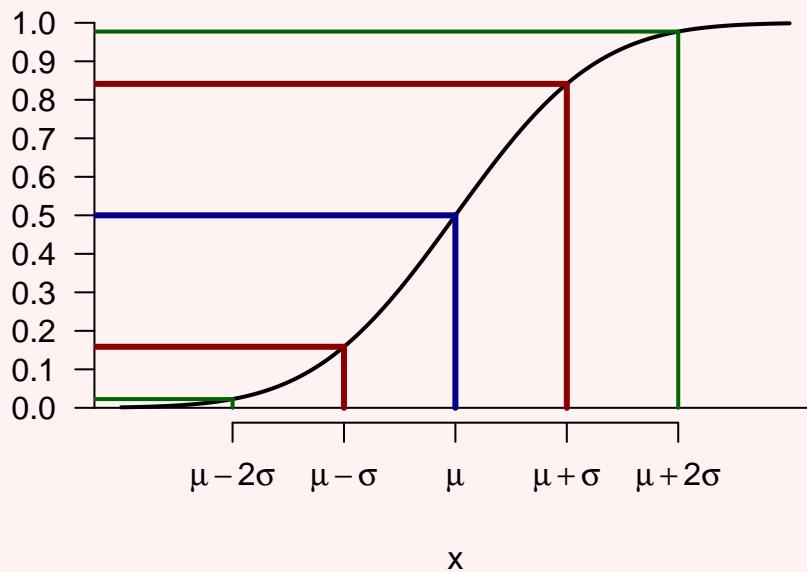
Die Form dieser Dichtefunktion ist sehr charakteristisch und trägt den Namen Gauß'sche Glockenkurve.

Dichtefunktion der Normalverteilung



Das Integral dieser Funktion, die Stammfunktion $F(x)$, also die **Verteilungsfunktion der Normalverteilung**, ist nicht mehr in geschlossener Form berechenbar und kann daher nur noch numerisch angenähert und daher graphisch dargestellt werden.

Verteilungsfunktion der Normalverteilung



Die korrespondierenden Werte auf der y-Achse der Verteilungsfunktion entsprechen den Flächen unterhalb der Kurve der Dichtefunktion bis zu der jeweiligen Stelle. Hier ist der Mittelwert μ farbcodiert, wie auch die relevanten Stellen $\mu \pm \sigma$ und $\mu \pm 2\sigma$.

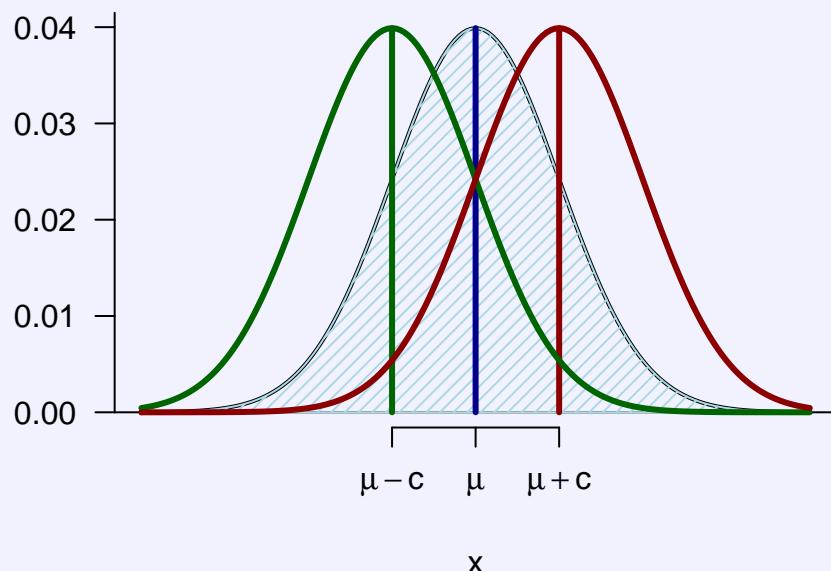
In den graphischen Darstellungen erkennt man sehr schön die Eigenschaften der Normalverteilung:

- Der Erwartungswert der Normalverteilung ist μ

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

Wird μ größer oder kleiner gemacht, wird dadurch die Kurve entlang der x-Achse nach links und rechts verschoben. Die Form der Glockenkurve bleibt gleich.

Dichtefunktion der Normalverteilung



- Die Gauß'sche Glockenkurve ist symmetrisch, wobei die Symmetrieachse vertikal durch den Mittelwert μ verläuft. Daher gilt auch arithmetischer Mittelwert = Median = μ für alle normalverteilten Daten, denn es liegt die Hälfte der Beobachtungen unterhalb und die andere Hälfte oberhalb des Mittelwerts. Daher gilt auch für die Verteilungsfunktion immer

$$F(\mu) = 0.5$$

- Die Varianz der Normalverteilung ist σ^2

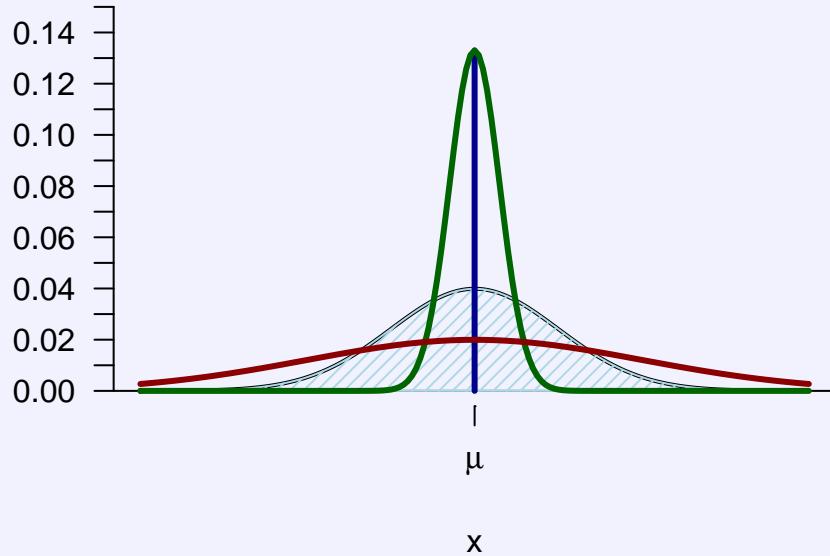
$$\mathbb{V}[X] = \sigma^2$$

und ihre Standardabweichung daher σ .

Wird σ größer oder kleiner gemacht, bleibt die Kurve um denselben Mittelwert zentriert. Die die Form der Glockenkurve ändert sich aber:

- Wird die Standardabweichung größer, so wird die Glockenkurve breiter und flacher, weil die Fläche unter der Kurve 1 sein muss.
- Wird die Standardabweichung kleiner, so wird die Glockenkurve schmäler und höher, weil die Fläche unter der Kurve 1 sein muss.

Dichtefunktion der Normalverteilung



- An den Stellen $\mu - \sigma$ und $\mu + \sigma$ befinden sich die Wendepunkte der Gaußschen Glockenkurve. Innerhalb dieser, also zwischen $\mu - \sigma$ und $\mu + \sigma$ liegen 68.3 % der Werte der Wahrscheinlichkeitsdichte, also etwa 2/3 der Daten.
- An den Stellen $\mu \pm 2\sigma$ schneiden die Wendetangenten der Glockenkurve die x-Achse. Innerhalb dieser, also zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$ liegen 95.4 % der Werte der Wahrscheinlichkeitsdichte, also etwa nur 5 von 100 Datenwerten außerhalb. Daher werden diese Werte auch als Näherungswerte der Quantile für die Konstruktion der 95% - Konfidenzintervalle und Prädiktionsintervalle herangezogen.
- Innerhalb von $\mu \pm 3\sigma$ liegen 99.73 % der Werte der Wahrscheinlichkeitsdichte, also etwa nur 5 von 100 Datenwerten außerhalb.
- Die **Standardnormalverteilung** kommt zustande, indem von der Werten der Zufallsgröße X jeweils der Mittelwert μ abgezogen wird, wodurch eine Translation um das Zentrum 0 entsteht, und anschließend durch die Standardabweichung σ dividiert wird, wodurch für jegliche Einheit und Messskala korrigiert wird, mathematisch

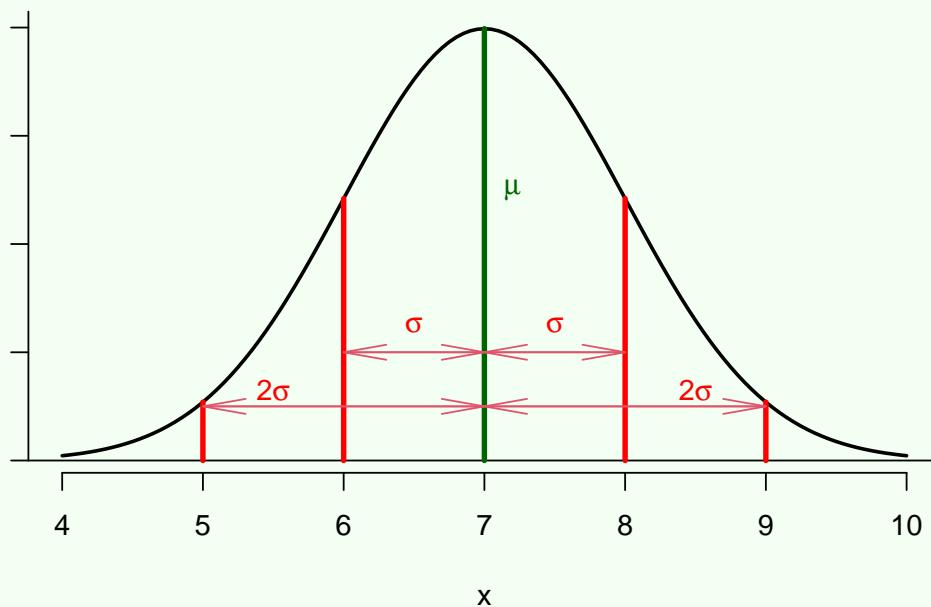
$$z = \frac{x - \mu}{\sigma}$$

Die Standardnormalverteilung ist also die Normalverteilung $\mathcal{N}(0, 1)$ mit Mittelwert 0 und Varianz und Standardabweichung 1, weshalb sie Standardnormalverteilung genannt wird.

Ablesen von Normalverteilungsparametern: Nährmedien

Wir beginnen mit einer rein graphischen Analyse und dem Ablesen von Zahlenwerten. Wir haben in folgenden Grafik die typische Verteilung der Nährmedienmasse in g dargestellt. Daraus wollen wir rein graphisch den Mittelwert und die Standardabweichung ablesen.

Dichtefunktion



Wir erkennen an der Stelle des Maximums, durch das auch die Symmetriearchse verläuft, unschwer den **Mittelwert** der Verteilung, $\mu = 7$ g.

Für die **Standardabweichung** müssen wir uns etwas mehr anstrengen und einerseits die Lage der Wendepunkte abschätzen, da wir wissen, dass die *Wendestellen* bei $\mu \pm \sigma$ zu finden sind und andererseits wissen, dass innerhalb von $\mu \pm 2 \cdot \sigma$ die mittleren 95% der Werte zu finden sind. Wir können also ablesen, dass die **Standardabweichung** $\sigma = 8 - 7 = 1$ g beträgt.

Rechnen mit Normalverteilung: Körpergrößen

Die Körpergrößen von Menschen sind erfahrungsgemäß normalverteilt. Als mittlere Körpergröße im westlichen Bereich wird ein Wert von etwa 170 cm angenommen.

Man weiß, dass 84% der Menschen eine Körpergröße von über 160.06 cm haben. Dass 84% der Menschen größer als 160.06 cm bedeutet dasselbe wie, dass 16% der Personen kleiner oder gleich groß wie 160.06 cm sind.

Daraus ermitteln wir im ersten Schritt die Standardabweichung mithilfe der Standardnormalverteilung, bei der die Quantile immer eindeutig definiert sind: `quantile_normal(0.16,0,1)` (Maxima) = `qnorm(0.16)` (R) = -0.9944579

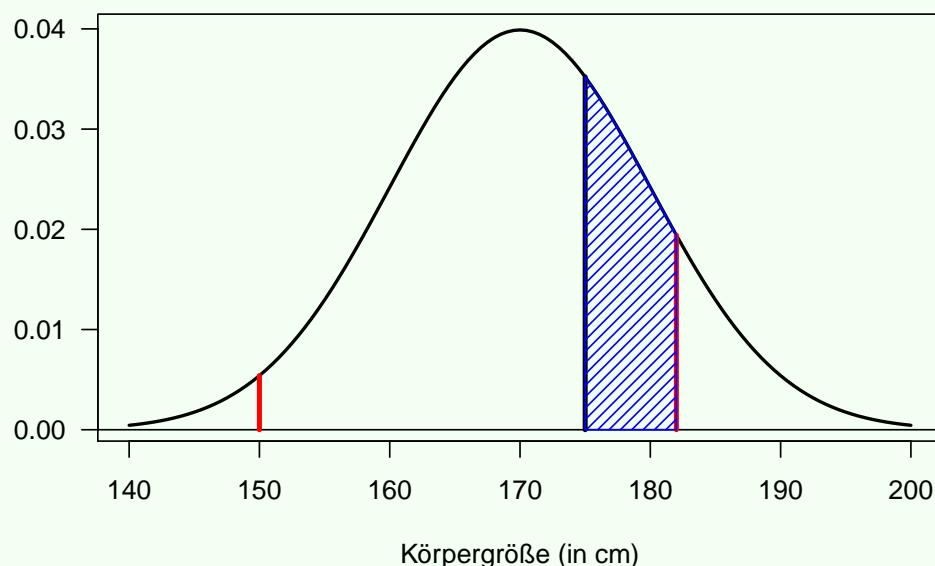
Nun da wir das Quantil der Standardnormalverteilung haben, benutzen wir die Formel der z-Transformation zur Umrechnung in unsere konkrete Normalverteilung der Körpergrößen und setzen dann unsere Informationen in diese Formel ein.

$$z = \frac{x - \mu}{\sigma}$$

$$-0.9944579 = \frac{160.06 - 170}{\sigma}$$

Nun muss dieser Ausdruck nur noch nach σ umgeformt und aufgelöst werden, um die Standardabweichung rechnerisch zu ermitteln: $\sigma = 9.9953957 \approx 10$ cm.

Dichtefunktion



Die konkrete Körpergröße von 150 cm hat eine Wahrscheinlichkeit von 0 als solche beobachtet zu werden, da bei einer stetigen Verteilung ein konkreter Einzelwert keine Fläche unterhalb der Dichtefunktion aufweist - ein Strich ist nämlich unendlich dünn. Nur für einen Bereich, ein Intervall, kann eine Fläche und damit eine Wahrscheinlichkeit ermittelt werden.

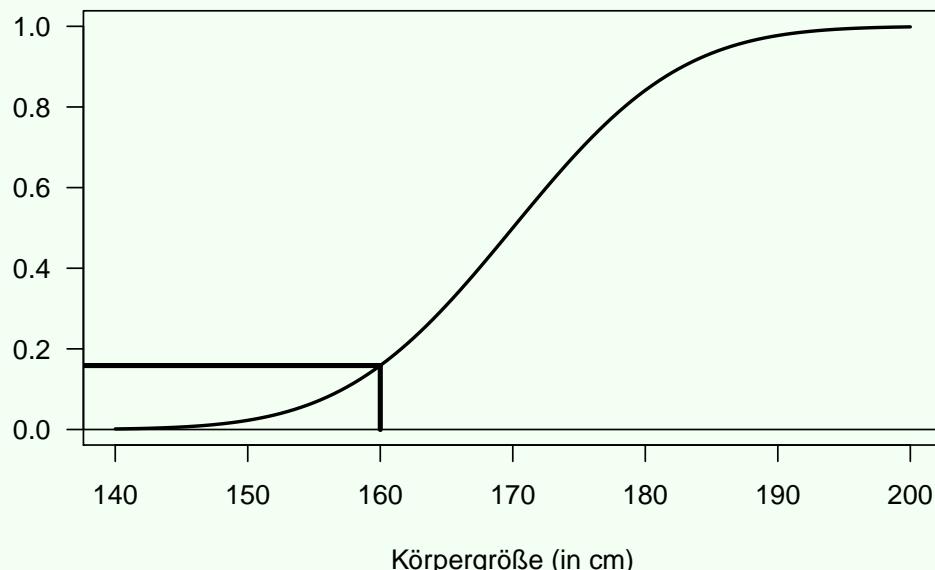
Eine Körpergröße zwischen 175 cm und 182 cm hat daher sehr wohl eine konkrete Wahrscheinlichkeit, nämlich $\mathbb{P}[175 \leq X \leq 182] = \int_{175}^{182} p(x)dx = 0.1935 = 19.35\%$.

Die Berechnung in Maxima erfolgt mithilfe des Befehls `cdf_normal(182,170,10)-cdf_normal(175,170,10)`.

Die Berechnung in R mithilfe von `(pnorm(182,mean = 170,sd=10)-pnorm(175,mean = 170,sd=10))`.

Da die Fläche unter der Kurve, also die Stammfunktion, der Dichtefunktion für die Ermittlung der Wahrscheinlichkeit so relevant ist, hat diese Funktion eine eigene Bezeichnung, die **kumulative Verteilungsfunktion**. In ihrer Darstellung wird an der y-Achse die Wahrscheinlichkeit und an der x-Achse die konkreten Werte, bis zu denen diese Wahrscheinlichkeit ermittelt wird, aufgetragen.

Verteilungsfunktion



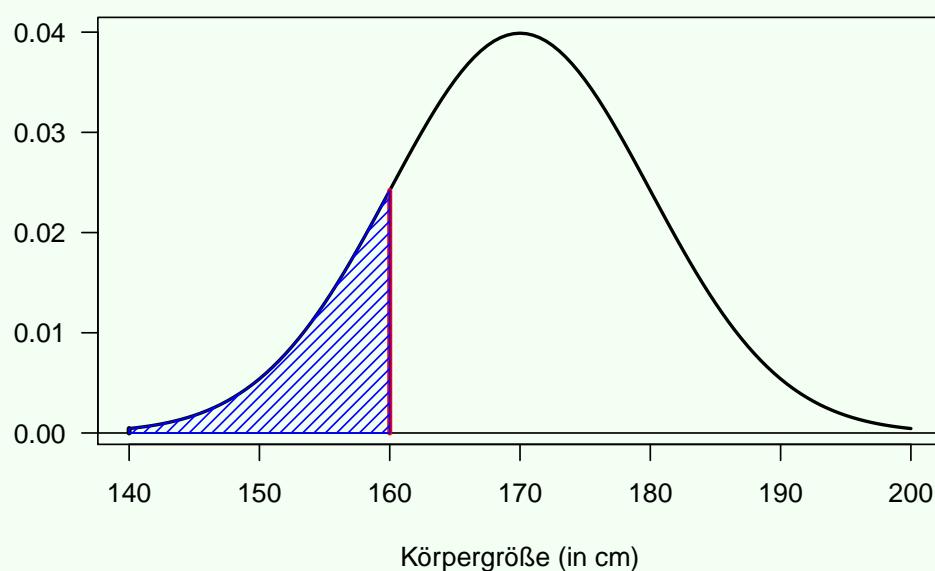
Hier wird an der Stelle 160 cm sehr wohl eine Wahrscheinlichkeit dargestellt, nämlich die Wahrscheinlichkeit, eine Körpergröße von höchstens 160 cm zu haben, $\mathbb{P}[-\infty \leq X \leq 160] = \int_{-\infty}^{160} p(x)dx = 0.1587 = 15.87\%$.

Die Berechnung in Maxima erfolgt mithilfe des Befehls `cdf_normal(160, 170, 10)`.

Die Berechnung in R erfolgt mithilfe des Befehls `pnorm(160, mean = 170, sd=10)`.

Diese Wahrscheinlichkeit entspricht in der obigen Darstellung jener Fläche unter der Kurve:

Dichtefunktion



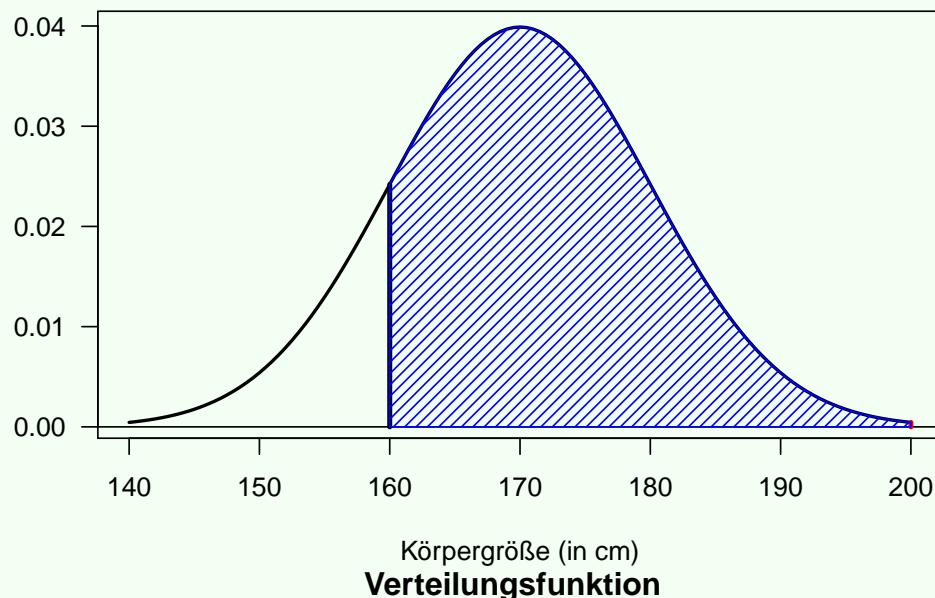
Will man umgekehrt wissen, mit welcher Wahrscheinlichkeit eine Körpergröße von mindestens 160 cm auftritt, brauche ich die Gegenwahrscheinlichkeit. Da 160cm ja Wahrscheinlichkeit 0 hat, muss ich es hier bei der Gegenwahrscheinlichkeit nicht weglassen. Im Unterschied zu den diskreten Dichten gilt also $\mathbb{P}[160 \leq X \leq \infty] = \int_{160}^{\infty} p(x)dx = 1 - \int_{-\infty}^{160} p(x)dx = 0.8413 = 84.13\%$.

Die Berechnung in Maxima erfolgt mithilfe des Befehls `1-cdf_normal(160,170,10)`.

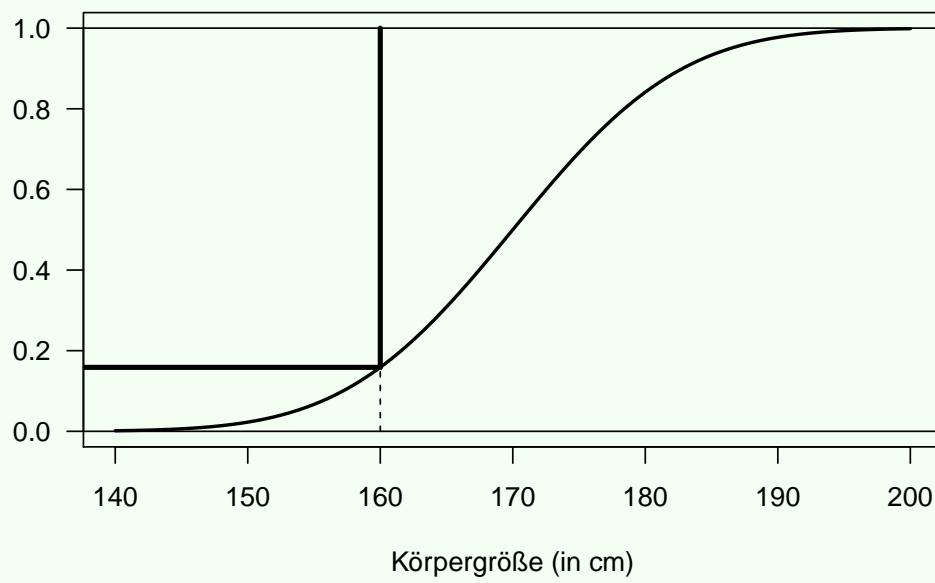
Die Berechnung in R erfolgt mithilfe des Befehls `1-pnorm(160,mean = 170,sd=10)`.

Diese Wahrscheinlichkeit entspricht in der obigen Darstellung jener Fläche unter der Kurve:

Dichtefunktion



Verteilungsfunktion

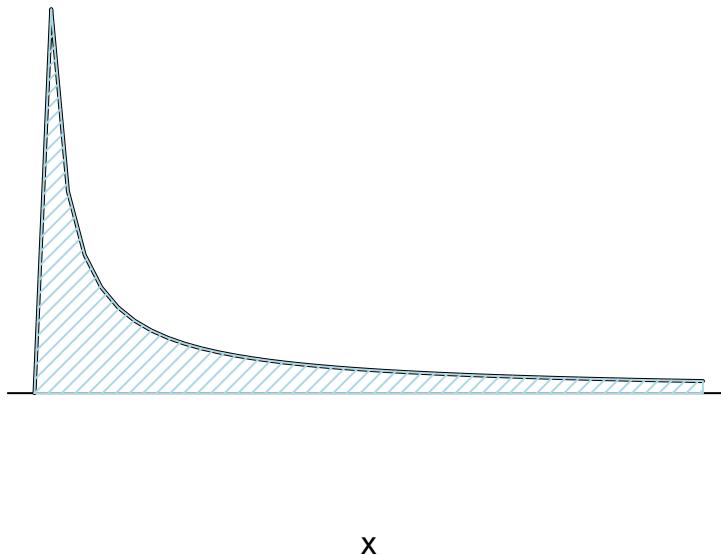


Log-Normalverteilung

Die Log-Normalverteilung ist die Verwandte der Normalverteilung, welche zur Anwendung kommt, wenn die Daten selbst nicht normalverteilt sind, aber durch Anwenden der Logarithmusfunktion als Transformation normalverteilt werden.

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

Dichtefunktion der Normalverteilung



Hier wird der logarithmische Verlauf augenscheinlich. Dadurch eignet sich die Log-Normalverteilung für Beobachtungen mit schwerem rechten Rand. Umgekehrt kann in vielen Fällen durch Logarithmieren bei solchen Beobachtungen in Normalverteilung zurück transformiert werden.

Student-t Verteilung

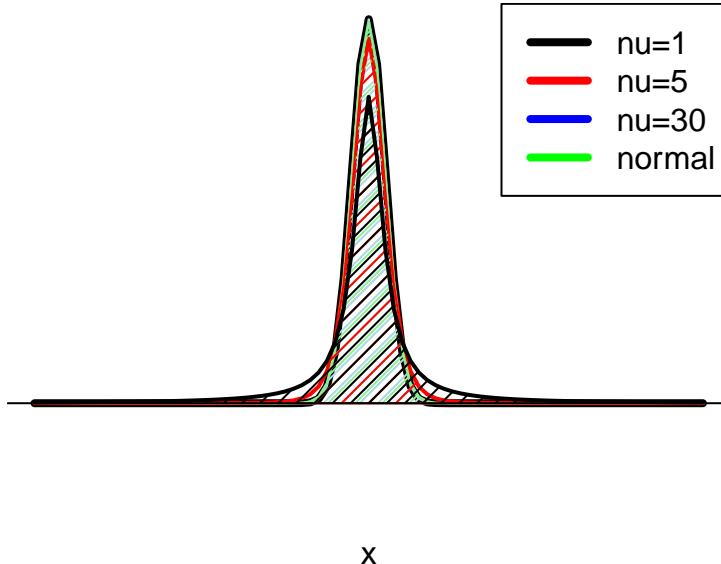
Die Student-t Verteilung ist eine symmetrische Verteilung, welche allerdings deutlich schwerere Ränder als die Normalverteilung hat, also Szenarien modelliert, in denen sich die Daten deutlich häufiger weit vom Zentrum der Verteilung entfernen als bei einer Normalverteilung.

Die Dichtefunktion der t-Verteilung ist

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \cdot \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Der Parameter der Student-t-Verteilung ist der *Freiheitsgrad ν* . Bei niedrigem Freiheitsgrad sind die Ränder schwerer als bei höheren Freiheitsgraden. Die Student-t Verteilung ist mit der Normalverteilung zusammenhängend, als dass die Normalverteilung ihre Grenzverteilung bildet und sie mit zunehmenden Freiheitsgraden der Normalverteilung immer ähnlicher wird.

Dichtefunktion der Student-t-Verteilung



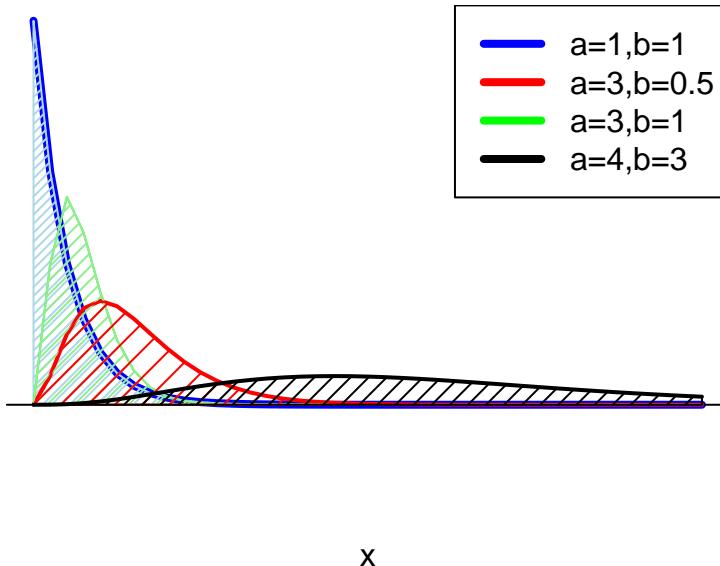
Gamma-Verteilung, Exponentialverteilung, χ^2 -Verteilung

Die Gamma-Verteilung besitzt die Dichtefunktion

$$\frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}}$$

und hat die Form abhängig von den Formparametern a und b.

Dichtefunktion der Gamma-Verteilung



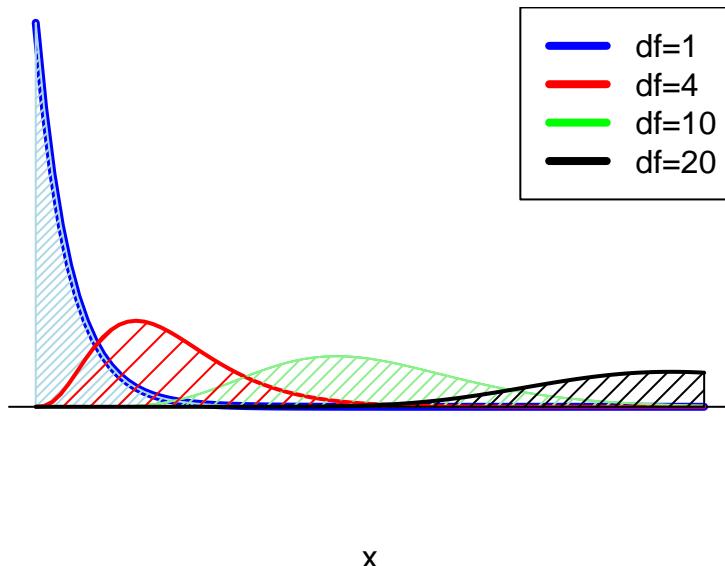
Zwei spezielle Varianten der Gamma-Verteilung sind von besonderer Relevanz:

- Die **Exponentialverteilung** beschreibt eine exponentiell abfallende Verteilung, die zustande kommt, wenn der Parameter $a=1$ und der Parameter $b=\frac{1}{\lambda}$ gesetzt wird. Die Exponentialverteilung hat die spezielle Eigenschaft **gedächtnislos** zu sein, was bedeutet, dass an jedem Wert der exponentiell abfallende Verlauf gleich wie zu Beginn ist.
- Die **χ^2 -Verteilung** kommt zustande, wenn der Parameter $a=\frac{n}{2}$ und der Parameter $b=2$ gesetzt wird. Diese Verteilung hängt außerdem auf spezielle Weise mit der Normalverteilung zusammen. Die Summe von n quadrierten Standardnormalverteilten Zufallsgrößen X_i ist χ_n^2 verteilt.

$$X_1^2 + X_2^2 + \dots + X_n^2 \sim \chi_n^2$$

Die Dichtefunktion der χ^2 -Verteilung mit $df = k$ Freiheitsgraden haben folgende Struktur

Dichtefunktion der Chi–Quadrat–Verteilung



Es gibt einige Eigenschaften dieser Verteilung, die wichtig für ihre Anwendung bei Hypothesentests und Konfidenzintervallen sind:

- Die χ^2 -Verteilung nimmt nur nichtnegative Werte an, $X \geq 0$.
- Die χ^2 -Verteilung ist aufgrund der unteren Schranke bei 0 inherent nichtsymmetrisch.
- Die χ^2 -Verteilung bewegt sich mit steigenden Freiheitsgraden nach rechts. Für $df \rightarrow \infty$ nähert sie sich sogar der Normalverteilung von ihrer Form her an, allerdings abgeschnitten an 0.

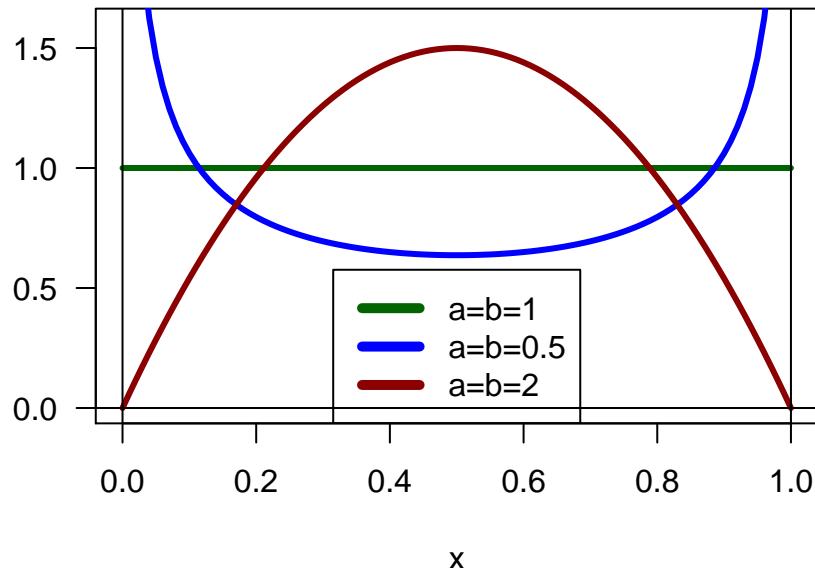
Beta-Verteilung

Die Beta-Verteilung ist im Prinzip eine Verteilung von Prozentwerten zwischen 0 und 1. Ihre Dichtefunktion $Be(\alpha, \beta)$ ist

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

wobei die Werte von $x \in [0, 1]$ also Prozentwerte oder selbst Wahrscheinlichkeiten sind. Die Formparameter $\alpha \in \mathbb{R}$ und $\beta \in \mathbb{R}^+$ bestimmen die Form und den Verlauf der Verteilung.

Beta distribution



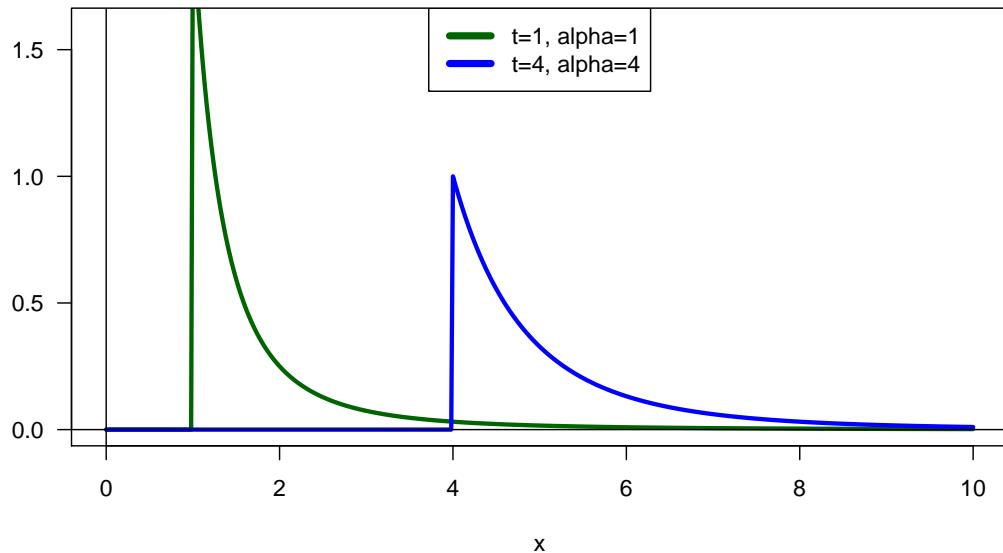
Pareto Verteilung:

Die Pareto Verteilung ist eine Verteilung für Überlebenswahrscheinlichkeiten und daher exponentiell fallend ab einem bestimmten Zeitpunkt x_0 :

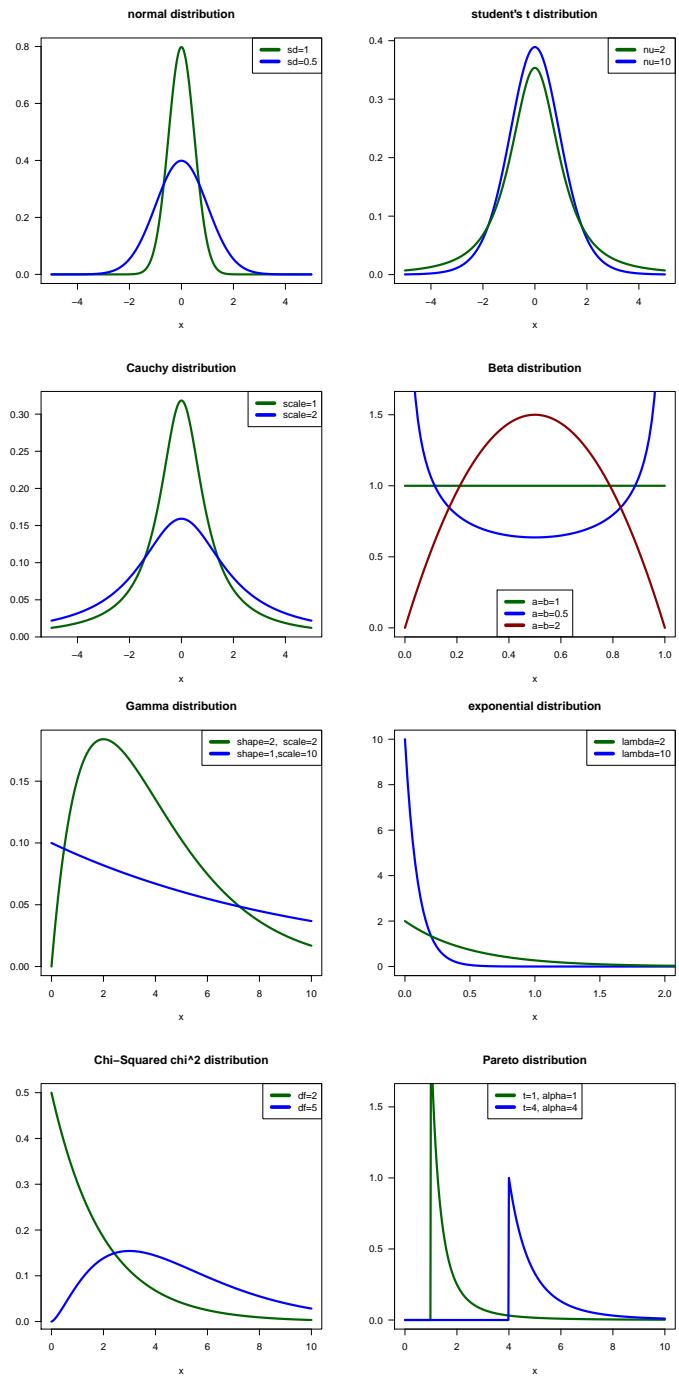
$$f(x; \alpha, x_0) = \alpha x_0^\alpha x^{-(\alpha+1)}$$

ist die Dichtefunktion der Pareto (Typ I) $Pa(\alpha, x_0)$ Verteilung mit "tail index" Parameter α . Hier sind alle Werte $x \geq x_0$ nach dem Schwellwert x_0 angesetzt und $\alpha \in \mathbb{R}^+$ beschreibt die Stärke des exponentiellen Zerfalls.

Pareto distribution



Übersicht über Stetige Verteilungen



Relations between Distributions visualised

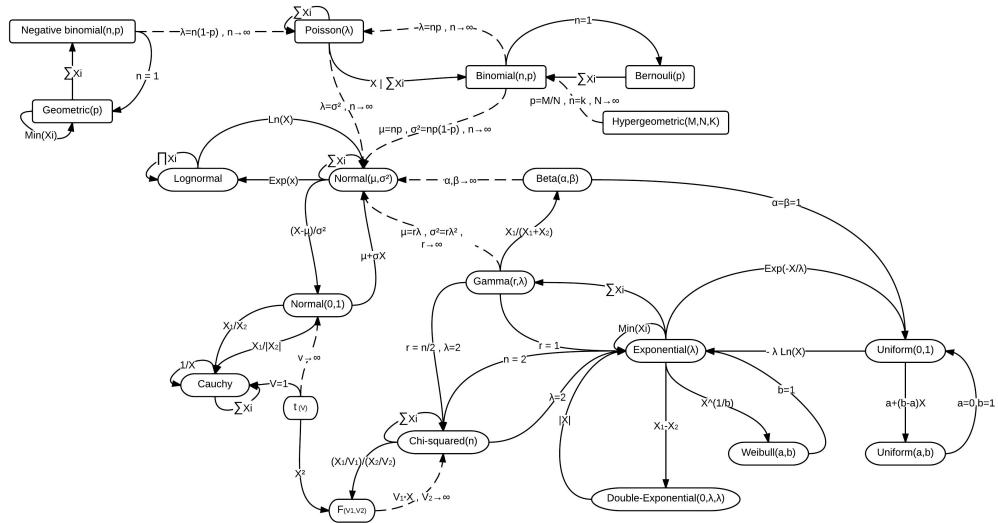


Abbildung 1: Source: www.math.wm.edu/~leemis/2008amstat.pdf

Umsetzung mithilfe von Softwarepaketen im Überblick

Umsetzung in Maxima:

Die Verteilungen sind in Maxima in dem Paket “**distrib**” implementiert.

Dort gibt es die Funktionen, welche den Grundbau für alle Verteilungen bilden:

- (pdf_*) **Probability Density Function** (**Dichtefunktion/Wahrscheinlichkeitsfunktion**)
- (cdf_*) **Cumulative Distribution Function** ((**kumulative**) **Verteilungsfunktion**)
- (quantile_*) **Quantile**
- (mean_*) **Mean** (**Mittelwert/Erwartungswert**)
- (var_*) **Variance** (**Varianz**)
- (std_*) **Standard Deviation** (**Standardabweichung**)
- (random_*) **Random Variate** (**Zufallszahlengenerierung**)

In Maxima stehen folgende häufig verwendete **diskrete Verteilungen** (engl. Discrete distributions) zur Verfügung, wobei jeweils das Symbol * in den obigen Befehlen durch den Namen der Verteilung in der Klammer ersetzt wird:

Binomial Verteilung	(* = binomial)
Poisson Verteilung	(* = poisson)
Bernoulli Verteilung	(* = bernoulli)
Geometrische Verteilung	(* = geometric)
Gleichverteilung	(* = discrete_uniform)
Hypergeometrische Verteilung	(* = hypergeometric)

In Maxima stehen folgende häufig verwendete **stetige Verteilungen** (engl. Continuous distributions) zur Verfügung, wobei jeweils das Symbol * in den obigen Befehlen durch den Namen der Verteilung in der Klammer ersetzt wird:

Normalverteilung	(* = normal)
Student's t Verteilung	(* = student_t)

Umsetzung in R:

Die Verteilungen sind in R in dem Paket “**stats**” implementiert.

Dort gibt es die Funktionen, welche den Grundbau für alle Verteilungen bilden:

- (d*) Density Function (**Dichtefunktion/Wahrscheinlichkeitsfunktion**)
- (p*) Cumulative Distribution Function calculating the probability ((**kumulative Verteilungsfunktion**) berechnet die **Wahrscheinlichkeit**, höchstens einen bestimmten Wert anzunehmen)
- (q*) Quantile
- (r*) Random Variate (**Zufallszahlengenerierung**)

In Maxima stehen folgende häufig verwendete **diskrete Verteilungen** (engl. Discrete distributions) zur Verfügung, wobei jeweils das Symbol * in den obigen Befehlen durch den Namen der Verteilung in der Klammer ersetzt wird:

- Binomial Verteilung** (* = binom)
- Poisson Verteilung** (* = pois)
- Geometrische Verteilung** (* = geom)
- Hypergeometrische Verteilung** (* = hyper)

In Maxima stehen folgende häufig verwendete **stetige Verteilungen** (engl. Continuous distributions) zur Verfügung, wobei jeweils das Symbol * in den obigen Befehlen durch den Namen der Verteilung in der Klammer ersetzt wird:

- Normalverteilung** (* = norm)
- Student's t Verteilung** (* = t)
- Gamma Verteilung** (* = gamma)
- Beta Verteilung** (* = beta)
- Exponantialverteilung** (* = exp)
- Pareto Verteilung** (* = pareto)

Beschreibende Statistik

Beschreibende, oder auch deskriptive Statistik, befasst sich mit nichts weiterem als statistisch erhobene Daten aufzubereiten und zu charakterisieren. Wie diese Daten erhoben wurden und inwiefern diese repräsentativ für eine gewisse Grundgesamtheit sind, ist hier noch nicht das Thema

Merkmale und Daten

Die Größen die wir beobachten, nennen wir hier Merkmale, und ihre möglichen Zustände Merkmalsausprägungen. Wir unterscheiden Merkmale grundsätzlich zwischen 2 Hauptarten von Merkmalen und innerhalb dieser zwischen Untertypen.

Merkmalskategorien

- **Qualitative Merkmale** sind Merkmale, die Kategorien bestimmen und nicht sinnvoll numerisch gemessen werden können. Hier unterscheidet man:
 - **nominal** skalierte Merkmale haben keine vordefinierte Anordnung der zugrund liegenden Kategorien: Geschlecht (M - W - X) , Farben (rot, grün, blau, ...) etc.
 - **ordinal** skalierte Merkmale haben ein klar zugrundeliegende Ordnung der Kategorien: Noten (“ Sehr Gut” ist besser als “Gut” ...), Qualitätsskalen etc.
- **Quantitative Merkmale** können sinnvoll numerisch gemessen und verglichen werden, das heisst die Merkmalsausprägungen sind auf natürliche Weise eine Wertebereich zuordenbar sind. Hier unterscheidet man:
 - **intervallskalierte** Daten enthalten die Werte 0 und möglicherweise negative Zahlenwerte. Daher können nur Differenzen zwischen Werten berechnet werden, aber nicht Werte durcheinander dividiert werden, um Vielfachheiten zu ermitteln. “-5°C ist die (-1)-fache Temperatur von 5°C” ergibt keinen Sinn, “-5°C ist die inf-fache Temperatur von 0°C” noch viel weniger. Beispiele sind Temperaturen in °C oder °F, Jahreszahlen, Kontostände etc.
 - **rationalsskalierte** Daten sind ausschließlich positive Werte bzgl. eines absoluten Nullpunkts orientiert. Daher kann “Vielfachheit” ermittelt und interpretiert werden. “2m ist die doppelte Länge von 1m” ist sinnvoll, “200K ist die halbe Temperatur von 400K” ebenfalls. Beispiele sind physikalische Größen wie Längen, Massen, Temperatur in K, welche ausschließlich positive Werte annehmen.

Darüberhinaus unterscheidet man Merkmale als

- **diskret**, also entweder zu endlich vielen Kategorien gehörend (nominal oder ordinal skaliert) oder klar getrennte Zahlen annehmend, also etwa als Zählvariable 0, 1, 2, ... erfolgreiche Experimente, aber nicht 2.78. Alle kategorialen Variablen wie Farben, etc. und Zählvariablen, welche ein Anzahl bestimmen, soferne diese ”klein genug“ ist.
- **stetig**, also beliebige Werte innerhalb eines Intervalls annehmend eingeschränkt durch die Messgenauigkeit. Körpergrößen, Massen etc. fallen in diese Kategorie, aber auch Zählungen von ”großen“Werten, wenn die Einzelzählung nicht mehr eine Kategorie, sondern ein tatsächlich ausgewerteter Zahlenwert ist.

Grundsätzlich bezeichnet man die zugrundeliegenden *Merkmale*, auch *Zufallsvariablen* genannt, da sie zufällig bei jeder Messung eine Ausprägung annehmen können, meist mit Grossbuchstaben, also X, Y, T und dergleichen. Die Messdaten solcher Merkmale werden dann oft mit den dazugehörigen Kleinbuchstaben x_i, y_i, t_i , $i = 1, \dots, n$ geschrieben.

Im folgenden seien x_1, x_2, \dots, x_n die Messdaten von n Messungen eines Merkmals X. Die Verteilung dieser empirischen Daten kann man auf verschiedenste Arten durch Maßzahlen charakterisieren.

Datenmatrix

Wir unterscheiden einige statistische Grundvokabel im Umgang mit Untersuchungseinheiten und Merkmalen. Dazu gehört die Unterscheidung von **Untersuchungseinheit (engl. unit)**, der einzelnen Probe, Person, etc., und dem **Merkmal**, der **Variable (engl. variable)**, der Eigenschaft, welche für jede Untersuchungseinheit gemessen bzw. erhoben wird. An den Untersuchungseinheiten werden die **Werte (Ausprägungen, Realisierungen)** der Merkmale festgestellt (gemessen, erhoben). Das Ergebnis sind dann **Daten (Beobachtungen, observations)**.

Statistische Daten haben daher meist die Struktur einer rechteckigen **Datenmatrix (data frame)**, in der die Zeilen den Untersuchungseinheiten und die Spalten den Variablen entsprechen.

Probennummer	Messung	Proteinanteil	Fetteinanteil
1	HPLC	9.07	39.31
2	HPLC	8.92	38.82
3	GC	9.45	43.37
4	GC	9.26	42.46
5	HPLC	9.84	38.02

Schätzung und Darstellung qualitativer Messungen

Seien a_1, a_2, a_3, \dots die möglichen Merkmalsausprägungen eines diskreten Merkmals X und bei einer Messreihe von n Messungen $n_i =$ Häufigkeit des Auftretens von a_i , so unterscheidet man die daraus resultierenden Häufigkeiten:

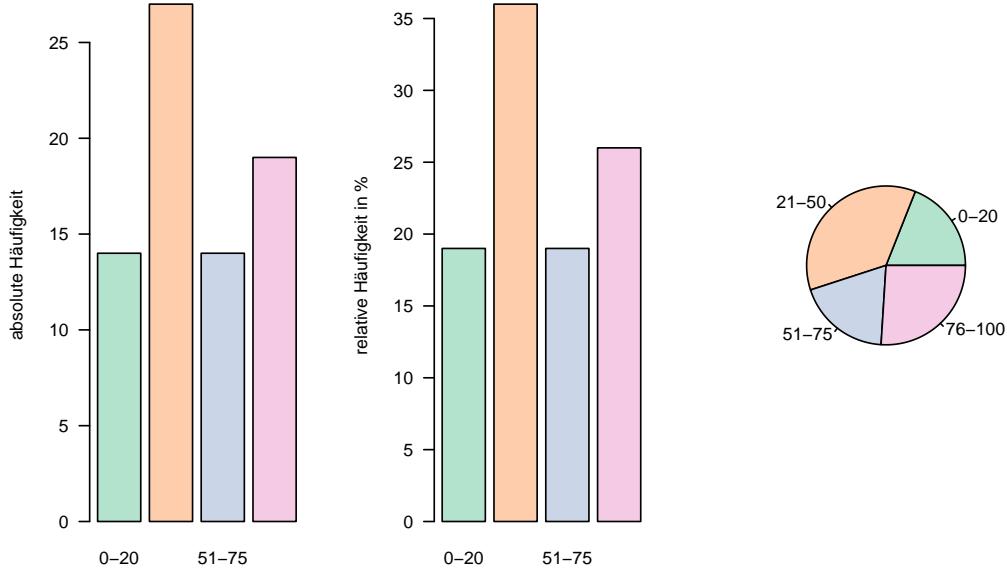
- n_i die **absolute Häufigkeit** der Merkmalsausprägung a_i . Es gilt klarerweise $\sum_i n_i = n$.
- $h_i = \frac{n_i}{n}$ die **relative Häufigkeit** der Merkmalsausprägung a_i . Die relativen Häufigkeiten summieren sich immer zu eins auf, $\sum_i h_i = 1$, da alle relativen Häufigkeiten aller Kategorien zusammen 100

Die entsprechende Visualisierung von Häufigkeiten erfolgt durch die Umrechnung von absoluten und relativen Häufigkeiten in Längen bei **Balkendiagrammen** und von relativen Häufigkeiten in Winkelanteile bei **Tortendiagrammen**. Hier ist zu beachten, dass das menschliche Auge Unterschiede von Längen viel besser unterscheiden kann als Unterschiede zwischen Winkeln, weshalb so gut wie immer Balkendiagrammen die bessere Wahl zur Visualisierung von Häufigkeiten sind!

Zusätzlich zu absoluten und relativen Häufigkeiten unterscheidet man zwischen kumulierten und nichtkumulierten Häufigkeiten.

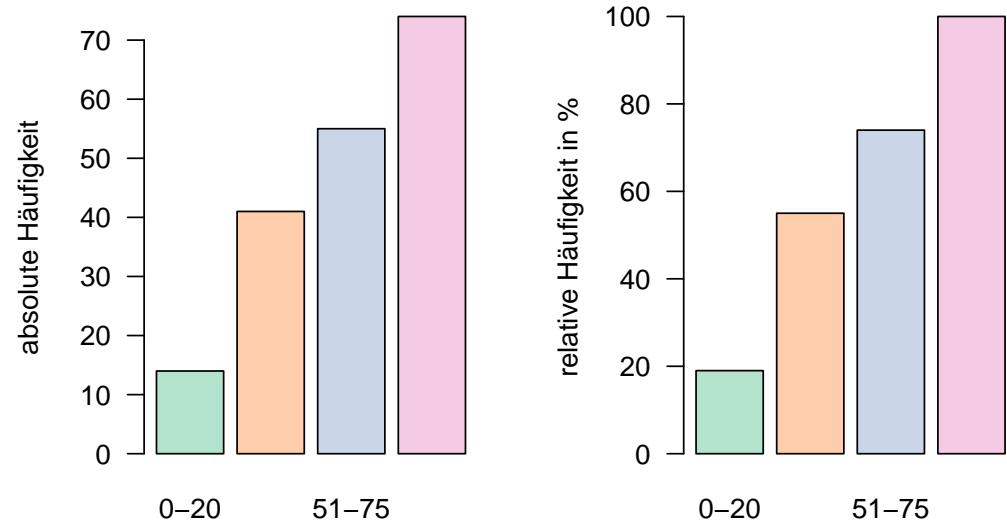
- **Nichtkumulierte Häufigkeiten**, n_i und $h_i = \frac{n_i}{n}$, werden je Kategorie angeben. Die Summe aller Einzelhäufigkeiten ist entweder die Stichprobengröße n bei absoluten oder 100 % =1 bei relativen Häufigkeiten.

	0-20	21-50	51-75	76-100
absolute Häufigkeiten	14	27	14	19
relative Häufigkeiten in %	19	36	19	26



- **Kumulative Häufigkeiten**, $N_i = \sum_{j \leq i} n_j$ und $H_i = \sum_{j \leq i} \frac{n_j}{n}$, können nur für geordnete Kategorien gebildet werden. Entsprechend der Anordnung ist die Häufigkeit der folgenden Kategorie die Häufigkeit der Kategorie addiert zu den Häufigkeiten aller vorhergehender Kategorien, sodass die letzte Kategorie die Stichprobengröße n bei absoluten oder 100 % = 1 bei relativen Häufigkeiten als kumulative Häufigkeit hat.

	0-20	21-50	51-75	76-100
absolute Häufigkeiten	14	41	55	74
relative Häufigkeiten in %	19	55	74	100



Die **Wahrscheinlichkeitsverteilungsfunktion**, auch **Wahrscheinlichkeitsdichtefunktion** (engl. probability density function *PDF*), der relativen Häufigkeiten

$$f(x) = \begin{cases} h_i & x = a_i \\ 0 & \text{sonst} \end{cases}$$

beschreibt, mit welcher Häufigkeit die Kategorie a_i vorkommt. Sie wird graphisch durch die obenstehenden Balkendiagramme dargestellt.

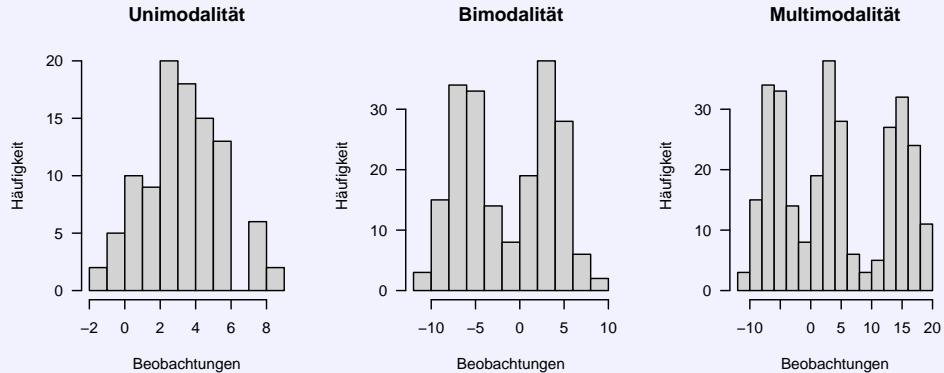
Eigenschaften von quantitativen Messungen

Wir unterscheiden unterschiedliche Eigenschaften von Messungen und ihrer zugrundeliegender Verteilung von Daten, welche wir durch statistische Schätzwerte einzuschätzen versuchen:

- **Lage** - "wo befinden sich die Daten?", "wo liegt ihr Zentrum?"
- **Streuung** - "wie stark streuen die Daten um ihr Zentrum?", "wie stark schwanken die Daten in ihrem Lagebereich?"
- **Symmetrie und Schiefe** - liegen die Werte symmetrisch um das Zentrum?", "gibt es deutlich mehr auf einer Seite als auf der anderen?"
- **Gewicht in den Rändern** - liegen viele Werte weit weg vom Zentrum?", liegen deutlich mehr Werte weit weg vom Zentrum als wir es bei Normalverteilung erwarten würden?"

Lageschätzer

- Der **Modus** oder **Modalwert** ist der Wert mit der höchsten Wahrscheinlichkeit der zugrundeliegenden Verteilung. Bei einer diskreten oder kategorialen Stichprobe ist es die am häufigsten vorkommende Kategorie. Dieser Wert muss nicht eindeutig sein, es kann auch mehrere Modi geben, was als Multimodalität im Unterschied zu nur einem Modus als Unimodalität bezeichnet wird.



- Der **arithmetische Mittelwert** $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ist der bekannteste Lageschätzer und sehr häufig in Verwendung.
- Der **Median** ist der in der Mitte befindliche Wert der Stichprobe im Sinne des 50 %-Quantils, also dass 50 % der Daten kleiner oder gleich dem Median und 50 % der Daten größer oder gleich dem Median sind. Bei einer ungeradzahligen Stichprobe ist er exakt der Wert in der Mitte, bei einer geradezahligen Stichprobe ist er der Mittelwert der beiden Werte in der Mitte.

Das **Problem des arithmetischen Mittelwerts** ist, dass er **sensitiv** gegenüber Asymmetrie, Ausreißern (Einzelwerte, deren Verhalten sich vom Verhalten der restlichen Daten unterscheidet) und schweren Rändern (systematischen Beobachtungen weit weg von der Mitte) ist und durch solche Werte verzerrt wird. Im Unterschied zum Mittelwert ist der **Median robust** gegenüber Asymmetrie, Ausreißern (Einzelwerte, deren Verhalten sich vom Verhalten der restlichen Daten unterscheidet) und schweren Rändern (systematischen Beobachtungen weit weg von der Mitte) und wird selbst von 50 % fehlerhaften Werten nur geringfügig beeinträchtigt.

Betrachten wir das mit einem konkreten Beispiel. Wir haben Messungen von Substanzwerten in mg: 2.301 4.571 3.814 3.647 4.132

Der arithmetische Mittelwert der Messwerte beträgt 3.693 und der Median 3.814. Durch einen Exceleinlesefehler wurde ein Dezimalpunkt der englischen Notation als Tausenderpunkt der deutschen Notation eingelesen.

2.301 mg, 4.571 mg, 3.814 mg, 3.647 mg, 4132 mg

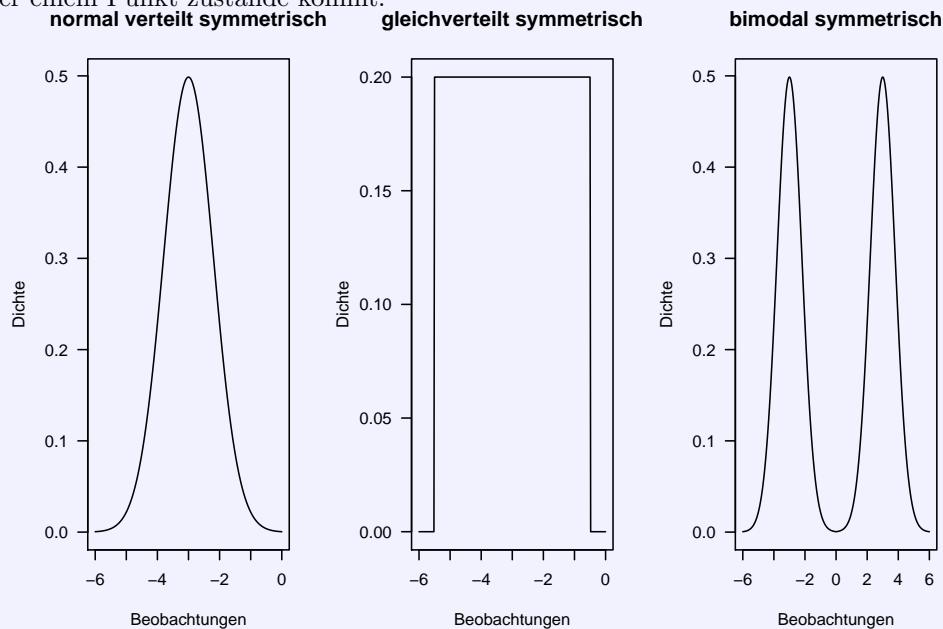
Der arithmetische Mittelwert der Messwerte beträgt 829.2666 und der Median 3.814. Diese Verschiebung des Mittelwerts bedeutet, dass er gegenüber jedem einzelnen Messwert anfällig ist. Der Median hingegen ist nicht betroffen, da der mittlere Wert nicht betroffen ist, was sein Robustheit ausmacht.

Streuungsschätzer

- Die **Standardabweichung** $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ und die Wurzel aus dieser Zahl, die **Standardabweichung** σ sind die Streuungsschätzer, die die Schwankung der Werte bezogen auf den arithmetischen Mittelwert messen. Wie der Mittelwert ist die Varianz und Standardabweichung *sensitiv* gegenüber abweichenden Werten.
- Die **Interquartilsdistanz** misst den Abstand zwischen dem 25 %-Quantil und dem 75 %-Quantil, also wie weit die mittleren 50 % der Daten auseinander liegen. Wie der Median, das 50 %-Quantil, ist die Interquartilsdistanz *robust*.
- Die **Spannweite** $x_{\max} - x_{\min}$ misst den Abstand zwischen dem größten und kleinsten Messwert und schätzt damit ab, welche Größenordnung von Werten von den Messungen insgesamt "überspannt" wird.

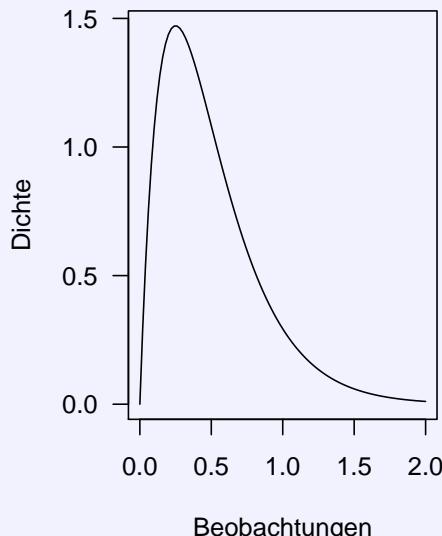
Symmetrie und Schiefe

Eine Funktion wird als **symmetrisch** bezeichnet, wenn sie durch Spiegelung eines Teilverlaufs an einer Achse oder einem Punkt zustande kommt.

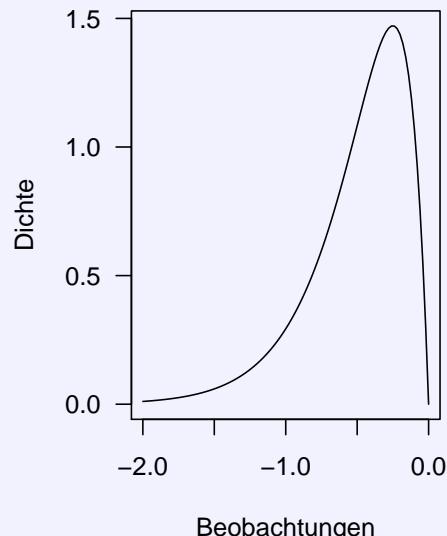


Wenn die Symmetrie zerstört wird, weil die Verteilung der Daten in einer Richtung „steiler“ und in der anderen Richtung „*schief auslaufend*“ verläuft, dann spricht man von einer schießen Verteilung. Je nachdem in welche Richtung der lange Rand schief ausläuft, heißt die Verteilung „**rechtsschief**“ oder „**linksschief**“.

rechtsschief



linksschief



In der Wahrscheinlichkeitsrechnung gibt es Funktionen, welche exakt symmetrisch sind, wie die Normalverteilungsdichtefunktion oder die Dichte der Gleichverteilung oder t-Verteilung.

In der Statistik kann beim Umgang mit realen Daten niemals exakte mathematische Symmetrie erreicht werden, sondern nur annähernde Symmetrie von Werteverteilungen. Erst wenn diese annähernde Symmetrie offensichtlich verletzt wird, spricht man von Schiefe in den Daten. Diese Eigenschaft wird anhand graphischer Darstellungen der Dichtefunktion, wie Histogramm oder Boxplot, ermittelt.

Kurtosis oder das Gewicht in den Rändern

Kurtosis misst, wie stark die Daten über das Konzept der Streuung hinaus von ihrem Zentrum abweichen, was man auch Gewicht der Ränder der Verteilung nennt. Als **Exzess-Kurtosis** wird dabei die Abweichung dieses Gewichts der Ränder vom Gewicht der Ränder der Normalverteilung als Referenz ermittelt.

Unterscheidung der Typen von Schätzern

- **Quantilsschätzer**

Quantilsschätzer Das α -**Quantil** (engl. α -quantile) x_α für ein beliebiges α aus $[0, 1]$ ist jener Wert, unter welchem ein Anteil α aller n Messwerte liegt,

$$\mathbb{P}[X \leq x_\alpha] = \alpha.$$

Die prominentesten Quantile sind der Median (50 %-Quantil) und das 1. und 3. Quartil (25 %- und 75 %-Quantil), sowie das Minimum (0 %-Quantil) und Maximum (100 %-Quantil).

Allen Quantilen ist die Eigenschaft gemeinsam, dass sie nicht nur für metrischen Daten, sondern auch für ordinale Daten als Schätzer verwendet werden können. Die Quartile, Median (50 %-Quantil) und das 1. und 3. Quartil (25 %- und 75 %-Quantil), sind insbesondere als robuste Schätzwerte von Dateneigenschaften wie Lage und Streuung metrischer Daten beliebt.

- **Momentenschätzer**

Wir haben die Momentenschätzung bereits in der Wahrscheinlichkeitstheorie angesprochen, verwenden hier die Konzepte aber wieder und wiederholen daher den Absatz.

Momentenschätzer

Als **n-tes Moment** bezeichnet man in der Statistik und Wahrscheinlichkeitsrechnung die Summe der n-ten Potenz der Datenwerte gewichtet mit ihrer Wahrscheinlichkeit, mathematisch

$$\mathbb{E}[X^n] = \begin{cases} \int f(x) \cdot x^n dx & \text{mit der Dichtefunktion } f(x) \\ & \text{der stetigen Zufallsvariable } X \text{ bzw.} \\ \sum_{i=0}^{\infty} p(x_i) \cdot x_i^n & \text{mit der Wahrscheinlichkeitsverteilung } p(x_i) \\ & \text{der diskreten Zufallsvariable } X. \end{cases}$$

Das bekannteste Moment einer Verteilung ist ihr 1. Moment, der **Erwartungswert** $\mathbb{E}[X] = \mu$. Sein empirischer Schätzwert ist der *arithmetische Mittelwert*.

Der **Erwartungswert** bzw. der **Mittelwert** einer Verteilung unterscheidet sich daher in seiner Berechnung ja nachdem, ob diese stetig oder diskret ist. Im Falle einer diskreten Wahrscheinlichkeit haben die einzelnen Ausgänge, also die Kategorien, tatsächliche Häufigkeiten und Wahrscheinlichkeiten, daher kann der Kategoriewert mit dieser Häufigkeit multipliziert werden und die dadurch entstehenden gewichteten Teilwerte werden aufaddiert.

Ein typisches Beispiel dafür ist, wenn ein Messgerät nur ganzzahlige Ergebnisse zurückliefert, was häufig bei einer Hauswaage oder der Altersermittlung der Falle ist. Bei der folgenden Altersverteilung von Studierenden eines Bachelorstudiengangs

Alter	Häufigkeit
18	5
19	17
20	9
21	6
22	3
27	1

ermittelt man den arithmetischen Mittelwert durch Gewichtung der Alterswerte durch die entsprechenden Häufigkeiten ihres Auftretens, also $18 \cdot \frac{5}{41} + 19 \cdot \frac{17}{41} + 20 \cdot \frac{9}{41} + 21 \cdot \frac{6}{41} + 22 \cdot \frac{3}{41} + 27 \cdot \frac{1}{41} = 20$ Jahre.

Im Falle unterschiedlicher Messungen, welche nicht als Kategorien zusammengefasst bzw. aufgefasst werden, entspricht die Berechnung des Mittelwerts einer Stichprobe genau der Berechnung des 1. Moments, wenn man als Verteilung eine Gleichverteilung über alle Beobachtungen annehmen würde. Das bedeutet, dass als Laplace-Wahrscheinlichkeit jedes beobachteten Wertes einer Stichprobe der Größe n die relative Häufigkeit $\frac{1}{n}$ verwendet wird. Dadurch entsteht die Formel des arithmetischen Mittelwerts

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

Der arithmetische Mittelwert von Messungen von Einwaagegewichten einer Probe 5.01 mg, 5.02 mg, 5.15 mg, 5.12 mg, 5.05 mg, 4.98 mg, 5.17 mg, 4.96 mg, 5.08 mg, 4.92 mg, 5.09 mg, 5 mg ergibt sich als 60.55 dividiert durch 12 = 5.05 mg.

Als **n-tes zentrales Moment** bezeichnet man in der Statistik und Wahrscheinlichkeitsrechnung die Summe der n-ten Potenz der Differenz der Datenwerte vom Erwartungswert $\mathbb{E}[X] = \mu$ gewichtet mit ihrer Wahrscheinlichkeit, mathematisch

$$\mathbb{E}[(X - \mu)^n] = \begin{cases} \int f(x) \cdot (x - \mu)^n dx & \text{mit der Dichtefunktion } f(x) \\ & \text{der stetigen Zufallsvariable } X \text{ bzw.} \\ \sum_{i=0}^{\infty} p(x_i) \cdot (x_i - \mu)^n & \text{mit der Wahrscheinlichkeitsverteilung } p(x_i) \\ & \text{der diskreten Zufallsvariable } X. \end{cases}$$

Das bekannteste zentrale Moment einer Verteilung ist das 2. zentrale Moment, die **Varianz** $\mathbb{E}[(X - \mu)^2]$. Ihr Schätzwert ist die *Stichprobenvarianz*. Die Varianz quadriert die Abweichungen vom Erwartungswert, wodurch das Resultat nicht mehr in seiner Größenordnung mit dem Erwartungswert vergleichbar ist. Um diesen Mangel zu beheben, zieht man die Wurzel aus der Varianz, was die Standardabweichung ergibt, welche wieder dieselbe Einheit und Skala wie der Erwartungswert hat.

Die wichtigsten empirischen Momente

1. Moment	Lage	Erwartungswert	\bar{x}	$= \frac{1}{n} \sum_{i=1}^n x_i$
2. zentrales Moment	Streuung	Varianz	s_n^2	$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
3. zentrales Moment	Symmetrie	Schiefe	$skew$	$= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}}$
4. zentrales Moment	Rändergewicht	(Exzess) Kurtosis	$kurt$	$= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$

Konkrete Berechnung der wichtigsten Schätzer von Lokation und Variation in R

Als Beispiel wollen wir die Lageschätzer und Streuungsschätzer für die Messungen der Probenentnahme einer biotechnologischen Anlage, gemessen in mg, ermitteln.

Messungen

1	5.67
2	4.78
3	6.91
4	5.48
5	4.89
6	5.22
7	5.63
8	5.88
9	5.71
10	4.80
11	6.19
12	5.28
13	4.91
14	6.12
15	5.32
16	5.48

Dafür werden die Datenwerte als Erstes in einen Vektor in R gespeichert.

```
datenwerte<-c(5.67,4.78,6.91,5.48,4.89,5.22,5.63,5.88,5.71,4.80,6.19,5.28,4.91,6.12,5.32,5.48)
```

Anschließend berechnen wir die unterschiedlichen Schätzwerte:

Der Mittelwert wird mithilfe der Funktion `mean` berechnet.

```
mean(datenwerte)
```

```
## [1] 5.516875
```

Der Mittelwert beträgt also 5.52 mg.

Der Median wird mithilfe der Funktion `median` berechnet.

```
median(datenwerte)
```

```
## [1] 5.48
```

Der Median beträgt also 5.48 mg.

Wir sehen hier, dass die Werte von Mittelwert und Median nah beisammen liegen, also kein starker Verzerrungeffekt von Ausreißern oder Asymmetrie vorhanden ist.

Die Varianz wird mithilfe der Funktion `var` berechnet.

```
var(datenwerte)
```

```
## [1] 0.3311296
```

Die Varianz beträgt also 0.33 mg^2 .

Die Standardabweichung wird mithilfe der Funktion `sd` berechnet.

```
sd(datenwerte)
```

```
## [1] 0.5754386
```

Die Standardabweichung beträgt also 0.58 mg.

Die Interquartilsdistanz wird mithilfe der Funktion `IQR` berechnet.

```
IQR(datenwerte)
```

```
## [1] 0.61
```

Die Interquartilsdistanz beträgt also 0.61 mg.

Wir sehen, dass Interquartilsdistanz und Standardabweichung auch von einer ähnlichen Größenordnung sind, was zeigt, dass es keine massive Beeinträchtigung durch Ausreißer oder Schiefe gibt.

Characteristic Dateneigenschaften

1. **Modalität** Haben die Daten ein einziges Zentrum oder bestehen sie aus unterschiedlichen Teilen mit verschiedenen Eigenschaften? Die Anzahl der "Gipfel"(peaks") bestimmt die Modalität.
2. **Lage (Lokation)** Die Lage des Zentrums der Daten wird durch einen representativen oder mittleren Wert beschrieben, was nur sinnvoll ist, wenn es nur ein Zentrum (unimodal) der Daten gibt. Bei multimodalen Daten wird dies für jeden Teil separat bestimmt.
3. **Schwankung (Variation)** Ein Maß für Variabilität von Messungen bezogen auf einen zentralen Lageparameter sind Schwankungs- oder Variationsschätzer. Auch diese sind nur sinnvoll für unimodale Daten und werden für multimodale Daten separat für jeden Teil ermittelt.
4. **Verteilung (Symmetrie und Ränder)** Wir betrachten bei Form und Verhalten zwei Hauptaspekte: der erste Aspekt ist, ob Daten symmetrisch oder asymmetrisch sind. Außerdem betrachten wir "Spitzigkeit" oder in anderen Worten, ob viel oder wenig Gewicht in den Rändern liegt, also mehr oder weniger Werte weit weg vom Zentrum sind als bei einer Normalverteilung.
5. **Ausreißer** Werte der Stichprobe, die ein systematisch anderes Verhalten als der Rest der Daten aufweisen, werden als Ausreißer bezeichnet. Oft aber nicht immer sind das Werte, die weit weg von der Mehrheit der anderen Beobachtungen liegen. Manchmal haben Werte weit weg vom Zentrum aber einige andere Werte in der Nähe, die sich ähnlich verhalten, was sie alle nicht zu Ausreißern sondern zu einem sich anders verhaltenden Datenteil macht (s. Multimodalität).
6. **Zeit** Wenn sich Eigenschaften der Daten in Abhängigkeit von der Zeit verändern, spricht man von Zeitreihen oder zeitabhängigem Verhalten. Das inkludiert, dass man Eigenschaften in einen generellen Trend, ein sich wiederholendes saisonales Verhalten und einen zufällig verteilten zeitunabhängigen Rest aufteilen kann.

Characteristic properties of data

1. **Modality.** Do the data have a single center or do they consist of several different parts with different characteristics? The number of peaks" determines the modality.
2. **Center.** The location of the middle of the data illustrated by a representative or average value. This makes sense for unimodal data or for multimodal data in each part separately.
3. **Variation.** A measure of the variability of measurements with respect to its central location value. This makes sense for unimodal data or for multimodal data in each part separately.
4. **Distribution.** How the spread of the data is shaped and behaves. Possibly, we consider whether data are symmetric or asymmetric. In addition we consider "peakedness" or in other words the weight of the tails which is the amount of data further away from the center than would be expected for Gaussian data.
5. **Outliers.** Values of the sample showing a behaviour which differs from the rest of the sample, often but not always are those values located very far away from the vast majority of the other sample values. Sometimes, values far from the center have close neighbours which have a similar behaviour. Then, they are not necessarily outliers, but maybe a different mode (subpart) of the data.
6. **Time.** Changing characteristics of the data over time. This includes a basic trend, seasonal behaviour and random errors.

Erweiterte Inhalte - Weit aus mehr Lage- und Streuungsschätzer

Location Estimators

Different estimators for the center or location of unimodal data exist.

{Important measures of location}		
	Sample	R
(Arithmetic) mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	mean
weighted mean	$\bar{x} = \sum_{i=1}^n w_i x_i$	weighted.mean
trimmed mean	$\bar{x} = \frac{1}{n} \sum_{i=q_{trim}}^{q_{1-trim}} x_{(i)}$	mean(trim=p)
Geometric mean	$\sqrt[n]{\prod_{i=1}^n x_i}$	exp(mean(log(x)))
Harmonic mean	$\frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$	1/mean(1/x)
Median	middle value of ordered data	median
Mode	value with the greatest frequency	mode
Midrange	$\frac{\max x_i + \min x_i}{2}$	-

The arithmetic mean is identical with the estimator of the first moment of the distribution. It is therefore the best motivated estimator mathematically which includes every single observation with equal weight $\frac{1}{n}$. This strength of including all values equally is also its weakness, as the **arithmetic mean** is therefore **sensitive against outliers, heavy tails and asymmetry**. For this reason the **weighted arithmetic mean** which assigns individual weights to all observations which must sum up to 1. This allows downweighting extreme observations and upweighting central ones. An automatized version of this is the **trimmed arithmetic mean** which automatically trims off the largest and smallest $p \cdot 100\%$ of the data and is therefore more robust against outliers or heavy tails.

The **geometric mean** and **harmonic mean** are related to the arithmetic mean. In general the inequality holds:

$$\frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \leq \sqrt[n]{\prod_{i=1}^n x_i} \leq \frac{1}{n} \sum_{i=1}^n x_i$$

Their applications however are more specific. The **geometric mean** is suitable for all scenarios where numbers would be multiplied instead of added. As with the corresponding geometric series, this applies to all kinds of growth scenarios. Fitting to the current times, we can look at CoVID-19 growth rates per day which are 21%, 20%, 18%, 19%, 17%, 19% and 20% within the first week without government measures. The amount of infected people based on these growth rates and the number of infected people before measuring the growth N_0 would be calculated as

$$N_0 \cdot 1.21 \cdot 1.20 \cdot 1.18 \cdot 1.19 \cdot 1.17 \cdot 1.19 \cdot 1.20.$$

Therefore, the average growth rate in such a scenario should be calculated as

$$\sqrt[7]{1.21 \cdot 1.20 \cdot 1.18 \cdot 1.19 \cdot 1.17 \cdot 1.19 \cdot 1.20}$$

instead of

$$\frac{1.21 + 1.20 + 1.18 + 1.19 + 1.17 + 1.19 + 1.20}{7}$$

which is exactly the geometric mean instead of the arithmetic one.

The **harmonic mean** is meant for all kinds of scenarios which involve ratios, such as calculating the average velocity $v = \frac{s_1+s_2}{t_1+t_2}$ based on measured times t_i and distances s_i .

The **median** is the **most robust** measurement of location available. Its breaking point is 0.5 which means that up to half of the data can be affected by measurement errors, yet the median will still provide a resonable location estimate.

A mode is a local maximum of the density curve. As we have seen before with unimodality (“one mode”) and multimodality (“several modes”), there can be more than one mode in the data.

Example for applying Location estimates

Which location measure do we use in which situation?

- A biologist wants to calculate the average growth rate of a bacteria culture on different plates which showed the following growth rates, 3%, 2%, 1.2%, 0.3%, 0.9%, 1.6%.

As these plates are separated units, their growth happens independent of each other. Therefore, the average growth can be calculated as the mean, as all measurements are of equal order of magnitude, or the median.

- A biologist wants to calculate the average growth rate of a bacteria culture on different plates sequentially which showed the following growth rates, 3%, 2%, 1.2%, 0.3%, 0.9%, 1.6%.
- A server tracks the times until response for user logins. Most response times are within milliseconds (ms), whereas some responses take several minutes. A informatician wants to calculate the average server response time.
- A bioinformatician wants to calculate the average in several steps of PCR for sequentially amplifying genetic material which showed the following growth rates, 3%, 2%, 1.2%, 0.3%, 0.9%, 1.6%.
- What is the average BMI of 5 persons (22.21,19.45,23.65,20.40,21.37)?
- A company wants to know about its average losses due to their employees' sickness leaves.
- An employees argues with his personnel manager about his average days on leave (due to sickness, vacation, maternity leave, etc.).

Measures of Variation

(More or less) important measures of variation

	Sample	Population	R
Range	$\max x_i - \min x_i$		range
Variance	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$	var
Standard deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$	sd
MAD (from the mean)	$\frac{1}{n} \sum_{i=1}^n x_i - \bar{x} $	$\frac{1}{N} \sum_{i=1}^N x_i - \bar{x} $	mad
MAD from the median	$\frac{1}{n} \sum_{i=1}^n x_i - \text{med } x_i $	$\frac{1}{N} \sum_{i=1}^N x_i - \text{med } x_i $	-
MedMed IQR	$\text{med } x_i - \text{med } x_i $ $Q_{0.75} - Q_{0.25} = \text{Box length}$		- IQR
Coefficient of variation	$CV = \frac{s}{\bar{x}}$	$CV = \frac{\sigma}{\mu}$	-

We measure variation in order to gain better insights towards our data's quality and precision. Saying that a server has an average response time of 10 ms is worthless, unless you state precisely how much this value can

vary. If the variation is containing mainly values between 9 ms and 11 ms you would consider this reasonable, whereas values which vary between 1 ms and 19 ms appear less reliable, although their location estimate stays the same.

The measure of variation must always fit to the corresponding measure of location. Therefore, any measures which refers to distance from the arithmetic mean \bar{x} can only be combined with the mean, but not the median. In the same way, measures which correspond to the median or quartiles can only be combined with the median.

Schätzung und Darstellung quantitativer Messungen

Zählvariablen

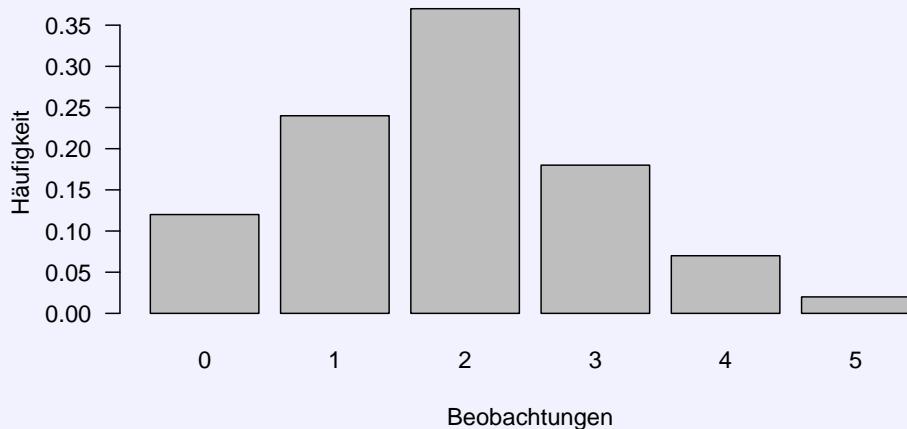
Zählvariablen beschreiben Zählungen von Anzahlen von gewünschten Ereignissen (“Erfolgen”), welche bei 0 als Wertebereich beginnen und 0 auch tatsächlich als sinnvoll vorkommende Merkmalsausprägung haben.

Die **Wahrscheinlichkeitsverteilungsfunktion**, auch **Wahrscheinlichkeitsdichtefunktion** (engl. probability density function *PDF*) genannt, der relativen Häufigkeiten

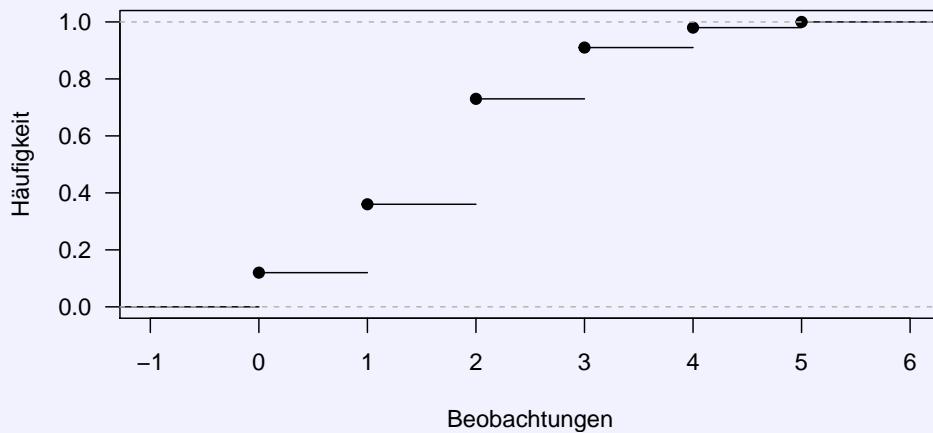
$$f(x) = \begin{cases} h_i & x = i \\ 0 & \text{sonst} \end{cases}$$

beschreibt, mit welcher Häufigkeit der gezählte Wert i vorkommt.

Wahrscheinlichkeitsverteilung



kumulative Verteilungsfunktion



Die Funktion der kumulativen Häufigkeiten (Summenhäufigkeiten) wird graphisch als **empirische kumulative Verteilungsfunktion** (engl. empirical cumulative distribution function ECDF) $F_n(x)$ dargestellt. Diese Funktion ermittelt, welcher Anteil der Beobachtungen kleiner oder gleich der Anzahl i ist.

Die Schritthöhe jeder Stufe der entstehenden Treppenfunktion ist die relative Häufigkeit h_i der jeweiligen Anzahl. Daher hat die kumulative Verteilungsfunktion unterhalb der kleinsten beobachteten Anzahl den Wert 0 und ab der höchsten beobachteten Anzahl den Wert 1, da 100 % der Werte kleiner oder gleich diesem Wert i ist.

sind.

Mathematisch betrachtet ist die empirische Verteilungsfunktion nichts anderes als eine Stammfunktion der empirischen Dichte f .

Stetige numerische Variablen

Wir nehmen an, dass die Merkmalsausprägungen innerhalb eines Intervalls $[a, b)$ liegen. Bei stetigen Merkmalen ist es sehr unwahrscheinlich, dass sich auch nur zwei Messungen genau gleichen. Eine diskrete Dichtefunktion wie zuvor zu konstruieren ergibt somit keinen Sinn.

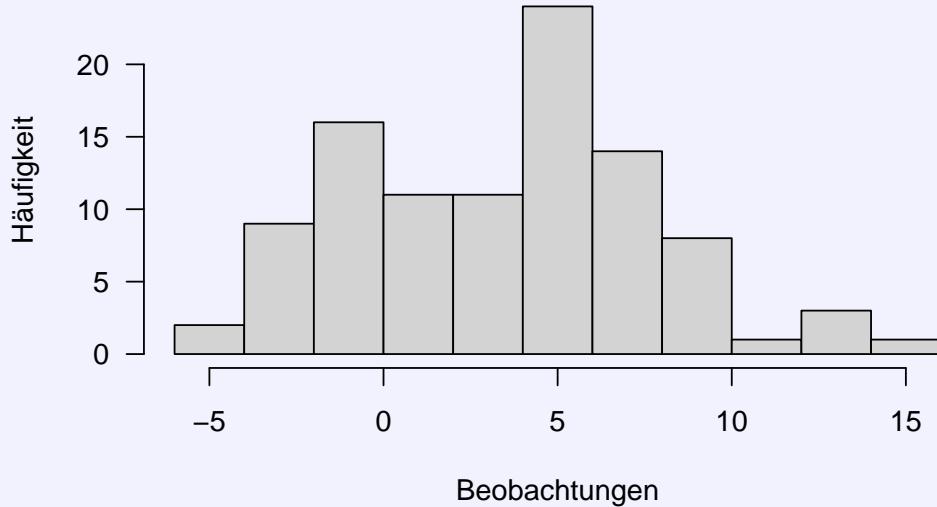
Um etwas Ähnliches zu erreichen, fasst man die Messungen in Klassen zusammen, d.h. man zerlegt das Intervall nach Ermessen in kleinere Teilintervalle $[a, b) = [a_0 = a, a_1) \cup [a_1, a_2) \cup \dots \cup [a_{k-1}, a_k = b)$ und betrachtet dann die dadurch entstehenden Häufigkeiten $n_i =$ die Anzahl der Messergebnisse im Intervall $[a_{i-1}, a_i)$.

Wir kennen bereits, dass wir mit diesen absoluten und relativen Häufigkeiten eine Wahrscheinlichkeitsdichtefunktion mit Balkendiagrammen darstellen können. Diese Balken werden ohne Abstand zwischen benachbarten Balken dargestellt und die graphische Darstellung heißt **Histogramm**. Dabei spielt die Einstellung der Klassenbreite, also wie breit die Intervalle $[a_{i-1}, a_i)$ sind, eine große Rolle für den optischen Eindruck. Empirisch erprobte Methoden lassen bei den meisten Softwarelösungen automatisch ein in 90 % der Fälle günstige Lösung wählen, während für die restlichen Fälle die Breite manuell angepasst werden muss.

{Histogramm}

Das Histogramm stellt die Dichtefunktion dar und erlaubt Unimodalität oder Multimodalität zu erkennen, indem man das Intervall mit der höchsten Wahrscheinlichkeit bzw. die Gipfel der Dichtefunktion sucht.

Histogramm

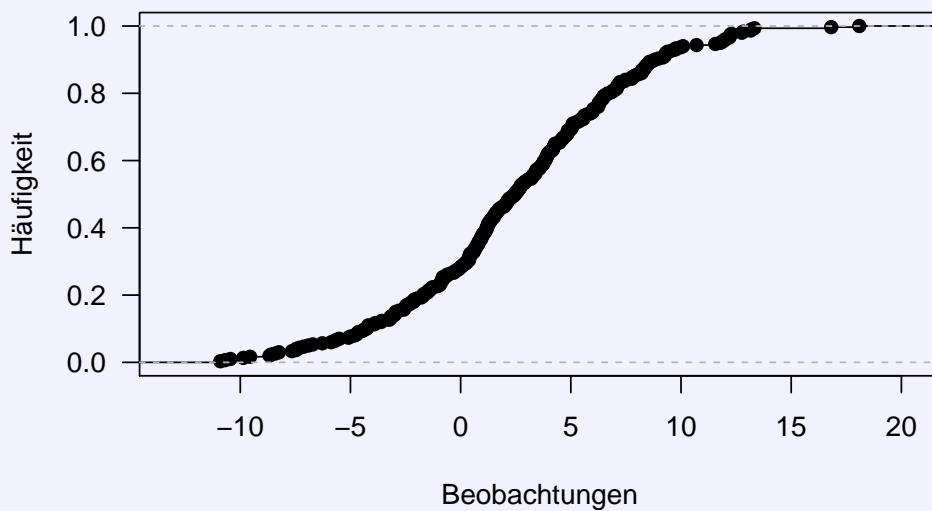


Das Histogramm ist die für das Auge beste graphische Umsetzung der Annäherung der Dichtefunktion und erlaubt etwa Eigenschaften wie die Modalität oder Symmetrie der Daten klarer als in anderen Darstellungen zu erkennen.

Empirische kumulative Verteilungsfunktion

Analog zu zuvor wird auch die **empirische kumulativen Verteilungsfunktion** (engl. empirical cumulative distribution function ECDF) $F_n(x)$ definiert als die Häufigkeit einen Wert kleiner oder gleich einer Zahl x aus dem Intervall $[a,b]$ zu beobachten.

empirische kumulative Verteilungsfunktion

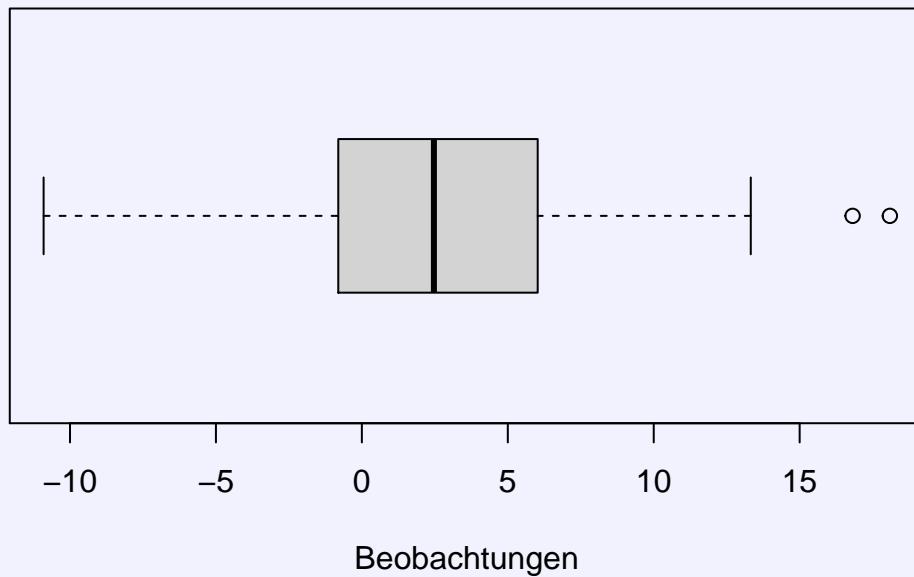


Die empirische Verteilungsfunktion wird auch als Quantilsfunktion bezeichnet, weil man bei ihr die Quantile auf der x-Achse zu den gegebenen Wahrscheinlichkeiten auf der y-Achse direkt ablesen kann.

Gemeinsam ergeben die 5 wichtigsten Quantile, der Median (50 %-Quantil) und das 1. und 3. Quartil (25 %- und 75 %-Quantil), sowie das Minimum (0 %-Quantil) und Maximum (100 %-Quantil), die graphische Darstellung des Boxplots. Wie bereits bei den Schätzern angesprochen, ist der **Boxplot robust**, da die **Quartile robust** sind. Dafür ist er so robust, dass er **Multimodalität nicht mehr erkennen** kann und von multimodalen Daten verzerrt wird.

{**Boxplot**} Der Boxplot stellt den Median als Mitte der Box, die Quartile als die beiden Enden der Box dar, was die mittleren 50 % der Daten klar kennzeichnet. Darüber hinaus werden durch die Whiskers, die maximal das 1,5-fache des Quartilsabstands umfassen, die zentralen ~95 % der Daten abgesteckt, wenn die Daten normalverteilt wären. Daher werden in Bezug auf diese Normalverteilungsannahme Werte außerhalb der Whiskers als potentielle Ausreißer bezeichnet. Wenn die Verteilung der Daten aber inherent schief oder von schweren Rändern geprägt ist, ist diese Ausreißereinteilung jedenfalls falsch.

Boxplot

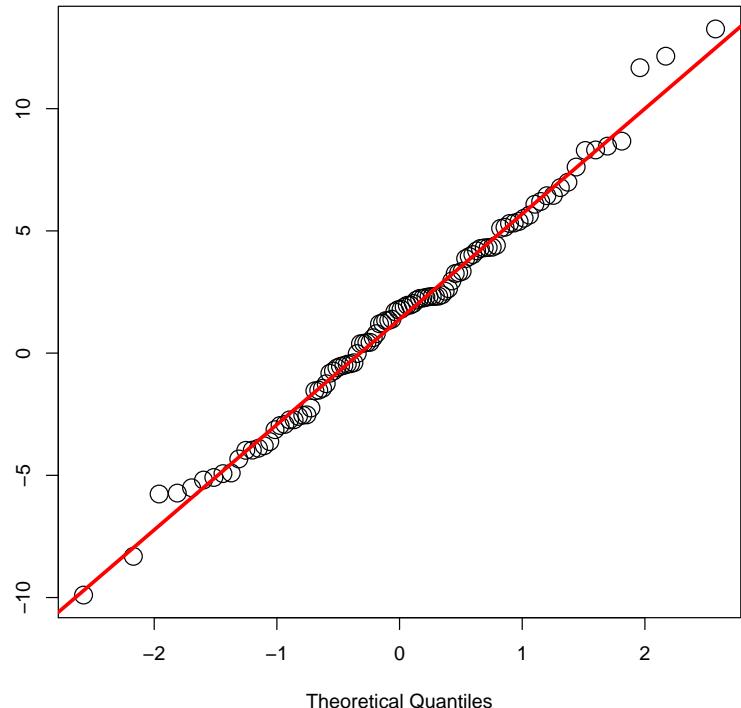


Quantil-Quantil-Plot (QQ-Plot)

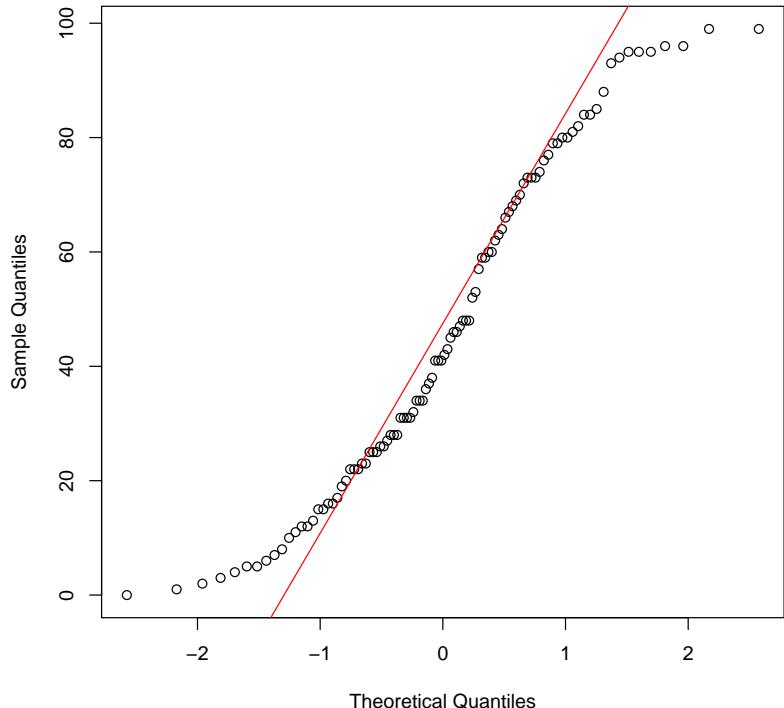
Quantil-Quantil-Plot (QQ-Plot) Ein **Quantil-Quantil-Plot (QQ-Plot)** ist ein Graph, der die Verteilung der Daten einer Stichprobe mit der Verteilung der Daten einer anderen Stichprobe oder einer theoretischen Verteilung der Datenwerte (z. B. Standardnormalverteilung) vergleicht.

```
qqnorm(x); qqline(x, col="red")
```

Normal Q–Q Plot



Normal Q–Q Plot



Fallbeispiele

Assessing normality Procedure for determining whether it is reasonable to assume that sample data are from a normally distributed population:

1. Normal quantile-quantile plot: Do the points lie reasonably close to a straight line or is there a *systematic pattern* that is not a straight-line pattern? (R: `qqnorm(x)` , `qqplot(x)`)

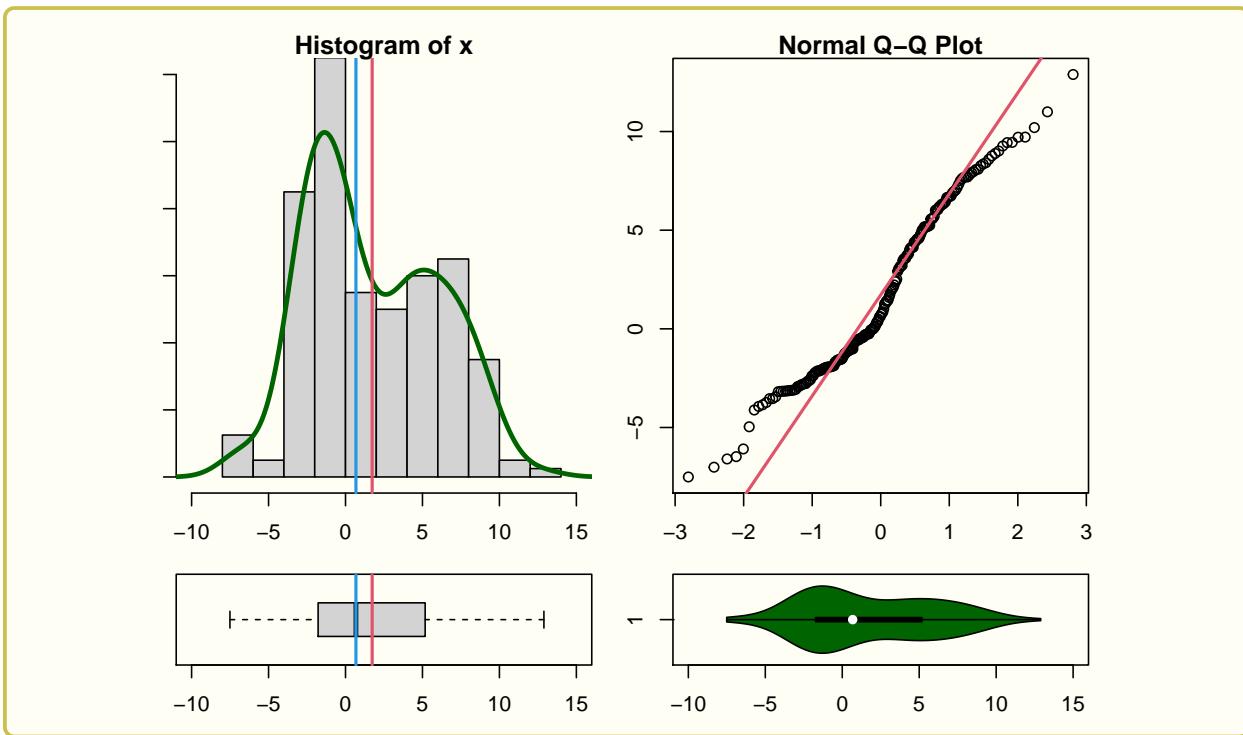
This requires some training of the eye to be able to assess normality, but it is more reliable than any test, as you can visually identify what was the problem for departing from normality.

2. We could also test for normality with a Shapiro-Wilk test. It measures whether the shape of the sample distribution is reasonably close to the normal distribution. However, every additional test we make in our analysis has its price (cf. multiple testing correction) (R: `shapiro.test(x)`)

Properties speaking against normality

- Multimodality - this is bad. Most test and models do not work for multimodal data. Here, we must continue with mixture models, classification or clustering.
- Outliers - this is good. The data would be normal with the exception of a few single values. Once detected, these outliers can be removed and we obtain normal data.
- Skewness - we are screwed or maybe not. Skewed data are definitely not normal. However, if they are right-skewed there is a chance that these data are log-normal and will become normal or much closer to normal when the logarithmic transformation is applied to the data.

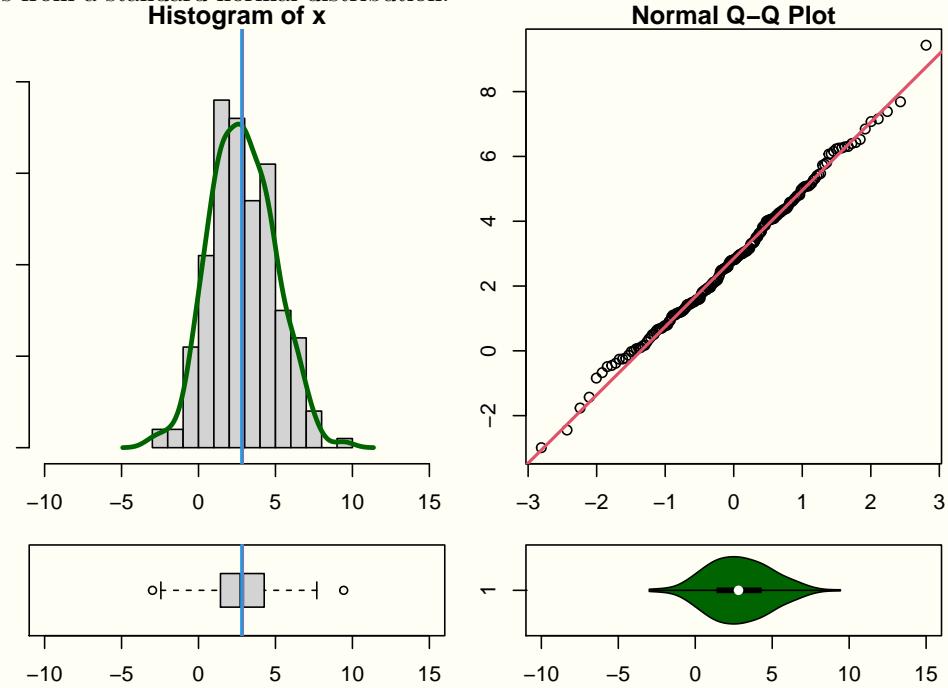
Multimodalität - Beispiel



Example: Assessing normality

Normal Data

100 draws from a standard normal distribution:



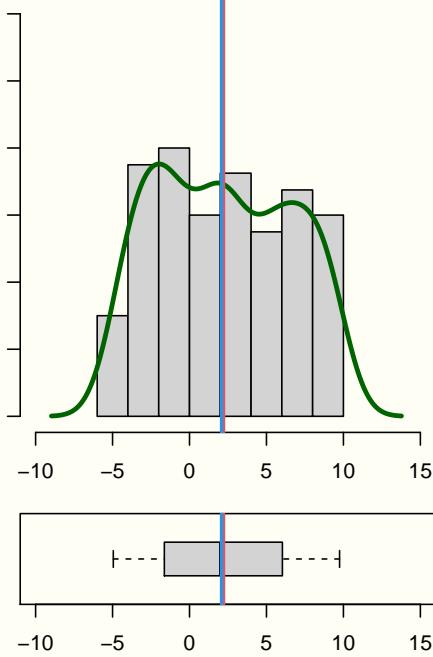
These data are unimodal centered around 0, as we can see from the histogram, density plot and QQ-plot. We can also observe that the data are symmetric from the histogram. The QQ-plot shows that both tails do not diverge from the reference line representing the normal distribution. Therefore, this data is normally distributed.

Example: Assessing normality

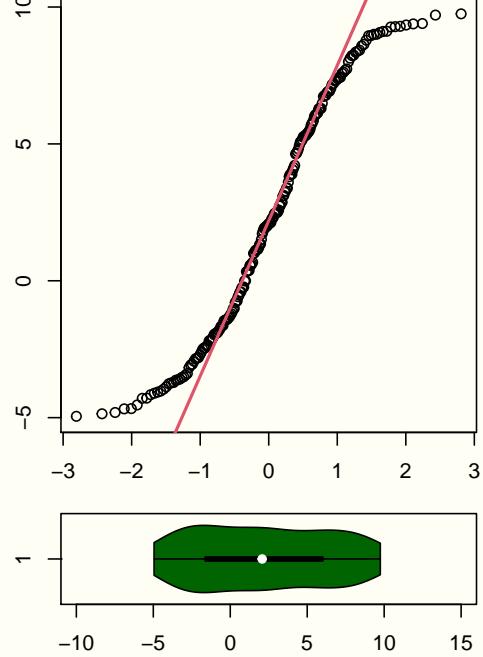
Uniform Data

100 draws from a uniform distribution on $[-3,3]$:

Histogram of x



Normal Q-Q Plot



-10 -5 0 5 10 15

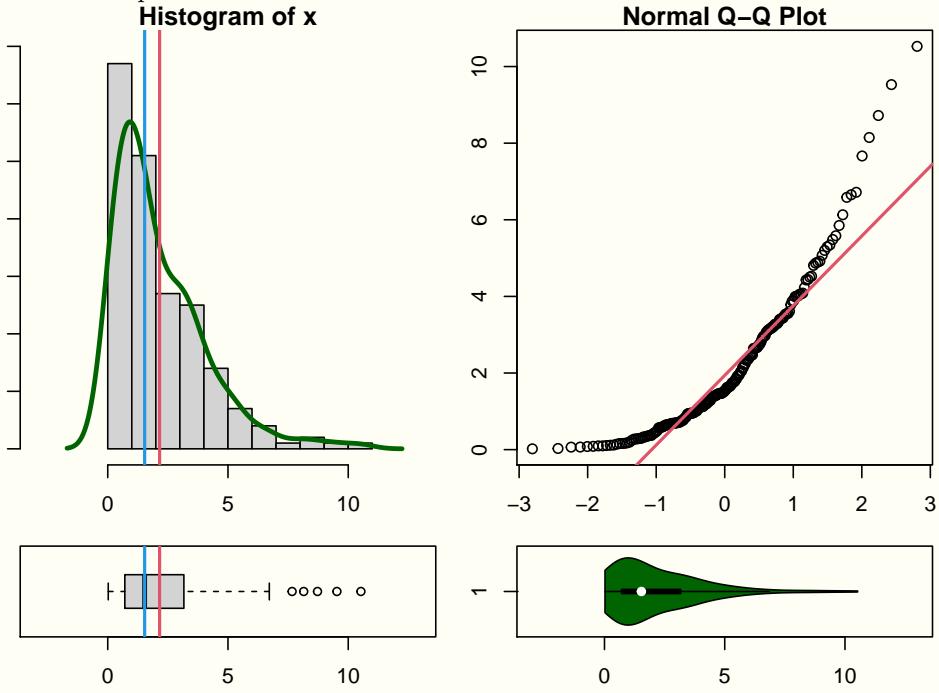
-3 -2 -1 0 1 2 3

These data are unimodal centered around 0, as we can see from the histogram, density plot and QQ-plot. We can also observe that the data are symmetric from the histogram. The QQ-plot shows that both tails diverge from the reference line representing the normal distribution. In both cases the tails are closer to the center than the normal tails would be which means that both tails are light tails. Therefore, this data is not normally distributed.

Example: Assessing normality

Exponential Data

100 draws from an exponential distribution:

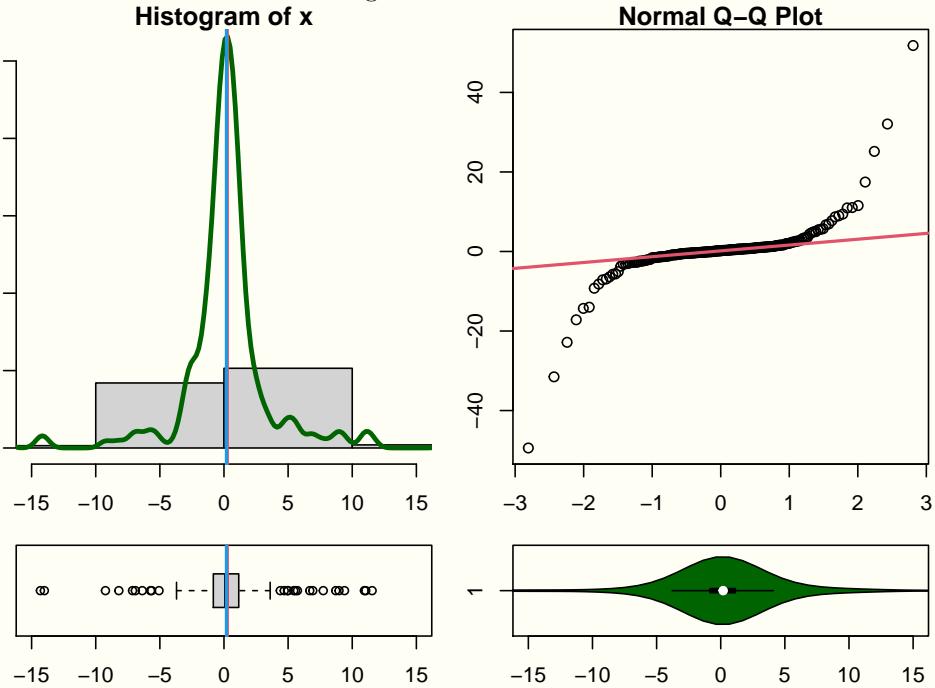


These data are unimodal or possibly bimodal, as we can see from the histogram, density plot and QQ-plot. The first mode is at -1, a second smaller mode at 3.5. One could also interpret the smaller mode a part of the extremely heavy tail on the right side of the distribution. We can also observe that the data are not symmetric from the histogram and QQ-plot. The QQ-plot shows that both tails diverge from the reference line representing the normal distribution. The left tail is closer to the center than the normal tails would be which means that it is a light tail. The right tail is much further away from the center than the normal tails would be which means that it is a heavy tail. Therefore, this data is not normally distributed.

Example: Assessing normality

Student's t Data

100 draws from a t distribution with 2 degrees of freedom:



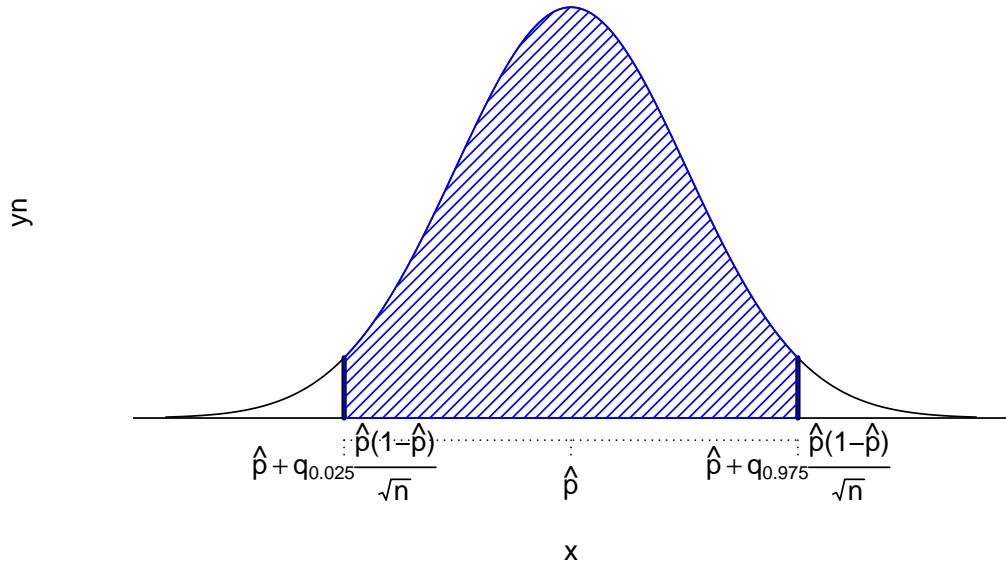
These data are unimodal centered around 0, as we can see from the histogram, density plot and QQ-plot. We can also observe that the data are symmetric from the histogram. The QQ-plot shows that both tails diverge from the reference line representing the normal distribution. In both cases the tails are further away from the center than the normal tails would be which means that both tails are heavy tails. Therefore, this data is not normally distributed.

Inferenzstatistik

Konzept der Konfidenzintervalle

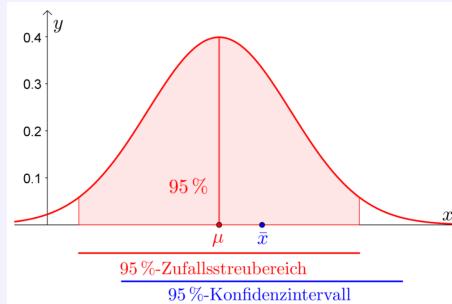
Konfidenzintervalle (= Vertrauensbereiche)

Wenn ein Wert (arithmetischer Mittelwert, Proportion, ...) aus einer Stichprobe geschätzt wird, um auf den zugrunde liegenden ‘wahren Wert’ des Zentrums, Anteils etc. Rückschlüsse zu ziehen, dann gibt der Konfidenzbereich den Bereich an, der diesen Wert mit einer Wahrscheinlichkeit von 95 % oder 99 %, allgemein $1 - \alpha$, überdeckt. Dann ist die Wahrscheinlichkeit, nicht zu überdecken, also den wahren Wert nicht zu enthalten, gleich α , etwa $\tilde{\alpha} = 5\%$.



Prädiktionsintervalle (= Vorhersagebereiche, Zufallsstrebereiche)

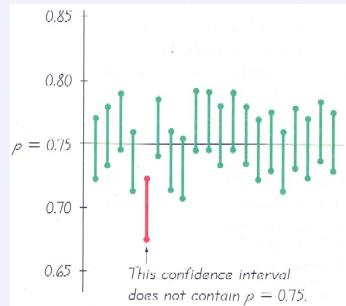
Wenn ein Wert (arithmetischer Mittelwert, Proportion, ...) aus einem bekannten Modell oder auf Basis bekannter gemessener Werte vorhergesagt werden soll, dann gibt der Vorhersagebereich den Bereich an, in dem sich der zu messende Wert mit einer Wahrscheinlichkeit von 95 % oder 99 % befindet.



Interpretationen von (95%) Konfidenzintervallen

- (wiederholte) Stichproben

“Würde mehrfach auf dieselbe Weise Stichproben gezogen werden, dann würden in 95% der Fälle die berechneten Konfidenzintervalle den wahren zugrunde liegenden (aber uns unbekannten) Wert überdecken.”



- Einzelstichprobe

“Mit 95%-iger Wahrscheinlichkeit enthält das ermittelte Konfidenzintervall den wahren zugrunde liegenden (aber uns unbekannten) Wert.”

- Akzeptanzbereich eines Hypothesentests

“Das Konfidenzintervall enthält alle Werte, die zum zugrunde liegenden Testszenario passen. Wenn ein beobachteter Wert außerhalb dieses Konfidenzintervalls liegt, kann man dieses Testszenario mit einer Irrtumswahrscheinlichkeit (p-Wert) von höchstens 5% verwerfen.”

Konfidenzintervall und Prädiktionsintervall für den Mittelwert

Wenn Daten einer Normalverteilung entspringen, dann ist der arithmetische Mittelwert ein passender Schätzwert für den Erwartungswert der Daten. Daher ist die Annahme von annähernd normalverteilten Daten eine Grundvoraussetzung für das Schätzen von Mittelwerten. Hierbei unterscheiden wir allerdings unterschiedliche Szenarien:

1. Prädiktionsintervall = Vorhersagebereich = Zufallsstrebereich für die Werte zukünftiger Messungen

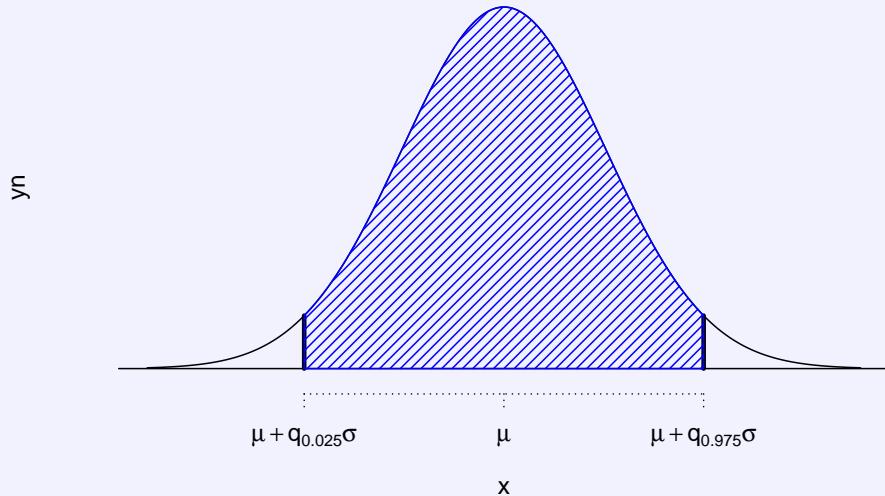
Prädiktionsintervall = Vorhersagebereich = Zufallsstrebereich für die Werte zukünftiger Messungen

Wenn der Mittelwert μ und die Standardabweichung σ bereits aus vorhergehenden Messung, aus den Grundeinstellung der Messgerätschaften etc. bekannt sind, dann gilt, lässt sich für eine zukünftige Messreihe der Mittelwert der Messungen mit einer gewissen Genauigkeit (**Konfidenz**) vorhersagen.

Unter Kenntnis von Mittelwert μ und die Standardabweichung σ gilt als Schätzung für die erwarteten Werte der zukünftigen Messungen, dass diese mit Wahrscheinlichkeit 95% vom Intervall

$$(\mu + q_{0.025} \cdot \sigma; \mu + q_{0.975} \cdot \sigma)$$

liegen.



Allgemein gilt für das Prädiktionsintervall mit **Konfidenz** $(1-\alpha)$, was einer **Signifikanz** (=Irrtumswahrscheinlichkeit bei Entscheidungen) von α entspricht:

$$(\mu + q_{\frac{\alpha}{2}} \cdot \sigma; \mu + q_{1-\frac{\alpha}{2}} \cdot \sigma)$$

Die Quantile $q_{\frac{\alpha}{2}}$ und $q_{1-\frac{\alpha}{2}}$ liegen also symmetrisch um den arithmetischen Mittelwert μ und jeweils die Hälfte des Signifikanzniveaus α wird anteilmäßig am oberen und am unteren Ende weggelassen.

2. Prädiktionsintervall = Vorhersagebereich = Zufallsstrebereich für den Mittelwert zukünftiger Messungen

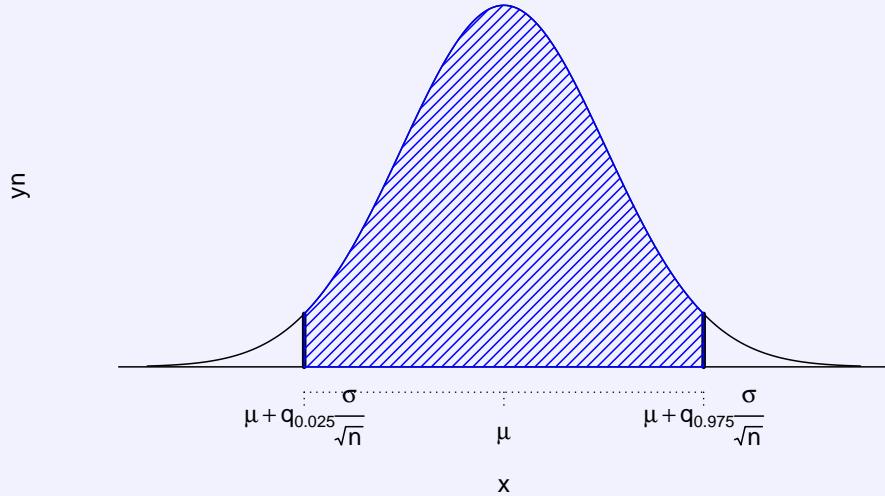
Prädiktionsintervall = Vorhersagebereich = Zufallsstrebereich für den Mittelwert zukünftiger Messungen

Wenn der Mittelwert μ und die Standardabweichung σ bereits aus vorhergehenden Messung, aus den Grundeinstellung der Messgerätschaften etc. bekannt sind, dann gilt, lässt sich für eine zukünftige Messreihe der Mittelwert der Messungen mit einer gewissen Genauigkeit (**Konfidenz**) vorhersagen.

Unter Kenntnis von Mittelwert μ und die Standardabweichung σ gilt als Schätzung für den erwarteten Wert des arithmetischen Mittelwerts \bar{x} , dass dieser mit Wahrscheinlichkeit 95% vom Intervall

$$\left(\mu + q_{0.025} \cdot \frac{\sigma}{\sqrt{n}}; \mu + q_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

liegt.



Allgemein gilt für das Prädiktionsintervall mit **Konfidenz** $(1-\alpha)$, was einer **Signifikanz** (=Irrtumswahrscheinlichkeit bei Entscheidungen) von α entspricht:

$$\left(\mu + q_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}; \mu + q_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Die Quantile $q_{\frac{\alpha}{2}}$ und $q_{1-\frac{\alpha}{2}}$ liegen also symmetrisch um den arithmetischen Mittelwert μ und jeweils die Hälfte des Signifikanzniveaus α wird anteilmäßig am oberen und am unteren Ende weggelassen.

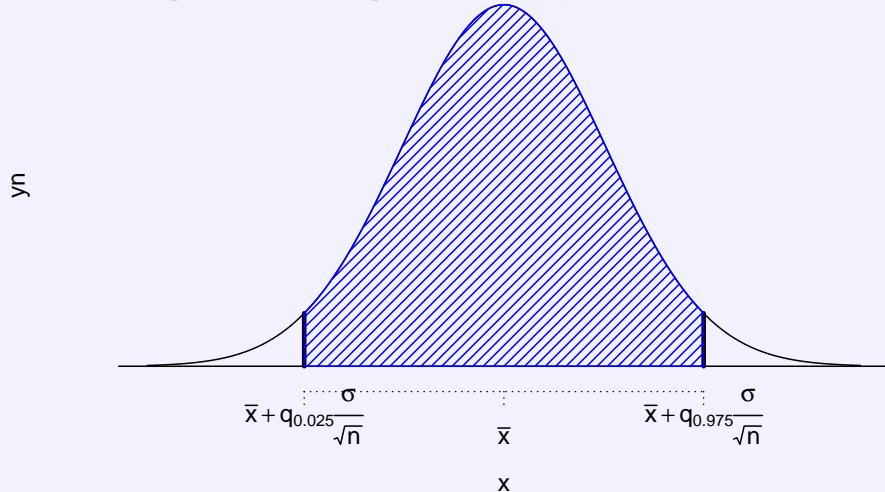
3. Konfidenzintervall = Vertrauensbereich mit bekanntem σ

Konfidenzintervall = Vertrauensbereich mit bekanntem σ

Für den arithmetischen Mittelwert \bar{x} gilt als Schätzung für den erwarteten Wert der Lage μ , dass dieser mit Wahrscheinlichkeit 95% vom Intervall

$$\left(\bar{x} + q_{0.025} \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + q_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

überdeckt wird, wobei $q_{0.025} = -1,96$ und $q_{0.975} = 1,96$ die Quantile der Normalverteilung sind.



Allgemein gilt für das Konfidenzintervall mit **Konfidenz** ($1-\alpha$), was einer **Signifikanz** (=Irrtumswahrscheinlichkeit bei Entscheidungen) von α entspricht:

$$\left(\bar{x} + q_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + q_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Die Quantile $q_{\frac{\alpha}{2}}$ und $q_{1-\frac{\alpha}{2}}$ liegen also symmetrisch um den arithmetischen Mittelwert \bar{x} und jeweils die Hälfte des Signifikanzniveaus α wird anteilmäßig am oberen und am unteren Ende weggelassen.

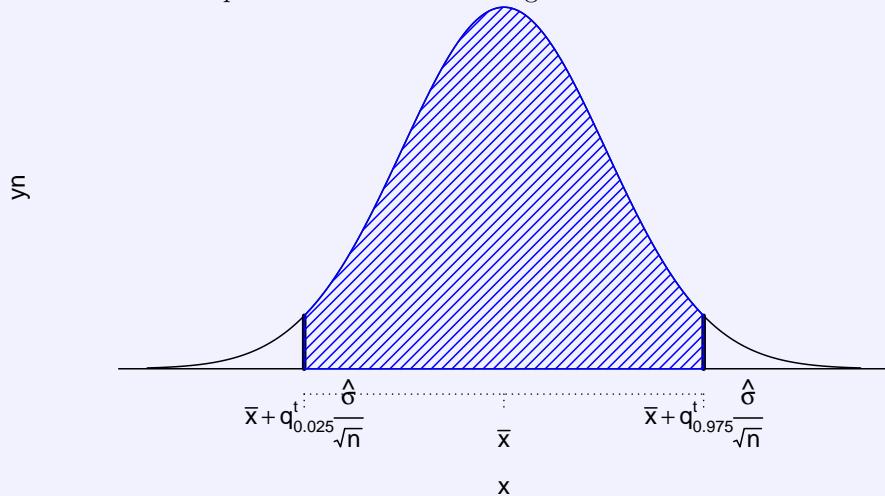
4. Konfidenzintervall = Vertrauensbereich mit unbekanntem σ

Konfidenzintervall = Vertrauensbereich mit unbekanntem σ

Für den arithmetischen Mittelwert \bar{x} gilt als Schätzung für den erwarteten Wert der Lage μ , dass dieser mit Wahrscheinlichkeit 95% vom Intervall

$$\left(\bar{x} + q_{0.025}^t \cdot \frac{\hat{\sigma}}{\sqrt{n}}; \bar{x} + q_{0.975}^t \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

überdeckt wird, wobei hier aber $q_{0.025}^t$ und $q_{0.975}^t$ die Quantile der t-Verteilung mit n Freiheitsgraden sind und $\hat{\sigma}$ die Stichprobenstandardabweichung ist.



Allgemein gilt für das Konfidenzintervall mit **Konfidenz** $(1-\alpha)$, was einer **Signifikanz** (=Irrtumswahrscheinlichkeit bei Entscheidungen) von α entspricht:

$$\left(\bar{x} + q_{\frac{\alpha}{2}}^t \cdot \frac{\hat{\sigma}}{\sqrt{n}}; \bar{x} + q_{1-\frac{\alpha}{2}}^t \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

Die Quantile $q_{\frac{\alpha}{2}}^t$ und $q_{1-\frac{\alpha}{2}}^t$ liegen also symmetrisch um den arithmetischen Mittelwert \bar{x} und jeweils die Hälfte des Signifikanzniveaus α wird anteilmäßig am oberen und am unteren Ende weggelassen.

Beispiel zu Prädiktionsintervall = Vorhersagebereich = Zufallsstrebereich für die Werte zukünftiger Messungen und deren Mittelwerte

Die Verteilung von Flüssigkeitsmengen in abgepackten Eprouvetten wird als annähernd normalverteilt angenommen. Sie liegt im Mittel bei einem Wert von 2.5 ml, wobei die Abfüllmenge mit einer Standardabweichung von 0.1 ml um diesen Wert schwanken kann. Es soll dann eine Stichprobe von 35 Proben entnommen werden und für diese vorhergesagt werden, in welchem Bereich voraussichtlich zu 95% die Werte der Stichprobe liegen werden und in welchem Bereich der Mittelwert liegen wird.

```
mittelwert=2.5  
standardabweichung=0.1  
stichprobengroesse=35  
quantilunten=qnorm(0.025,mean = 0,sd=1)  
quantiloben=qnorm(0.975,mean = 0,sd=1)
```

Dann erfolgt die Berechnung der Grenzen mithilfe der folgenden Befehle für das Prädiktionsintervall = Zufallsstrebereich der Datenwerte:

untere Intervallgrenze des Prädiktionsintervalls der Werte:
 $mittelwert+quantilunten*standardabweichung = 2.3040036 \text{ ml}$

obere Intervallgrenze des Prädiktionsintervalls der Werte:
 $mittelwert+quantiloben*standardabweichung = 2.6959964 \text{ ml}$

Dann erfolgt die Berechnung der Grenzen mithilfe der folgenden Befehle für das Prädiktionsintervall = Zufallsstrebereich des Mittelwerts:

untere Intervallgrenze des Prädiktionsintervalls der Mittelwerte:
 $mittelwert+quantilunten*standardabweichung/sqrt(stichprobengroesse) = 2.4668706 \text{ ml}$

obere Intervallgrenze des Prädiktionsintervalls der Mittelwerte:
 $mittelwert+quantiloben*standardabweichung/sqrt(stichprobengroesse) = 2.5331294 \text{ ml}$

Beispiel zu Konfidenzintervall = Vertrauensbereich für die tatsächlichen Mittelwerte\

Die Verteilung von Flüssigkeitsmengen in abgepackten Eprouvetten wird als annähernd normalverteilt angenommen. Doch im Unterschied zu vorher weiß man den Mittelwert nicht.

Im Szenario 1 wissen wir vom Hersteller, dass die Abfüllmenge mit einer Standardabweichung von 0.1 ml um den Abfüllwert schwanken kann.

Im Szenario 2 kennen wir auch die Standardabweichung nicht, sondern müssen beides erst ermitteln. Es wurde eine Stichprobe von 35 Proben entnommen und für diese bereits der Stichprobenmittelwert $\bar{x} = 2.55$ ml und die Stichprobenstandardabweichung $\hat{\sigma} = 0.11$ ml berechnet. Nun soll ermittelt werden, in welchem Bereich zu 95%-iger Wahrscheinlichkeit der wahre Mittelwert μ in beiden Szenarien liegen wird.

```
mittelwert=2.55
theoretischedstandardabweichung=0.1
stichprobenstandardabweichung=0.11
stichprobengroesse=35
normalquantilunten=qnorm(0.025,mean = 0,sd=1)
normalquantiloben=qnorm(0.975,mean = 0,sd=1)
tquantilunten=qt(0.025,df=(stichprobengroesse-1))
tquantiloben=qt(0.975,df=(stichprobengroesse-1))
```

Dann erfolgt die Berechnung der Grenzen mithilfe der folgenden Befehle für das Konfidenzintervall = Vertrauensbereich des Mittelwerts:

untere Intervallgrenze des Konfidenzintervalls der Mittelwerte bei bekannter Standardabweichung:
mittelwert+normalquantilunten*theoretischedstandardabweichung/sqrt(stichprobengroesse)
= 2.5168706 ml

obere Intervallgrenze des Konfidenzintervalls der Mittelwerte bei bekannter Standardabweichung:
mittelwert+normalquantiloben*theoretischedstandardabweichung/sqrt(stichprobengroesse)
= 2.5831294 ml

Dann erfolgt die Berechnung der Grenzen mithilfe der folgenden Befehle für das Konfidenzintervall = Vertrauensbereich des Mittelwerts:

untere Intervallgrenze des Konfidenzintervalls der Mittelwerte bei unbekannter Standardabweichung:
mittelwert+tquantilunten*stichprobenstandardabweichung/sqrt(stichprobengroesse) =
2.5122137 ml

obere Intervallgrenze des Konfidenzintervalls der Mittelwerte bei unbekannter Standardabweichung:
mittelwert+tquantiloben*stichprobenstandardabweichung/sqrt(stichprobengroesse) =
2.5877863 ml

Prädiktionsintervall und Konfidenzintervall für Proportionen (relative Häufigkeiten) = Vertrauensbereiche für Kategorienhäufigkeiten

Die Binomialverteilung zählt die Anzahl der “interessanten Ereignisse” bei n Experimenten, wenn

- nur zwei mögliche Ereignisse - “Erfolg” und “Misserfolg” - existieren
- die Erfolgswahrscheinlichkeit immer gleichbleibend p ist
- n Experimente, die nur in Erfolg oder Misserfolg ende können, aneinandergereiht werden

Die Binomialverteilung hat eine Wahrscheinlichkeitsfunktion für die gezählten Erfolge k von

$$\mathbb{P}[X = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

mit den Parameter, die der Stichprobengröße n und Wahrscheinlichkeit für “interessanten Ereignisse” p der Daten entsprechen, gilt:

$$\begin{aligned}\mathbb{E}[X] &= n \cdot p \\ \mathbb{V}[X] &= n \cdot p \cdot (1-p)\end{aligned}$$

Wir brauchen diese Werte jetzt, um damit das Konfidenzintervall und das Prädiktionsintervall zu konstruieren.

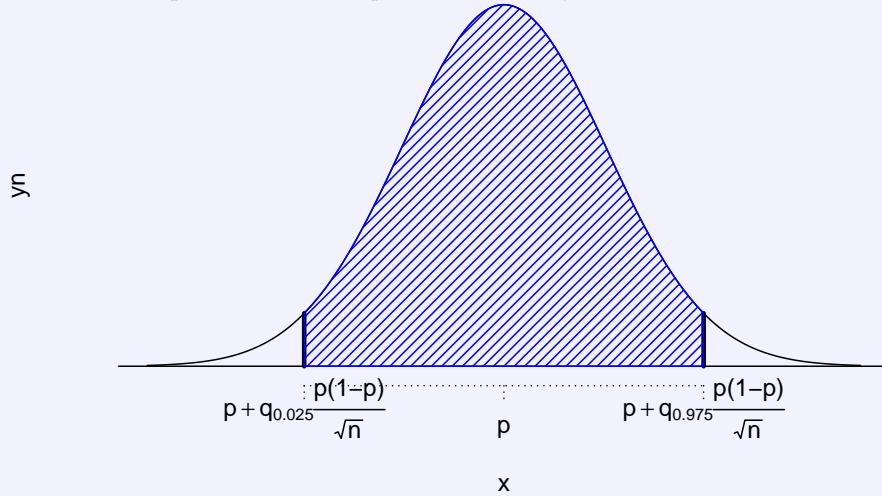
1. **Prädiktionsintervall für Proportionen (relative Häufigkeiten) = Vorhersagebereich/Zufallsstrebereich für Kategorienhäufigkeiten**

Prädiktionsintervall für Proportionen (relative Häufigkeiten) = Vorhersagebereich/Zufallsstrebereich für Kategorienhäufigkeiten

Unter Kenntnis der Proportion p und der zukünftigen Stichprobengröße n gilt als Schätzung für die Wahrscheinlichkeit einer Kategorie p , dass die zukünftig beobachtete Proportion \hat{p} mit Wahrscheinlichkeit 95% vom Intervall

$$\left(p + q_{0.025} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}; p + q_{0.975} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \right)$$

überdeckt wird, wobei $q_{0.025} = -1.96$ und $q_{0.975} = 1.96$ die Quantile der Normalverteilung sind.



Allgemein gilt für das Prädiktionsintervall mit **Konfidenz** ($1-\alpha$), was einer **Signifikanz** (=Irrtumswahrscheinlichkeit bei Entscheidungen) von α entspricht:

$$\left(p + q_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}; p + q_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \right)$$

Die Quantile $q_{\frac{\alpha}{2}}$ und $q_{1-\frac{\alpha}{2}}$ liegen also symmetrisch um den arithmetischen Mittelwert p und jeweils die Hälfte des Signifikanzniveaus α wird anteilmäßig am oberen und am unteren Ende weggelassen.

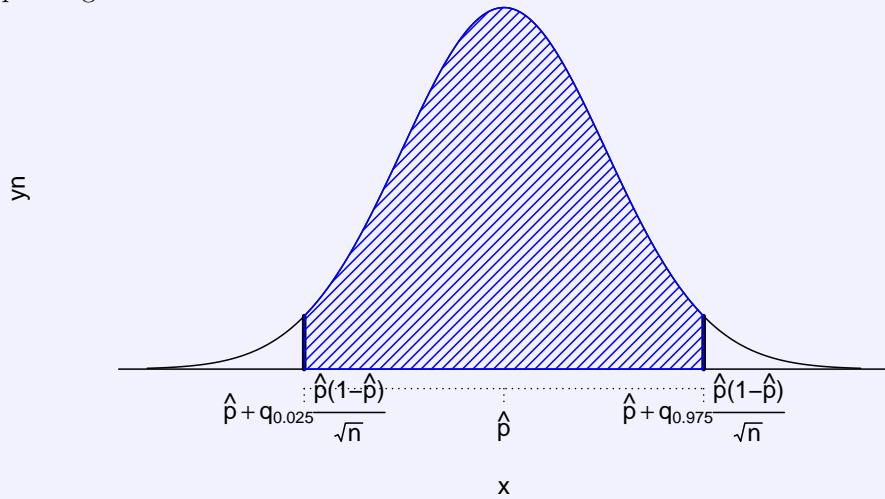
2. Konfidenzintervall für Proportionen (relative Häufigkeiten) = Vertrauensbereiche für Kategorienhäufigkeiten

Konfidenzintervall für Proportionen (relative Häufigkeiten) = Vertrauensbereiche für Kategorienhäufigkeiten

Für die Proportion $\hat{p} = \frac{h_1}{n}$ gilt als Schätzung für die Wahrscheinlichkeit einer Kategorie p , dass die wahre Proportion p mit Wahrscheinlichkeit 95% vom Intervall

$$\left(\hat{p} + q_{0.025} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}; \hat{p} + q_{0.975} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right)$$

überdeckt wird, wobei $q_{0.025} = -1,96$ und $q_{0.975} = 1,96$ die Quantile der Normalverteilung sind und die Stichprobengröße n ist.



Allgemein gilt für das Konfidenzintervall mit **Konfidenz** $(1-\alpha)$, was einer **Signifikanz** (=Irrtumswahrscheinlichkeit bei Entscheidungen) von α entspricht:

$$\left(\hat{p} + q_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}; \hat{p} + q_{1 - \frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right)$$

Die Quantile $q_{\frac{\alpha}{2}}$ und $q_{1 - \frac{\alpha}{2}}$ liegen also symmetrisch um den arithmetischen Mittelwert \bar{p} und jeweils die Hälfte des Signifikanzniveaus α wird anteilmäßig am oberen und am unteren Ende weggelassen.

Beispiel zu Prädiktionsintervall = Vorhersagebereich = Zufallsstrebereich für die Proportionen

Aus langjähriger Praxis weiß ein Hausarzt, dass ein bestimmtes Medikament in 85% der Fälle bei seinen Patienten eine positive Wirkung bei der Behandlung von grippalen Infekt-Symptomen zeigt. Er möchte abschätzen, bei wie vielen seiner 135 Patienten, denen er das Medikament verschreibt voraussichtlich eine Wirkung eintreten wird.

```
proportion=0.85  
stichprobengroesse=135  
quantilunten=qnorm(0.025,mean = 0,sd=1)  
quantiloben=qnorm(0.975,mean = 0,sd=1)
```

Dann erfolgt die Berechnung der Grenzen mithilfe der folgenden Befehle für das Prädiktionsintervall = Zufallsstrebereich der Proportionen:

untere Intervallgrenze des Prädiktionsintervalls der Proportionen:

```
proportion+quantilunten*sqrt((proportion*(1-proportion))/stichprobengroesse) =  
0.7897667 ml
```

obere Intervallgrenze des Prädiktionsintervalls der Proportionen:

```
proportion+quantiloben*sqrt((proportion*(1-proportion))/stichprobengroesse) =  
0.9102333 ml
```

Beispiel zu Konfidenzintervall = Vertrauensbereich für die Proportionen

Ein neuartiges Medikament wird an einer Gruppe von 135 Patienten getestet und zeigt bei 105 von ihnen eine positive Wirkung. Anhand dieser Zahlen soll ein Konfidenzintervall für die wahre Wirkungswahrscheinlichkeit des Medikaments ermittelt werden.

```
proportion=105/135  
stichprobengroesse=135  
quantilunten=qnorm(0.025,mean = 0,sd=1)  
quantiloben=qnorm(0.975,mean = 0,sd=1)
```

Dann erfolgt die Berechnung der Grenzen mithilfe der folgenden Befehle für das Prädiktionsintervall = Zufallsstrebereich der Proportionen:

untere Intervallgrenze des Prädiktionsintervalls der Proportionen:

```
proportion+quantilunten*sqrt((proportion*(1-proportion))/stichprobengroesse) =  
0.707648 ml
```

obere Intervallgrenze des Prädiktionsintervalls der Proportionen:

```
proportion+quantiloben*sqrt((proportion*(1-proportion))/stichprobengroesse) =  
0.8479076 ml
```

Additions to Confidence versus prediction regions

Prediction region

Under the assumption of a specific distribution of the data and the “known” value of the parameter to be measured, the prediction interval is determined. This interval has a $(1-\alpha) \cdot 100\%$ probability to contain the estimate of the parameter in a future sample of this distribution.

Confidence region

Based on a sample of observation an estimate of the “unknown” value of the parameter of interest and the confidence interval is calculated. This confidence interval contains the true underlying value of the parameter with probability $(1-\alpha) \cdot 100\%$.

Mathematical Definition of Confidence Intervals

Confidence Interval A set of values which contains the estimate of an unknown population parameter with a certain percentage (confidence level $1-\alpha$), if many samples were drawn repeatedly. Formally, the confidence interval at confidence level $1 - \alpha$ for a random sample X with probability distribution depending on parameter(s) θ is an interval with random endpoints $(\ell(X), u(X))$ which fulfills

$$\mathbb{P} [\ell(X) \leq \theta \leq u(X)] = 1 - \alpha$$

additional assumptions for $\ell(X)$ and $u(X)$ can be:

- symmetric around the “center” of the interval (mean or median)
- the same amount of the distribution $\frac{\alpha}{2}$ lies below $\ell(X)$ and above $u(X)$

How to calculate confidence regions?

Confidence region A confidence region is a set of points which cover the parameter to estimate with a certain probability

Confidence regions require quantiles and thus distributions

these distributions can come from

- the sample itself (Bootstrap, sampling distribution),
- asymptotic assumptions (CLT) or
- distribution assumptions.

Konzept der Hypothesentests

Hypothesentesten ist eine Methode zur Entscheidungsfindung basierend auf Daten und quantitativen Beobachtungen. Neben Punktschätzungen und Konfidenzintervallen dient in der Inferenzstatistik das **Testen von statistischen Hypothesen** der Entscheidungsfindung und dem Erlangen von Schlussfolgerungen. Dabei unterscheidet man grundsätzlich zwischen unterschiedlichen Typen:

- **Parameterhypothesen** testen eine Behauptung über den (oder die) Parameter von einer zugrunde liegenden Verteilung. Hier wird angenommen, dass die Verteilung der Daten wenigstens bezüglich ihres Typs bekannt ist (beispielsweise, dass es sich um normal verteilte Daten handelt).
- **Verteilungshypothesen** testen die Art der Verteilung selbst. Hier ist der Verteilungstyp also nicht bekannt und möchte man testen, ob eine bestimmte Verteilung oder eine bestimmte Verteilungsfamilie (beispielsweise, die Familie der Normalverteilungen) zu den Daten als zugrunde liegendes Modell passt.

Des Weiteren unterscheiden wir nach der Art des Tests, je nachdem, wie die Eigenschaften vorliegen:

- **Parametrische Hypothesentests** testen eben jene Parameterhypothesen einer zugrunde liegenden Verteilung. Die Voraussetzung, dass die Verteilung der Daten gilt, ist hier also essenziell.

Beispielhaft dafür ist den Mittelwert zu testen unter der Annahme, dass die Daten normal verteilt sind. Dieser Test heißt t-Test und wir werden uns noch sehr genau damit auseinander setzen.

- **Nichtparametrische Hypothesentests** testen ebenfalls Eigenschaften einer zugrunde liegenden Verteilung oder auch die Verteilung selbst. Sie kommen zur Anwendung, wenn die Voraussetzung, dass die Verteilung der Daten gilt, nicht erfüllt werden kann oder wie bei Verteilungshypothesen noch gar nicht bekannt oder nachgewiesen ist.

Beispielhaft dafür ist anstatt der Mittelwerts die Lage einer Stichprobe zu testen, indem man als robustes Maß Ränge anstatt der Originaldaten heranzieht. Die Fragestellung ist gleich wie beim t-Test, aber die Verteilungsannahme muss nicht erfüllt werden. Ein Test ist etwa der Wilcoxon Rangsummen Test und wir werden uns diesen als Alternative für nichtnormalverteilte Daten ansehen.

- **Simulationsbasierte Hypothesentests** testen ebenfalls Eigenschaften einer zugrunde liegenden Verteilung oder auch die Verteilung selbst, kommen aber im Unterschied zu nichtparametrischen Tests zum Einsatz, wenn zu wenige Daten verfügbar sind oder aber das Vergleichen von Daten notwendig wird, wenn selbst die Annahmen an nicht-parametrische Test nicht erfüllt sind. Sie kommen auch immer dann zum Einsatz, wenn keine parametrischen oder nicht-parametrischen Testszenarien existieren und haben ihren Aufschwung mit der Verfügbarkeit stark Rechenleistung bei Computern genommen.

{Begriffe zu Hypothesentests}

Hypothese (engl. *hypothesis*): Eine Annahme über die Daten oder ihre zugrunde liegende Verteilung und Eigenschaften, welche für den Test relevant ist. Diese muss auf eine spezifische Eigenschaft fokussiert sein, also etwa die Lage der Daten, ihre Streuung oder ihr Typus von Verteilung, aber nicht alles davon zugleich.

(assumption about the underlying structure of the population and distribution from which the sample is drawn)

- **Nullhypothese** (engl. *null hypothesis*) (H_0): Die Referenzannahme des Modells als Standardszenario, unter dem die Daten kreiert worden sind.

(basic structure and distribution assumed for the data, if nothing interesting occurs ('standard'))

- **Alternativhypothese** (engl. *alternative hypothesis*) (H_A): Die eigentliche Annahme von Interesse, welche eine Veränderung des 'Status Quo' der Nullhypothese bedeutet. Die Alternativhypothese ist häufig einfach die logische Verneinung der Nullhypothese, oder diejenige Behauptung, deren Zutreffen ein bestimmtes Handelnerfordert oder die gravierenderen Konsequenzen (positive oder negative) nach sich zieht. Tatsächliche erfolgt die Konstruktion aber eher von der Fragestellung der Alternativhypothese ausgehend, dass die Nullhypothese ihre logische Verneinung ist, als umgekehrt.

(interesting case being reasonably different from the standard case ('something happens'))

Teststatistik (engl. *test statistics*): Ein Wert oder Schätzer, welcher unter der Annahme des geltenden Nullhypotesenszenarios aus den beobachteten Daten berechnet wird. Dieser fasst zusammen, wie 'gut' die Daten zur Nullhypothese passen. Diese Teststatistik hat selbst wiederum eine Verteilung, die sich aus der Art des Tests und den damit verbundenen Annahmen ergibt, und anhand derer die Bewertung der Alternativhypothese geschieht.

(value/estimate calculated from the sample under the distribution assumption of H_0)

Was bei Hypothesentest zu beachten ist!

- **Beweisen kann nur die Mathematik und Logik als Formalwissenschaft, alle anderen Wissenschaften können mithilfe der Statistischen Hypothesentestmethodik nur nachweisen! You cannot prove anything, except for maths!**
- **ACHTUNG: Nullhypothese und Alternativhypothese sind NICHT vertauschbar! Die Nullhypothese kann niemals nachgewiesen, sondern nur widerlegt werden. Die Nullhypothese nicht widerlegen zu können, ist KEIN Nachweis. Daher muss die Fragestellung in der Alternativhypothese abgebildet werden.\ Caveat: Null hypothesis and alternative hypothesis are not balanced!\ You cannot accept the null hypothesis, only reject it!** Failure to reject the null hypothesis does not necessarily mean that it is true!

Zur besseren Vorstellung bemühen wir ein Beispiel aus der Philosophie und Wissenschaft. Lange Zeit nahm die Menschheit an, dass die Erde ein flache Scheibe und das Zentrum des Universums sei. Alle Beobachtungen von Sternen und Planeten wurden im Lichte dieser Annahme evaluiert, also quasi unter dem **Nullhypotesenszenario**. Erst nachdem zunehmend Evidenz für die **Alternativhypothese**, dass die Erde nicht flach oder Zentrum des Universums ist, gesammelt wurde, konnte die Nullhypothese verworfen werden und das geozentrische Weltbild wurde durch das heliozentrische ersetzt. Nicht genügend Beobachtungen zu haben, die die Erde als Zentrum des Universums widerlegen, weist das geozentrische Weltbild allerdings nicht nach und schon gar nicht kann es durch Mangel an Widersprüchen bewiesen werden!

Possible wordings in the final conclusion of research papers or a thesis:

- *There is sufficient evidence to warrant rejection of the claim that ...*
- *There is not sufficient evidence to warrant rejection of the claim that ...*

Das Konzept der Signifikanz und p-Werte

Signifikanz

Statistische Signifikanz: Etwas wird als ‘signifikant’ eingestuft, wenn es sehr unwahrscheinlich ist, dass dieses Ergebnis/Messung/Schätzwert durch reinen Zufall im Nullhypothesenzenario zustande gekommen ist. Hier wird als Gegenteil von ‘reinem Zufall’ ein systematisch anderer Prozess also die Alternativhypothese angenommen. Als Schwellwert dafür, ab wann es ‘zu unwahrscheinlich’ ist, dass Daten unter der Nullhypothese generiert wurden, dient das **Signifikanzniveau α** .

Der **p-Wert (engl. p-value)** ist die Wahrscheinlichkeit im Szenario der Nullhypothese die beobachtete Teststatistik ermittelt aus dem Daten oder eine noch extremer abweichende zu beobachten. Das ist die Wahrscheinlichkeit, einen Fehler zu begehen, wenn die Nullhypothese für bestimmte Daten verworfen wird. Verglichen wird der p-Wert mit dem Signifikanzniveau α , das somit den größtmöglichen p-Wert angibt, bei dem die Nullhypothese noch verwerfen wird.

Beispiel: Ein p-Wert von 0.3 wird mit dem Signifikanzniveau verglichen, dabei sind Wert wie $\alpha = 0.05$ oder 0.01 oder 0.001 gängige Praxis. 0.3 ist aber jedenfalls größer als jeder dieser Wert und daher liegt keine Signifikanz vor. Da die Irrtumswahrscheinlichkeit 30% betragen würde, wird die Nullhypothese nicht verworfen.

Ein p-Wert von 0.03 wird mit dem Signifikanzniveau $\alpha = 0.05$ verglichen. Hier zeigt sich, dass die Irrtumswahrscheinlichkeit bei Verwerfen der Nullhypothese nun unter der obersten Toleranzschwelle 0.05 liegt und daher H_0 verworfen wird. Wäre jedoch $\alpha = 0.01$ würden wir beim selben p-Wert H_0 nicht verworfen.

English version:

A result is significant if it is unlikely that the observed outcome occurred by chance under a null hypothesis given a threshold of significance (**significance level**).

significance level α : When assuming the standard structure under H_0 , we obtain probabilities for every observed sample estimate (independent of the alternative hypothesis). If the probability of this estimate falls beneath a certain predefined (!) level (=significance level), the null hypothesis is rejected.

p-value: the probability of observing the sample estimate (=test statistics) or a more extreme value, if drawing randomly from the distribution defined by the null hypothesis

This is the actual probability of making an error when rejecting the null hypothesis in a specific scenario for the given data is compared against the **significance level α** which represents the highest p-value allow to still reject the null hypothesis

A p-value of 0.03 is compared against the significance level. For $\alpha = 0.05$ the null hypothesis is rejected as the probability of making a mistake is 3% and thus lower than the highest tolerated error probability 5%.

For $\alpha = 0.01$ the null hypothesis is **not** rejected as the probability of making a mistake 3% is higher than the tolerated highest bound of 1%.

Gängige Bezeichnungen für p-Werte und Signifikanzniveaus

Signifikanz

p-value	Signifikanzniveau α	Notation	Bezeichnung
	1		
e.g. 0.24	0.05		nicht signifikant (not significant)
e.g. 0.024	0.01	*	schwach signifikant (weakly significant)
e.g. 0.0024	0.001	**	signifikant (significant)
e.g. $2.4 \cdot 10^{-6}$	0.00	***	hoch signifikant (highly significant)

Fehler beim Hypothesentesten

Fehlertabelle

	H_0 WAHR	H_0 FALSCH
H_0 beibehalten	✓	Type II Fehler
H_0 verwerfen	Type I Fehler	✓

- **Falsch positiv Rate** = Signifikanzniveau (α): Wahrscheinlichkeit einen Typ I Fehler begehen, fälschlicherweise die Nullhypothese zu verwerfen (positives Resultat")
- **Falsch negativ Rate** (β): Wahrscheinlichkeit einen Typ II Fehler zu begehen, fälschlicherweise die Nullhypothese beizubehalten (negative outcome)
- **Schärfe = Power** ($1 - \beta$): Wahrscheinlichkeit, die Nullhypothese korrekterweise zu verwerfen

Das **Signifikanzniveau** (α) wird durch Datenanalytiker *vor* Durchführung der Analyse und des Tests festgelegt. Typisch sind die Wert, die wir als Schwellen der Signifikanz bereits kennen, etwa $\alpha = 0.05$, $\alpha = 0.01$ oder $\alpha = 0.001$.

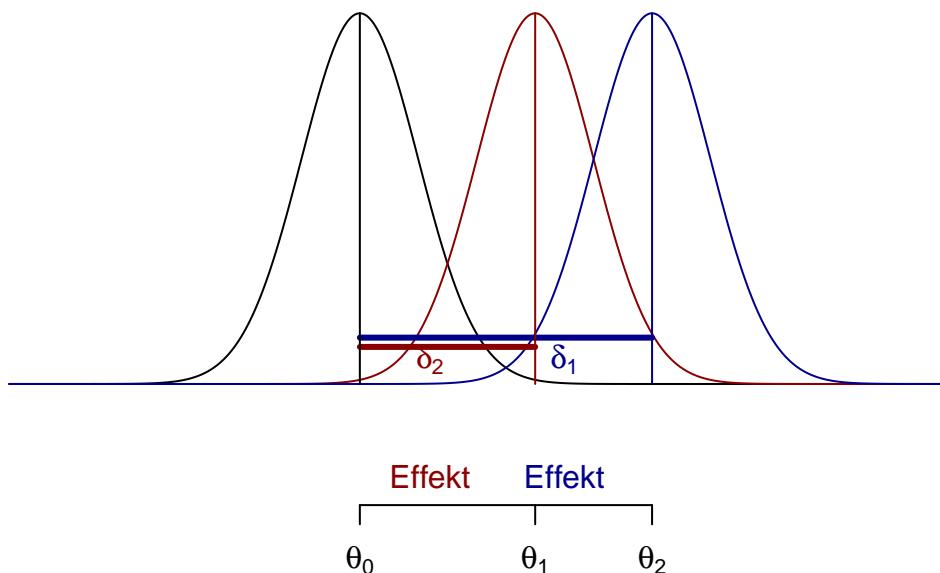
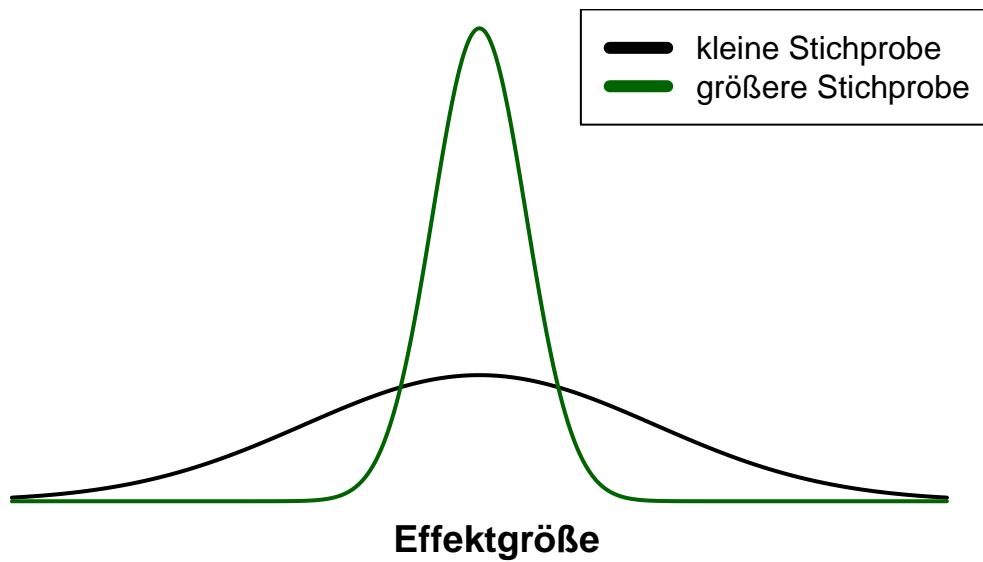
Trennschärfe=Power

Die **Schärfe = Power** $1 - \beta$ ist unbekannt, aber wir wissen, was darauf Einfluss hat:

- **Stichprobengröße**: je mehr Beobachtungen, desto größer die Trennschärfe des Tests, das die Streuung der Teststatistik kleiner wird.
- **Effektgröße**: je größer der zu detektierende Effekt (= Unterschied zwischen dem gemessenen und getesteten Wert oder mehreren gemessenen Werten) the larger the effect (difference of values) desto größer die trennschärfe
- **Signifikanzniveau**: je höher das Signifikanzniveau, desto größer die Trennschärfe = je größer der TypI Fehler, desto kleiner der TypII Fehler
- **Parametrisierung**: parametrische Tests haben höhere Trennschärfe als nichtparametrische Tests

Je kleiner der α Fehler gewählt wird, desto größer wird der β Fehler!

Stichprobengröße



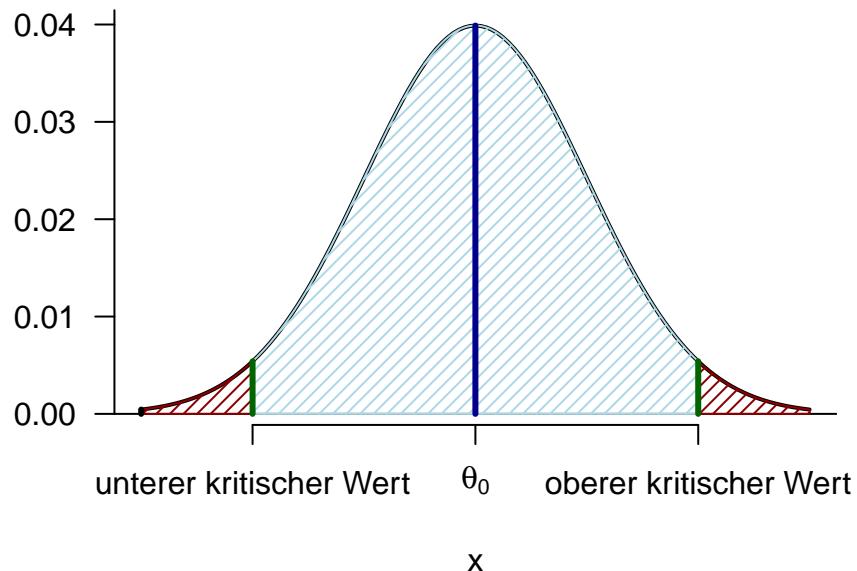
Zusammenhang zwischen Tests und Konfidenzintervallen Testing regions

Bei einem gegebenen Signifikanzniveau $\alpha=0.05$ ergeben sich die **kritischen Werte** als diejenigen Werte, ab denen die Teststatistik so weit von der Annahme der Nullhypothese entfernt ist, dass diese ab diesen Schwellwerten verworfen werden kann. Diese hängen mit der zugrunde liegenden **Verteilung der Teststatistik** zusammen, welche NICHT die Verteilung der Daten ist, aber von dieser etwa im parametrischen Fall abhängen kann.

Testbereiche

- Der **Akzeptanzbereich** umfasst die Menge aller Werte, für die die Nullhypothese valide ist und nicht mit zu großem Typ I Fehler der Basisannahme widerprochen wird. (Bereich, in dem H_0 nicht verworfen werden kann - hier hellblau)
- Der **Verwerfungsbereich** umfasst die Menge aller Werte, für die die Nullhypothese nicht valide ist ohne einen zu großen Typ I Fehler zu begehen. (Bereich, in dem H_0 verworfen werden kann - hier dunkelrot)

Testbereiche



Äquivalenz mit Prädiktionsbereich Unter den entsprechenden Verteilungsannahmen für die Daten und die Eigenschaft des Parameters der Nullhypothese ist im parametrischen Fall das Prädiktionsintervall mit dem zugehörigen Konfidenzniveau $1 - \alpha$ identisch mit dem Akzeptanzbereich des Hypothesentests.

English version: Statistical Errors

Error Table

	H_0 TRUE	H_0 FALSE
keep H_0	✓	Type II error
reject H_0	Type I error	✓

- **False positive rate** = significance level (α): probability of committing Type I error, incorrectly rejecting H_0 (positive outcome)
- **False negative rate** (β): probability to commit Type II error, incorrectly accepting H_0 (negative outcome)
- **Power** ($1 - \beta$): probability of correctly rejecting H_0
- **Parametrisation**: parametric tests have larger power than non-parametric ones

The **significance level** (α) is chosen by the statistician *before* the analysis and typically values such as $\alpha = 0.05$, $\alpha = 0.01$ or $\alpha = 0.001$ are chosen

The **Power** $1 - \beta$ is unknown, but affected by:

- **sample size**: the more observations, the larger the power
- **effect size**: the larger the effect (difference of values) to be detected, the larger the power
- **significance level**: the smaller the significance level, the smaller the power

The smaller the α error is chosen, the larger the β error becomes!

Testing regions

- **Region of acceptance**: set of value where the null hypothesis remains valid (cannot be rejected)
- **Region of rejection**: set of values where the null hypothesis is rejected

Equivalence to Prediction Regions Under distribution assumptions identical to the null hypothesis of the corresponding test the prediction interval is identical with the region of acceptance of the test.

Richtung der Hypothesentests - Formulierung der Fragestellung

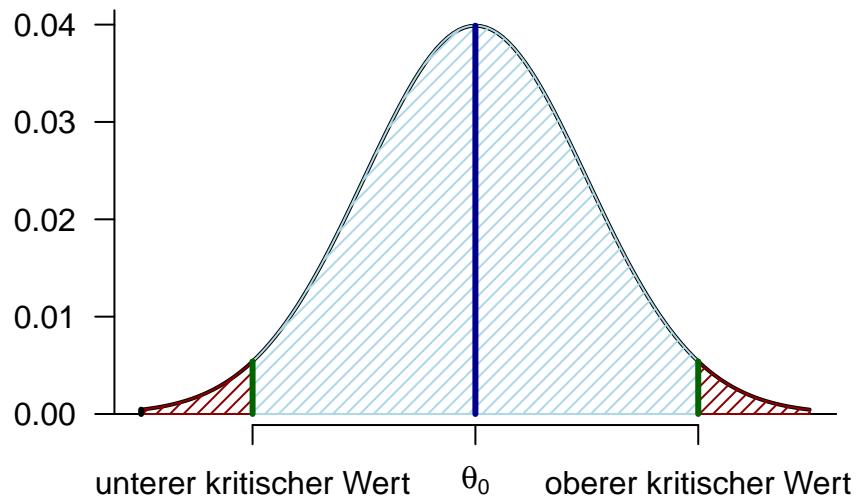
Zweiseitige Tests Bei einem gegebenen Signifikanzniveau $\alpha=0.05$ ergeben sich die **kritischen Werte** als diejenigen Werte, ab denen die Teststatistik so weit von der Annahme der Nullhypothese entfernt ist, dass diese ab diesen Schwellwerten verworfen werden kann. Beim zweiseitigen Test sind sie so platziert, dass jeweils ein Anteil $\frac{\alpha}{2}$ der Verteilung unterhalb der unteren Schranke und oberhalb der oberen Schranke liegt.

Die Formulierung von Nullhypothese und Alternativhypothese lautet stets:

$H_0: \theta = \theta_0$ "Der Parameter ist gleich einem bestimmten Referenzwert θ_0 "

$H_A: \theta \neq \theta_0$ "Der Parameter ist unterschiedlich von einem bestimmten Referenzwert θ_0 "

Zweiseitiger Test



Rechtsseitige Tests Bei einem gegebenen Signifikanzniveau $\alpha=0.05$ ergeben sich die **kritischen Werte** als diejenigen Werte, ab denen die Teststatistik so weit von der Annahme der Nullhypothese entfernt ist, dass diese ab diesen Schwellwerten verworfen werden kann. Beim rechtsseitigen Test sind sie so platziert, dass der gesamte Anteil α der Verteilung unterhalb der unteren Schranke liegt.

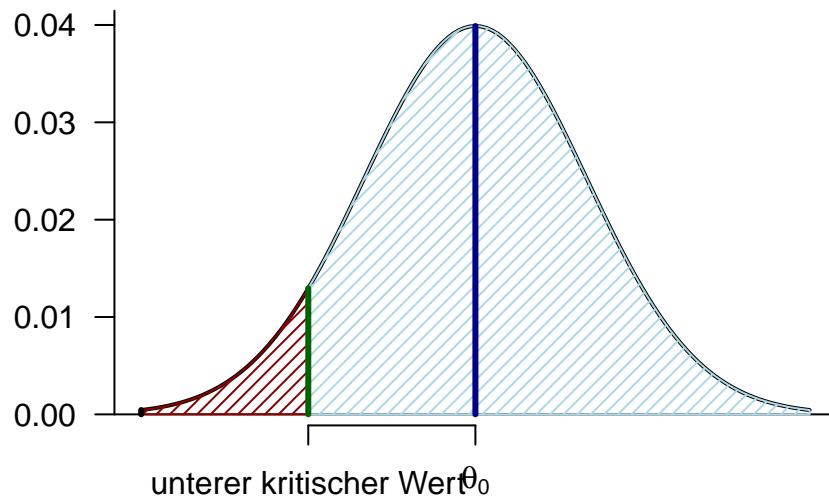
Die Formulierung von Nullhypothese und Alternativhypothese lautet stets:

$H_0: \theta \geq \theta_0$ "Der Parameter ist größer oder gleich einem bestimmten Referenzwert θ_0 "

$H_A: \theta < \theta_0$ "Der Parameter ist kleiner als ein bestimmter Referenzwert θ_0 = Eine Referenzschwelle θ_0 wird unterschritten"

Wir beachten, dass hier der untere Wert des Konfidenzintervalls immer $-\infty$ ist. Der obere Wert ist der kritische Wert, welcher hier näher beim Parameter θ_0 liegt als die beiden kritischen Werte beim zweiseitigen Akzeptanzbereich. Einseitiges Testen hat also größere Trennschärfe, wenn uns nur eine Richtung interessiert!

Rechtsseitiger Test



Linksseitige Tests Bei einem gegebenen Signifikanzniveau $\alpha=0.05$ ergeben sich die **kritischen Werte** als diejenigen Werte, ab denen die Teststatistik so weit von der Annahme der Nullhypothese entfernt ist, dass diese ab diesen Schwellwerten verworfen werden kann. Beim linksseitigen Test sind sie so platziert, dass der gesamte Anteil α der Verteilung oberhalb der oberen Schranke liegt.

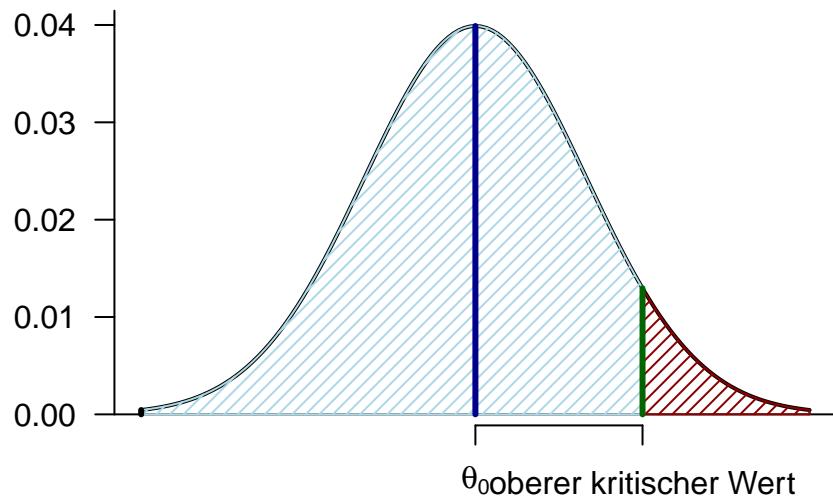
Die Formulierung von Nullhypothese und Alternativhypothese lautet stets:

$H_0: \theta \leq \theta_0$ "Der Parameter ist kleiner oder gleich einem bestimmten Referenzwert θ_0 "

$H_A: \theta > \theta_0$ "Der Parameter ist größer als ein bestimmter Referenzwert θ_0 = Eine Referenzschwelle θ_0 wird überschritten"

Wir beachten, dass hier der obere Wert des Konfidenzintervalls immer ∞ ist. Der untere Wert ist der kritische Wert, welcher hier näher beim Parameter θ_0 liegt als die beiden kritischen Werte beim zweiseitigen Akzeptanzbereich. Einseitiges Testen hat also größere Trennschärfe, wenn uns nur eine Richtung interessiert!

Linksseitiger Test



Erweiterte Inhalte zu Verteilungen von Teststatistiken

Zentraler Grenzverteilungssatz - Central Limit Theorem

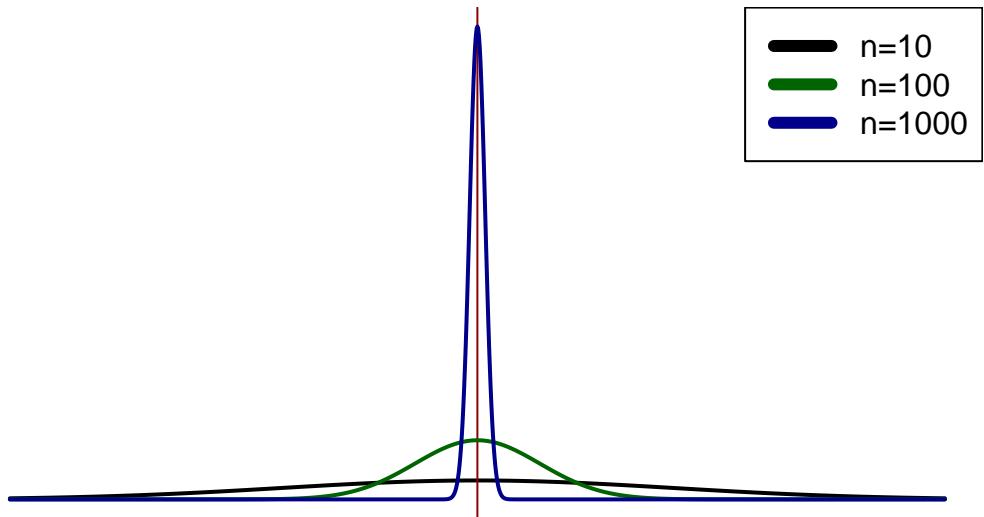
Central Limit Theorem Let X_1, X_2, \dots be a sequence of i.i.d. random variables with expectation $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then as n grows towards infinity, the random variables $\sqrt{n}(\bar{X} - \mu)$ converge in distribution to a normal distribution $N(0, \sigma^2)$.

$$\bar{X} \sim^P N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow$$

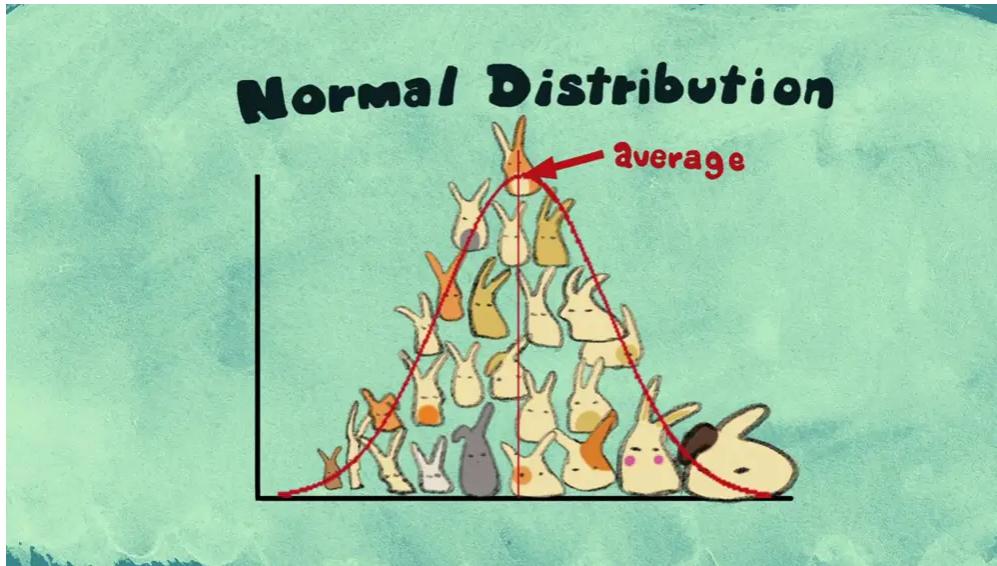
$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim^P N(0, 1)$$

Central Limit Theorem

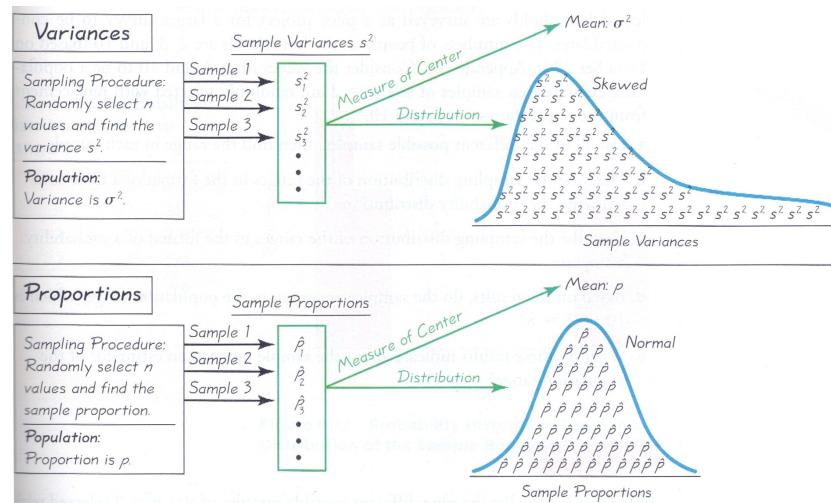
Sample size matters



CLT for dummies - Bunnies and Dragons



Sampling distributions



Bayesian Hypothesis tests

Bayesian posterior distribution provides the probability of the parameter given the observed data

Bayes factors

The Bayes Factor is the ratio of posterior probabilities of parameters under 2 hypotheses.

$$BF = \frac{\mathbb{P}[\theta \in \Theta_0 | x]}{\mathbb{P}[\theta \in \Theta_1 | x]}$$

This is symmetric w.r.t. the two hypotheses and on ratio scale, thus $BF = 2$ can be interpreted as H_0 is twice as likely as H_1 . If you want asymmetry you can add weights (i.e. "loss" when making the wrong decision). The decision is then based on $\frac{k_0}{k_1} \cdot BF$

Resampling Methods: Bootstrap distribution

The bootstrap estimator simulates the sampling distribution of an estimator by sampling with replacement from the original sample and calculating the estimator repeatedly for this 'new' sample. This means that 'artificial' additional samples are drawn from the distribution defined by the ECDF in order to approximate the underlying estimators distribution in cases when the CLT does not kick in.

Resampling Methods: Permutation Distribution

The permutation distribution is

- the exact distribution of any reasonably constructed estimator for a small enough sample, after calculating all possible values of the test statistic under permutations of the data points' labels
- or the approximation of the exact distribution of any reasonably constructed estimator for a sample too large to calculate all possible permutations

Wichtige Tests im Überblick

Parametrische und nicht-parametrische Tests im Vergleich

Parametrische Tests Parametrische Tests setzen eine bestimmte parametrische Verteilung der Daten voraus. Passende Schätzer für den Parameter führen zur Teststatistik. Unter der angenommenen Verteilung der Daten hat auch die Teststatistik eine spezielle Verteilung. Sehr oft hängt die Benennung des Tests mit der Verteilung der Teststatistik (nicht der Daten!) zusammen.

Beispiel: Student's t Test

- Annahme: Die Daten sind annähernd normalverteilt.
- Teststatistik: $t = \frac{\bar{x} - \mu_0}{\sigma}$
- Die resultierende Verteilung der Teststatistik t ist eine **student's t** Verteilung.

Durch die Verteilungsannahmen haben parametrische Tests die bestmögliche Trennschärfe von allen Tests mit demselben Frageziel, z.B. Lage der Daten. Allerdings gilt diese nur, wenn die Verteilungsannahmen auch erfüllt sind. Daher müssen sie VOR Durchführung des Tests überprüft werden.

Zwar gibt es nur wenige parametrische Szenarien, für die passende parametrische Test existieren. Dennoch sind parametrische Tests die am häufigsten verwendeten Hypothesentests, da diese wenige Szenarien die gängigsten Fragestellungen abdecken.

Nichtparametrische Tests Nichtparametrische Tests kommen ohne die Annahme einer konkreten parametrischen Verteilung der Daten aus. Sie kommen zur Anwendung, wenn die Voraussetzung, dass die Verteilung der Daten gilt, nicht erfüllt werden kann oder wie bei Verteilungshypothesen noch gar nicht bekannt oder nachgewiesen ist.

Allerdings kommen auch sie nicht ohne jedwede Annahmen aus:

- Ein Mindeststichprobengröße ist erforderlich, die deutlich oberhalb jener für parametrische Tests liegt. Etwa 80-100 Beobachtungen sind vernünftig als Startpunkt.
- Die Verteilung der Daten muss unimodal sein, sonst sind auch rangbasierte Lageschätzer nicht sinnvoll.

Typischerweise sind sie auf Rang- und Ordnungsstatistiken aufgebaut. Beispielsweise kann man anstatt des Mittelwerts die Lage einer Stichprobe auch zu testen, indem man als robustes Maß Ränge anstatt der Originaldaten heranzieht. Die Fragestellung ist gleich wie beim t-Test, aber die Verteilungsannahme muss nicht erfüllt werden. Ein Test ist etwa der Wilcoxon Rangsummen Test und wir werden uns diesen als Alternative für nichtnormalverteilte Daten ansehen.

Overview of tests by testing mechanism

- Classical **parametric tests**

Assume a certain distribution of the data characterised by parameters

calculate a sufficient statistics which follows a different type of distribution

- **non-parametric tests**

Do not assume a certain distribution

work with general properties like ranks and rank statistics

- **resampling methods**

sample an exact or approximate distribution of an estimator based on 'newly drawn samples' out of the original data sample

Überblick über die wichtigsten Tests nach Zielsetzung (Eigenschaft, die getestet wird)

Mittelwerttests und Proportionentests

- parametrische Tests

unter Annahme der Normalverteilung der Daten

- Gauss test für 1 oder 2 Stichproben
vergleicht die Mittelwerte von von Stichproben miteinander oder mit einem Referenzwert
- Exakter Student's t test für 1 oder 2 Stichproben
- Welch's test (approximativer student's t Test) für 1 oder 2 Stichproben
- ANOVA (=Analysis of Variance) für 2 oder mehr Stichproben

- non-parametric tests

- Wilcoxon Rangsummentest (für 1 Stichprobe)
- Wilcoxon-Mann-Whitney test (für 2 Stichproben)
- Kruskal-Wallis test (für 2 oder mehr Stichproben)

- Bayesian test for the mean (and proportions) (for 1 or more samples)
- resampling methods (bootstrapping, permutation test) (for 1 or more samples)
- parametrischer Test

unter Annahme der Binomialverteilung der Daten

- Gauss Test für 1 oder 2 Stichproben
vergleicht die Häufigkeiten von von Stichproben miteinander oder mit einem Referenzwert

- nichtparametrischer Test

unter Erstellung der Kontingenztafel der Daten

- χ^2 Test für 1 oder 2 Stichproben
special case: test for proportions (assuming Binomial distribution for which the proportion is estimated)

Varianztests

- parametrische Tests
 - unter Annahme der Normalverteilung der Daten
 - χ^2 test für 1 Stichprobe
eine Varianz gegen einen vorgegebenen Wert vergleichen
 - **F test** für 2 Stichproben
zwei Varianzen miteinander vergleichen
 - ANOVA (Analysis of Variance) für 2 oder mehr Stichproben
Spezialfall der F-Test um die Mittelwerte verschiedener Stichproben oder Modelle miteinander zu vergleichen
- Bayesian test for variance and Bayesian ANOVA
- sampling-based tests based on permutation or bootstrapping

Verteilungstest

- χ^2 -test für Homogenität
- χ^2 -test für Unabhängigkeit

Vergleich von Häufigkeitsverteilung zwischen 2 kategorialen Variablen mit 2 oder mehr Kategorien pro Variable

- Kolmogorow-Smirnow test, Cramer-von-Mises test

allgemeiner Test, ob die Daten eine spezifische vorgegebene Verteilung haben.

Achtung: Daten und Verteilungen müssen eindimensional und unimodal sein!

- Shapiro-Wilks test

Test, ob Daten eine Normalverteilung haben

Achtung: Daten müssen eindimensional und unimodal sein! Der Test detektiert schwere Ränder und Ausreißer nicht Multimodalität. Daher ist Datenexploration trotzdem wichtig!

Proportionentest

Proportionentest mit 1 Stichprobe

parametrischer Test

Daten Die angenommene Verteilung der Daten ist die **Binomial Verteilung** $Bin(n, p)$

Ziel Testen des **Parameters** p = die **Proportion**/der **Anteil** der 'Erfolge' bei n Versuchen

Schätzwert Relative Häufigkeit der Erfolge $\hat{p} = \frac{m}{n}$ mit Anzahl der Erfolge m

Die Teststatistik für unseren Proportionentest lautet

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Die angenommene Verteilung des Parameters ist die **Normalverteilung**

Teststatistik

Nullhypothese Unter der Nullhypothese wird bei nur 1 Stichprobe als Referenzwert der Erfolgswahrscheinlichkeit p_0 angenommen
zweiseitig $H_0 : p = p_0$
linksseitig $H_0 : p \geq p_0$
rechtsseitig $H_0 : p \leq p_0$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert

zweiseitig $H_A : p \neq p_0$
linksseitig $H_A : p < p_0$
rechtseitig $H_A : p > p_0$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

Wir verwenden bei der Schätzung der Prädiktionsintervalle, Konfidenzintervalle, Akzeptanz- und Verwerfungsbereiche sowie für die Ermittlung des p-Werts jeweils die Normalverteilung. Hier tritt der zentrale Grenzwertsatz (Central Limit Theorem) in Kraft, wonach die Verteilung des Mittelwerts der Daten unabhängig von der Datenverteilung eine Normalverteilung ist, wenn entsprechend viele Stichproben gezogen werden.

```
prop.test(x, n, p = NULL,  
          alternative = c("two.sided", "less", "greater"),  
          conf.level = 0.95, correct = TRUE)
```

Proportionentest

Wir verwerfen die Nullhypothese, wenn

- im Fall des **zweiseitigen Tests** der Wert der Teststatistik das untere $\alpha/2$ -Quantil $q_{\frac{\alpha}{2}}$ untertrifft oder das obere $\alpha/2$ -Quantil $q_{1-\frac{\alpha}{2}}$ übersteigt.
- im Fall des **linksseitigen Tests** der Wert der Teststatistik das untere α -Quantil q_α untertrifft.
- im Fall des **rechtsseitigen Tests** der Wert der Teststatistik das obere α -Quantil $q_{1-\alpha}$ übertrifft.

Example: Finding a CI for a proportion - Global warming

Proportion of Adults Believing in Global Warming

In a Pew Research Center poll, 1051 of 1501 randomly selected adults in the United States believe in global warming, so the sample proportion is $\hat{p} = 0.70$. Is this significantly different from random guessing?

Umsetzung in R:

```
binom.test(c(1051, 450), p = 0.5)
##
##  Exact binomial test
##
## data: c(1051, 450)
## number of successes = 1051, number of trials = 1501, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.6763127 0.7233011
## sample estimates:
## probability of success
##                      0.7001999

prop.test(1051, 1501, p = 0.5)
##
##  1-sample proportions test with continuity correction
##
## data: 1051 out of 1501, null probability 0.5
## X-squared = 239.84, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6761947 0.7231682
## sample estimates:
##          p
## 0.7001999
```

Die **Teststatistik** für dieses Szenario beträgt

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.7 - 0.5}{\sqrt{\frac{0.7(1-0.7)}{1501}}} \approx 16.9087$$

und ihre quadrierte Variante

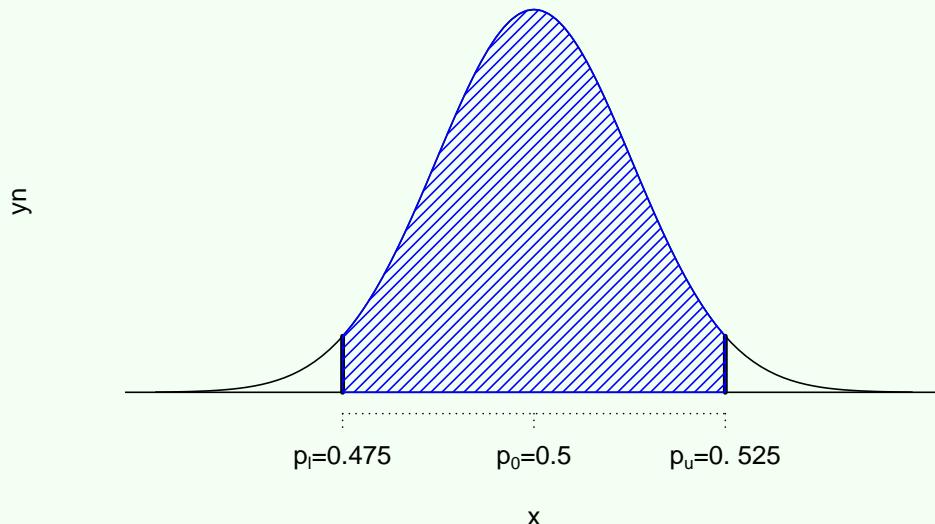
$$\left(\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \right)^2 = \left(\frac{0.7 - 0.5}{\sqrt{\frac{0.7(1-0.7)}{1501}}} \right)^2 \approx 286.5856$$

Diese weicht hier von der Teststatistik aus R ab, da im Hintergrund mit Kontingenztafeln und dem χ^2 -Test für Homogenität gerechnet wird, s.u..

Akzeptanzbereich=Prädiktionsintervall mit $1 - \alpha = 0.95$ Konfidenzniveau:

$$p_l = p_0 - q_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}} = 0.5 - 1.96 \sqrt{\frac{0.5(1-0.5)}{1501}} \approx 0.4747$$

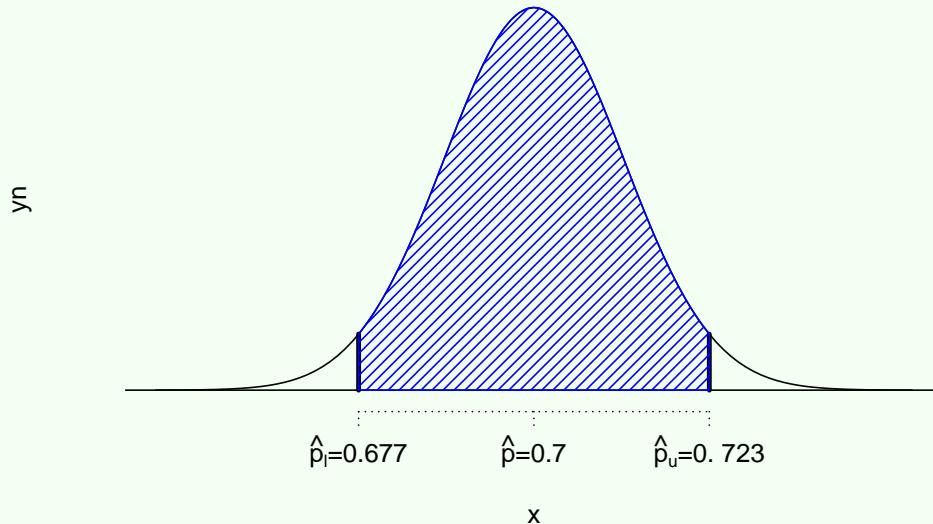
$$p_u = p_0 + q_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}} = 0.5 + 1.96 \sqrt{\frac{0.5(1-0.5)}{1501}} \approx 0.5253$$



Konfidenzintervall mit $1 - \alpha = 0.95$ Konfidenzniveau:

$$p_l = \hat{p} - q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7 - 1.96 \sqrt{\frac{0.7(1-0.7)}{1501}} \approx 0.6768$$

$$p_u = \hat{p} + q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7 + 1.96 \sqrt{\frac{0.7(1-0.7)}{1501}} \approx 0.7232$$



Interpretation: With a probability of 95%, the interval $[0.677, 0.723]$ contains the relative frequency of believers in global warming in the US. As the 0.5 mark is not contained in this interval, we may safely say that the majority of adults in the US believe in global warming. Thus, people have a strong belief which is significantly different from random guessing ($p_0 = 0.5$).

A statement summarizing the results 70% of United States adults believe that the earth is getting warmer. That percentage is based on a Pew Research Center poll of 1501 randomly selected adults in the United States. In theory, in 95% of such polls, the percentage should differ by no more than 2.3 percentage points in either direction from the percentage that would be found by interviewing all adults in the United States.

Estimating the Required sample size

Which sample size is needed?

- Sample size depends on the required **accuracy** (formula for $\text{Var}[\hat{p}]$ has n in the denominator, thus the standard deviation of \hat{p} decreases proportionally to \sqrt{n}).
- Calculation can be done by solving the corresponding formula

$$n = \frac{q_{1-\frac{\alpha}{2}}^2 \hat{p}(1-\hat{p})}{\underbrace{(p_u - \hat{p})^2}_E},$$

where $E = p_u - \hat{p} = \hat{p} - p_l$ denotes the desired margin of error.

- If no estimate \hat{p} is known, we simply assume $\hat{p} = 0.5$, yielding

$$n = \frac{q_{1-\frac{\alpha}{2}}^2}{4E^2}.$$

Medikamententest Genericon gegen Original

Bei einem Medikamententest bekommen 150 Testpersonen ein Genericon verabreicht. Bei 87 von ihnen tritt eine positive Wirkung ein. Für die Zulassung eines Genericons ist notwendig, dass es nicht schlechter als das Originalpräparat mit einer Wirkungsrate von 70% ist.

Medikamententest Genericon gegen Original

Hier wird die **Alternativhypothese** formuliert als ‘nicht schlechter’, was dasselbe bedeutet wie ‘Wirkungsrate größer oder gleich wie beim Original’.

```
prop.test(87,150,p=0.7,alternative = "greater")
##
## 1-sample proportions test with continuity correction
##
## data: 87 out of 150, null probability 0.7
## X-squared = 9.7222, df = 1, p-value = 0.9991
## alternative hypothesis: true p is greater than 0.7
## 95 percent confidence interval:
## 0.509528 1.000000
## sample estimates:
## p
## 0.58
```

Proportionentest zwischen 2 Stichproben

parametrischer Test

Daten Die angenommene Verteilung der Daten ist die **Binomial Verteilung** $Bin(n, p)$

Ziel Testen des **Parameters** p = die **Proportion**/der **Anteil** der 'Erfolge' bei **n Versuchen**

Schätzwert Die über 2 Stichproben gepoolte relative Häufigkeit der Erfolge $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ und separate relative Häufigkeiten $p_i = \frac{x_i}{n_i}$

Die Teststatistik für unseren Proportionentest lautet

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

, wobei hier der Unterschied der Proportionen $p_1 - p_2$ in H_0 festgelegt ist. Dabei bedeutet $p_1 - p_2 = 0$, dass $p_1 = p_2$.

Die Voraussetzung ist, dass $x_i = n_i p_i \geq 5$ und $n_i - x_i = n_i(1 - p_i) \geq 5$ für beide Stichproben gelten.

Teststatistik

Die angenommene Verteilung des Parameters ist die **Normalverteilung**

Nullhypothese Der Unterschied der Proportionen $p_1 - p_2$ wird getestet und daher bedeutet $p_1 - p_2 = 0$, dass $p_1 = p_2$ gilt.
zweiseitig $H_0 : p_1 = p_2$
linksseitig $H_0 : p_1 \geq p_2$
rechtsseitig $H_0 : p_1 \leq p_2$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert

zweiseitig $H_A : p_1 \neq p_2$

linksseitig $H_A : p_1 < p_2$

rechtseitig $H_A : p_1 > p_2$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

Wir verwenden bei der Schätzung der Prädiktionsintervalle, Konfidenzintervalle, Akzeptanz- und Verwerfungsbereiche sowie für die Ermittlung des p-Werts jeweils die Normalverteilung. Hier tritt der zentrale Grenzverteilungssatz (Central Limit Theorem) in Kraft, wonach die Verteilung des Mittelwerts der Daten unabhängig von der Datenverteilung eine Normalverteilung ist, wenn entsprechend viele Stichproben gezogen werden.

Proportionentest

Wir verwerfen die Nullhypothese, wenn

- im Fall des **zweiseitigen Tests** der Wert der Teststatistik das untere $\alpha/2$ -Quantil $q_{\alpha/2}$ untertrifft oder das obere $\alpha/2$ -Quantil $q_{1-\alpha/2}$ übersteigt.
- im Fall des **linksseitigen Tests** der Wert der Teststatistik das untere α -Quantil q_α untertrifft.
- im Fall des **rechtsseitigen Tests** der Wert der Teststatistik das obere α -Quantil $q_{1-\alpha}$ übertrifft.

Example: Inference about two proportions

Do Airbags Save Lives? The table below lists results from a simple random sample of front-seat occupants involved in car crashes (based on data from “Who Wants Airbags?” by Meyer and Finney, *Chance*, Vol. ~18, No. ~2). Use a 0.05 significance level to test the claim that the fatality rate of occupants is lower for those in cars equipped with airbags.

	Airbag	No Airbag
Occupant Fatalities	41	52
Total Number of Occupants	11,541	9,853

Mathematical Formulation of the hypotheses:

$$\begin{aligned} H_0 : \quad p_1 &\geq p_2 \\ H_1 : \quad p_1 &< p_2 \end{aligned}$$

Requirements:

1. Two independent simple random samples \Rightarrow OK!
2. $41 > 5$ and $11541 > 5$ and $52 > 5$ and $9801 > 5 \Rightarrow$ OK!

```
deaths<-c(41,52); total<-c(11541,9853)
prop.test(deaths,total,alternative="less")

 2-sample test for equality of proportions with continuity correction

data: deaths out of total
X-squared = 3.2667, df = 1, p-value = 0.03535
alternative hypothesis: less
95 percent confidence interval:
-1.0000000000 -0.0001238461
sample estimates:
prop 1      prop 2
0.003552552 0.005277580
```

Medikamententest Präparat gegen Placebo

Bei einem Medikamententest bekommen 150 Testpersonen ein Placebo und 200 das Originalpräparat verabreicht. Bei 87 Personen der Placebogruppe und 139 Personen der Präparatsgruppe tritt eine positive Wirkung ein. Für die Zulassung eines Medikaments ist notwendig, dass es besser als das Placebo wirkt.

Es gibt hier 2 Arten, die Alternativhypothese zu formulieren, welche für die Entscheidung äquivalent sind:

1. Original ist besser als das Placebo = Wirkungsrate beim Original ist größer als beim Placebo

```
prop.test(x=c(139, 87), n=c(200, 150), alternative = "greater")
```

```
##  
## 2-sample test for equality of proportions with continuity correction  
##  
## data: c(139, 87) out of c(200, 150)  
## X-squared = 4.4652, df = 1, p-value = 0.0173  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## 0.02395317 1.00000000  
## sample estimates:  
## prop 1 prop 2  
## 0.695 0.580
```

2. Original ist besser als das Placebo = Wirkungsrate beim Placebo ist kleiner als beim Original

```
prop.test(x=c(87, 139), n=c(150, 200), alternative = "less")
```

```
##  
## 2-sample test for equality of proportions with continuity correction  
##  
## data: c(87, 139) out of c(150, 200)  
## X-squared = 4.4652, df = 1, p-value = 0.0173  
## alternative hypothesis: less  
## 95 percent confidence interval:  
## -1.00000000 -0.02395317  
## sample estimates:  
## prop 1 prop 2  
## 0.580 0.695
```

Entscheidung: auf 5% Niveau Nullhypothese verworfen, 1% Niveau Nullhypothese beibehalten
Daher ist das ein schwach signifikantes Szenario!

Mittelwertschätzung und Tests auf Lage

1 Stichproben t-Test

parametrischer Test

Daten Die angenommene Verteilung der Daten ist die **Normalverteilung** $N(\mu, \sigma)$

Ziel Testen des **Parameters μ** = der **Mittelwert** von n Messungen

Schätzwert arithmetische Mittelwert $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$

Teststatistik Wir unterscheiden hier 2 Varianten:

- Varianz σ ist bekannt

Die Teststatistik lautet

$$z = \frac{\mu - \hat{x}}{\sigma / \sqrt{n}}$$

Die angenommene Verteilung der Teststatistik ist die **Normalverteilung**.

- Varianz σ ist unbekannt

Die Teststatistik lautet

$$t = \frac{\mu - \hat{x}}{\hat{s} / \sqrt{n}}$$

Die angenommene Verteilung der Teststatistik ist die **student's t-Verteilung** mit $df = n - 1$ Freiheitsgraden.

Nullhypothese Unter der Nullhypothese wird bei nur 1 Stichprobe als Referenzwert der Mittelwert μ_0 angenommen
zweiseitig $H_0 : \mu = \mu_0$

linksseitig $H_0 : \mu \geq \mu_0$

rechtsseitig $H_0 : \mu \leq \mu_0$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert

zweiseitig $H_A : \mu \neq \mu_0$

linksseitig $H_A : \mu < \mu_0$

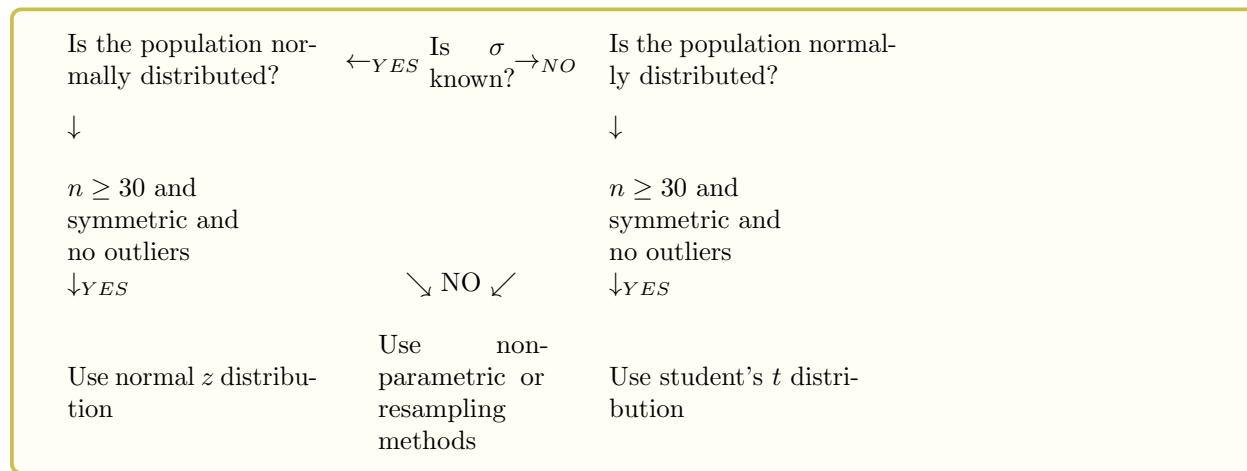
rechtseitig $H_A : \mu > \mu_0$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

Wir verwenden bei der Schätzung der Prädiktionsintervalle, Konfidenzintervalle, Akzeptanz- und Verwerfbereiche sowie für die Ermittlung des p-Werts jeweils die student's t-Verteilung. Das ist eine Anwendung der mathematischen Statistik, wonach die Verteilung des Mittelwerts von normalverteilten Daten eine t-Verteilung ist, wenn entsprechend viele Stichproben gezogen werden.

```
t.test(x = daten, mu = 0, conf.level = 0.95,
       alternative = c("two.sided", "less", "greater"))
```

Choosing the appropriate distribution



Example: Testing a claim about a mean with σ unknown

Boat Safety

Data set 1 from Triola contains the following information about body weights of people on a boat: $n = 40$, $\bar{x} = 171.1822934$ lb, $s = 23.738052$ lb. Do not assume that the value of σ is known. Use these results to test the claim that men have a mean weight greater than 166.3 lb, which was the weight in the National Transportation and Safety Board's recommendation M-04-04. Use a 0.05 significance level.

Requirement check: (1) simple random sample, (2) σ is not known, (3) $n > 30$ or the population is normally distributed. \Rightarrow OK!

Getestet wird hier also die **einseitige Hypothese**

$$H_0 : \mu \leq 166.3$$

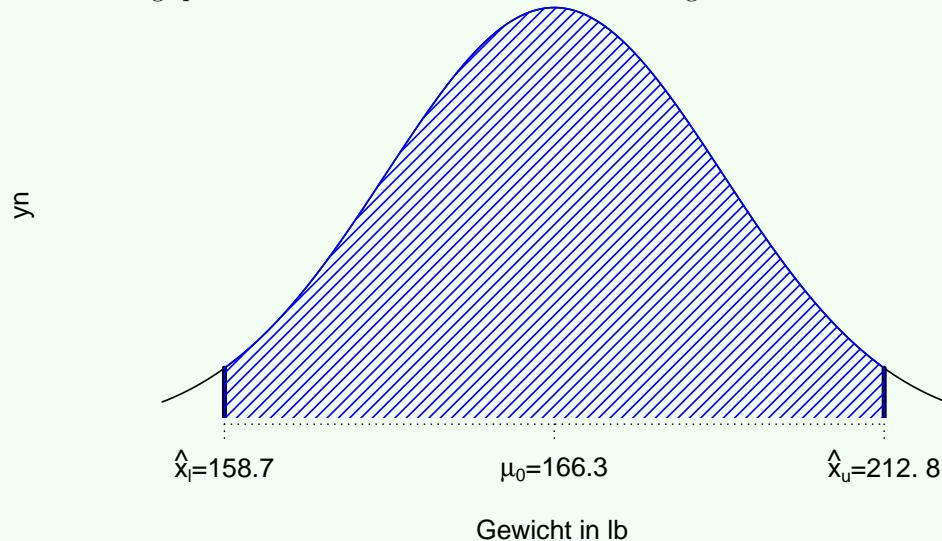
$$H_1 : \mu > 166.3$$

Die **Teststatistik** für dieses Szenario beträgt $t = \frac{\bar{x} - \mu_{\bar{x}}}{s/\sqrt{n}} = 1.3008$

Der symmetrische **Akzeptanzbereich=Prädiktionsintervall** mit $1 - \alpha = 0.95$ Konfidenzniveau für die zukünftigen Datenwerte lautet:

$$\begin{aligned}\hat{x}_l &= \mu_0 - q_{1-\frac{\alpha}{2}}^t \sigma = 166.3 - 2.022690923.738052 \approx 118.2853 \text{ lb} \\ \hat{x}_u &= \mu_0 + q_{1-\frac{\alpha}{2}}^t \sigma = 166.3 + 2.022690923.738052 \approx 214.3147 \text{ lb}\end{aligned}$$

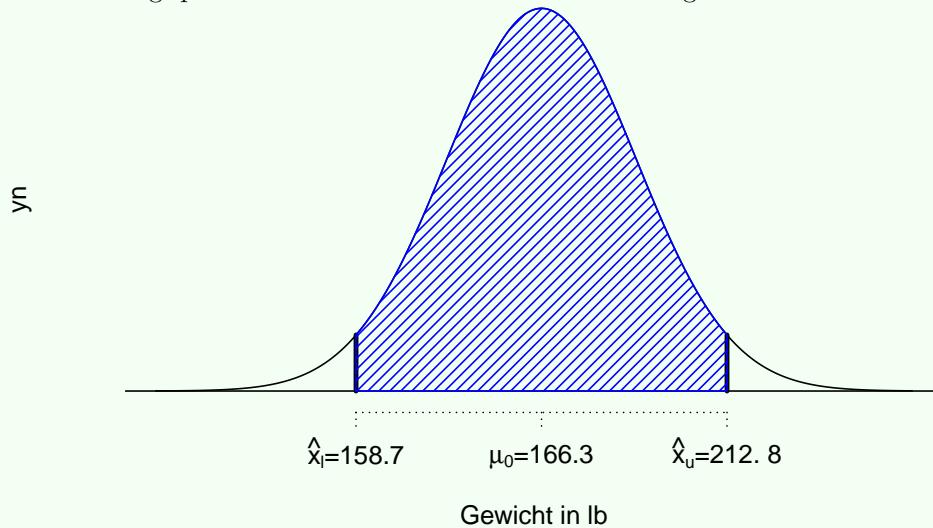
mit q^t als t-Verteilungsquantilen mit $\nu = df = 40 - 1 = 39$ Freiheitsgraden.



Der symmetrische **Akzeptanzbereich=Prädiktionsintervall** mit $1 - \alpha = 0.95$ Konfidenzniveau für den zukünftigen Mittelwert:

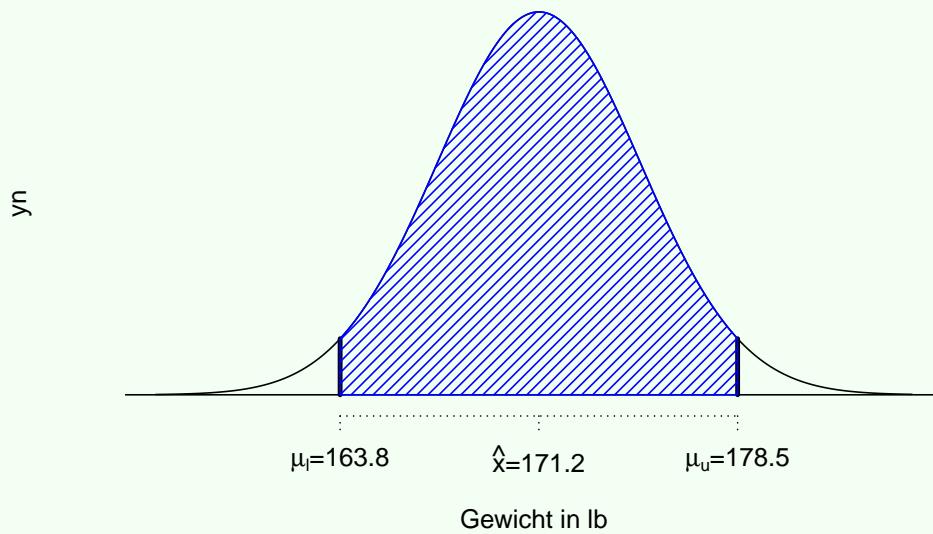
$$\begin{aligned}\hat{x}_l &= \mu_0 - q_{1-\frac{\alpha}{2}}^t \sqrt{\frac{\sigma^2}{n}} = 166.3 - 2.0226909 \sqrt{\frac{563.4951104}{40}} \approx 158.7082 \text{ lb} \\ \hat{x}_u &= \mu_0 + q_{1-\frac{\alpha}{2}}^t \sqrt{\frac{\sigma^2}{n}} = 166.3 + 2.0226909 \sqrt{\frac{563.4951104}{40}} \approx 173.8918 \text{ lb}\end{aligned}$$

mit q^t als t-Verteilungsquantilen mit $\nu = df = 40 - 1 = 39$ Freiheitsgraden.



Konfidenzintervall mit $1 - \alpha = 0.95$ Konfidenzniveau für die Einschätzung des Erwartungswerts auf Basis des Mittelwertes der Stichprobe:

$$\begin{aligned}\mu_l &= \hat{x} - q_{1-\frac{\alpha}{2}}^t \sqrt{\frac{\sigma^2}{n}} = 171.1822934 - 2.0226909 \sqrt{\frac{563.4951104}{40}} \approx 163.5905 \text{ lb} \\ \mu_u &= \hat{x} + q_{1-\frac{\alpha}{2}}^t \sqrt{\frac{\sigma^2}{n}} = 171.1822934 + 2.0226909 \sqrt{\frac{563.4951104}{40}} \approx 178.7741 \text{ lb}\end{aligned}$$



Boat Safety

Data set 1 from Triola contains the following information about body weights of people on a boat: $n = 40$, $\bar{x} = 171.1822934$ lb, $s = 23.738052$ lb. Do not assume that the value of σ is known. Use these results to test the claim that men have a mean weight greater than 166.3 lb, which was the weight in the National Transportation and Safety Board's recommendation M-04-04. Use a 0.05 significance level.

Requirement check: (1) simple random sample, (2) σ is not known, (3) $n > 30$ or the population is normally distributed. \Rightarrow OK!

- Test statistics $t = \frac{\bar{x} - \mu_{\bar{x}}}{s/\sqrt{n}} = 1.3008$
- P-value method: Find area to the **right** of the test statistic $t = 1.301$: P-value is 0.1004812 .
- Confidence interval method: **90% confidence interval**: $164.8584254 \text{ lb} < \mu < \infty \text{ lb}$, which contains the assumed $\mu = 166.3$.

Because we fail to reject the null, we conclude that there is not sufficient evidence to support a conclusion that the population mean is greater than 166.3 lb, as in the National Transportation and Safety Board's recommendation.

Example in R As above, we can obtain the same information about the hypothesis, the p-value and the confidence interval from the R output.

```
t.test(x, mu=166.3, alternative = "greater")
One Sample t-test

data: x
t = 1.3008, df = 39, p-value = 0.1005
alternative hypothesis: true mean is greater than 166.3
95 percent confidence interval:
164.8584      Inf
sample estimates:
mean of x
171.1823
```

Einwaage

Ein Laborant wiegt eine Substanz ein, wobei er im Mittel 10 mg ein Einwaagegewicht erreichen soll. Im Protokoll verzeichnet er die folgenden Messungen:

10.5 mg, 10.2 mg, 9.8 mg, 9.5 mg, 9.7 mg, 10.1 mg

Diese codieren wir in R mit

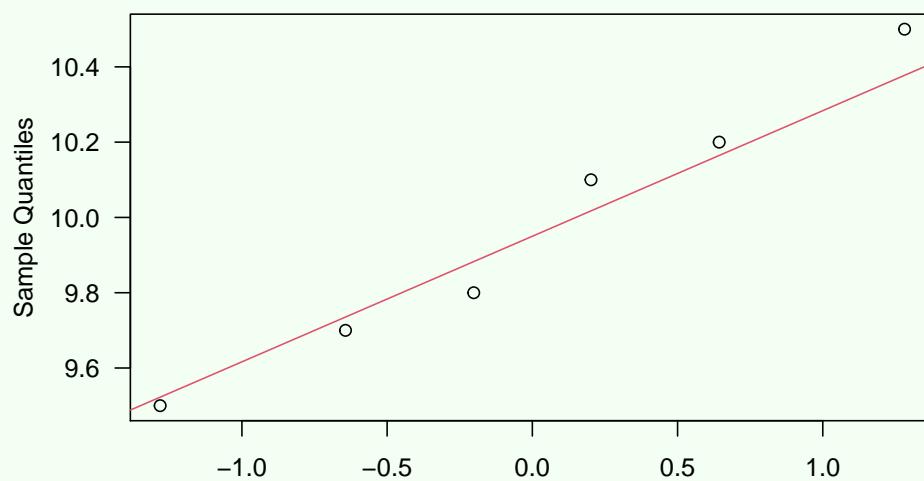
```
messungen<-c(10.5,10.2,9.8,9.5,9.7,10.1)
```

Nun möchte er feststellen, ob er im Mittel von seiner Vorgabe auf einem Signifikanzniveau von 1% abweicht.

Wir beginnen mit der Überprüfung der Normalverteilung der Daten, um uns für den parametrischen oder nichtparametrischen Test zu entscheiden.

```
qqnorm(messungen,las=1)  
qqline(messungen,col=2)
```

Normal Q-Q Plot



Theoretical Quantiles

Da die Daten annähernd normalverteilt und ohne Ausreißer sind, dürfen wir den t-Test verwenden.

```
t.test(x=messungen,mu=10,alternative = "two.sided",conf.level = 0.99)  
  
##  
## One Sample t-test  
##  
## data: messungen  
## t = -0.2225, df = 5, p-value = 0.8327  
## alternative hypothesis: true mean is not equal to 10  
## 99 percent confidence interval:  
## 9.362592 10.570741  
## sample estimates:  
## mean of x  
## 9.966667
```

Hätten wir das nicht erkannt, gäbe es auch den nicht-parametrischen Test als Alternative, wobei dieser hier problematisch ist, da wir nicht wissen, ob das Ergebnis nicht signifikant ist, weil wir zu wenig Beobachtungen haben, den weniger trennscharfen Test verwendet haben oder weil die Nullhypothese tatsächlich gültig ist.

Der Wilcoxon-signed-rank Test ist hier der passende Test.

```
wilcox.test(x=messungen, mu=10, alternative = "two.sided", conf.level = 0.99)
## Warning in wilcox.test.default(x = messungen, mu = 10, alternative =
## "two.sided", : cannot compute exact p-value with ties
##
## Wilcoxon signed rank test with continuity correction
##
## data: messungen
## V = 9, p-value = 0.833
## alternative hypothesis: true location is not equal to 10
```

Required sample size

Which sample size is needed?

- Sample size depends on the required **accuracy** (formula for $\mathbb{V}[\bar{x}]$ has n in the denominator, thus the standard deviation of \bar{x} decreases proportionally to \sqrt{n}).
- Calculation can be done by solving formula for the confidence interval

$$n = \left(\frac{q_{1-\frac{\alpha}{2}} \sigma}{E} \right)^2.$$

where again $E = \mu_u - \bar{x} = \bar{x} - \mu_l$ denotes the desired margin of error.

Example: Required sample size

IQ Scores of Statistics Students Assume that we want to estimate the mean IQ score for the population of statistics students. How many must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

For a 95% confidence interval, we have $\alpha = 0.05$, so $q_{1-\frac{\alpha}{2}} = 1.96$. Because we want the sample mean to be within 3 IQ points of μ , the margin of error is $E = 3$. Also, $\sigma = 15$. We get:

$$n = \left(\frac{q_{1-\frac{\alpha}{2}} \sigma}{E} \right)^2 = \left(\frac{1.96 \times 15}{3} \right)^2 = 96.04 \approx 97 \quad (\text{rounded up}).$$

Interpretation:

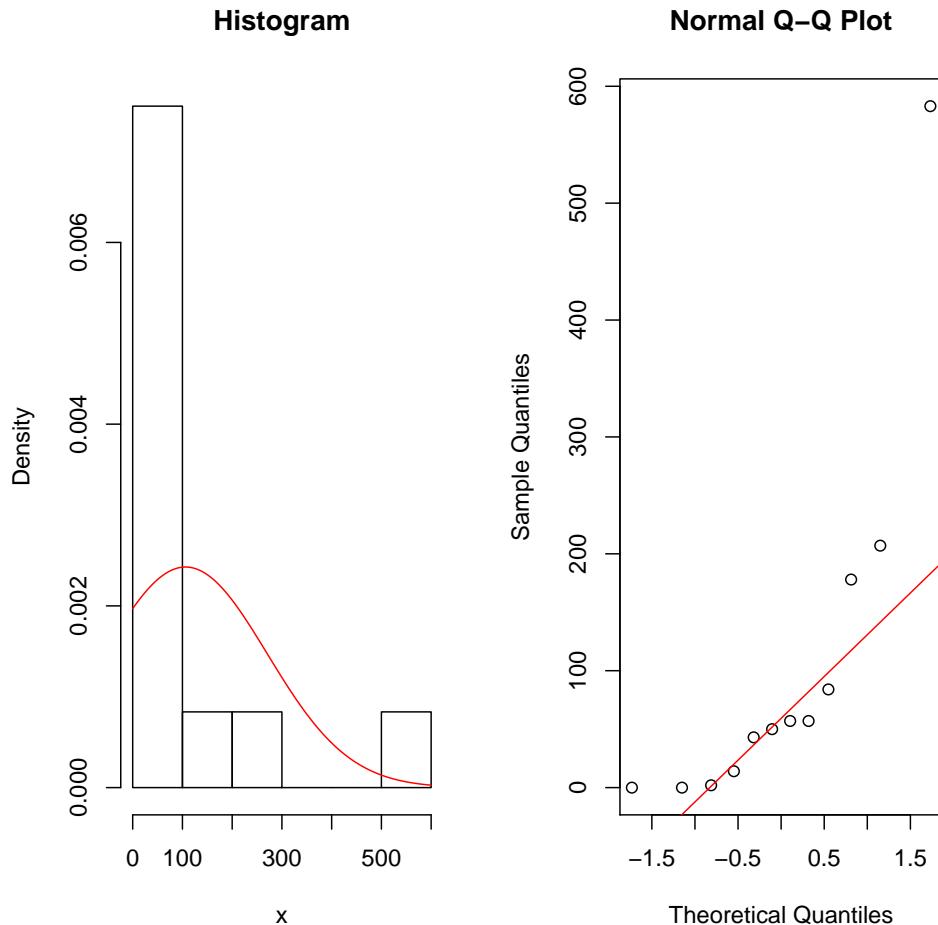
Among the thousands of statistics students, we need to obtain a simple random sample of at least 97 students. Then we need to get their IQ scores. With a simple random sample of only 97 stats students, we will be 95% confident that the sample mean \bar{x} is within 3 IQ points of the true population mean μ .

Example: What can go wrong!

Confidence Interval for Alcohol in Video Games Twelve different video games showing substance use were observed. The duration times (in seconds) of alcohol were recorded, with the times listed below (based on data from “Content and Ratings of Teen-Rated Video Games,’’ by Haninger and Thompson, *Journal of the American Medical Association*, Vol.~291, No.~7). The design of the study justifies the assumption that the sample can be treated as a simple random sample. Use the sample data to construct a 95% confidence interval estimate of μ , the mean duration time that the video showed the use of alcohol.

84 14 583 50 0 57 207 43 178 0 2 57

Caveat: $n = 12 < 30$, thus we must determine whether the data appear to be from a normal population.



Example: Finding a CI for μ (σ unknown)

The requirements are not satisfied \Rightarrow STOP!

Let's continue anyway:

- $n = 12$
- $\bar{x} = 106.25$
- $s \approx 164.33$
- $\alpha = 0.05, df = n - 1 = 11 \Rightarrow t_{\alpha/2} \approx 2.20$
- $E = t_{\alpha/2} \frac{s}{\sqrt{n}} \approx 104.42$
- $\bar{x} - E < \mu < \bar{x} + E \Rightarrow 1.84 < \mu < 210.66$

This result is highly questionable because it assumes incorrectly that the requirements are satisfied! Other methods such as nonparametric estimation or bootstrap resampling are needed, the latter yielding a confidence interval of $35.3 < \mu < 205.6$ (Triola).

2 Stichproben t-Test und Welch's t-Test

parametrischer Test

Daten Die angenommene Verteilung der Daten ist die **Normalverteilung** $N(\mu, \sigma)$

Ziel Testen des **Parameters** μ = der **Mittelwert** von n Messungen

Schätzwert arithmetische Mittelwert $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$

Teststatistik Wir unterscheiden hier 2 Varianten:

- 2 Stichproben-t-Test
Varianzen beider Stichproben sind gleich
Die Teststatistik lautet

$$z = \frac{(\mu_1 - \mu_2) - (\hat{x}_1 - \hat{x}_2)}{\hat{s}/\sqrt{n}}$$

Die angenommene Verteilung der Teststatistik ist die **student's t-Verteilung** mit $df = n_1 + n_2 - 2$ Freiheitsgraden.

- Welch's t-Test
Varianzen beider Stichproben sind verschieden
Die Teststatistik lautet

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Die angenommene Verteilung der Teststatistik ist die **student's t-Verteilung** mit $df = n_1 + n_2 - 2 - \Delta$ Freiheitsgraden.

Nullhypothese Unter der Nullhypothese wird bei nur 1 Stichprobe als Referenzwert der Mittelwert μ_0 angenommen
zweiseitig $H_0 : \mu_1 = \mu_2$
linksseitig $H_0 : \mu_1 \geq \mu_2$
rechtsseitig $H_0 : \mu_1 \leq \mu_2$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert
zweiseitig $H_A : \mu_1 \neq \mu_2$
linksseitig $H_A : \mu_1 < \mu_2$
rechtseitig $H_A : \mu_1 > \mu_2$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

Wir verwenden bei der Schätzung der Prädiktionsintervalle, Konfidenzintervalle, Akzeptanz- und Verwerfungsbereiche sowie für die Ermittlung des p-Werts jeweils die student's t-Verteilung. Das ist eine Anwendung der mathematischen Statistik, wonach die Verteilung des Mittelwerts von normalverteilten Daten eine t-Verteilung ist, wenn entsprechend viele Stichproben gezogen werden.

```
# to test if the variances are the same
var.test(x, y, ratio = 1, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
# to perform the actual Welch's t-test for 2 samples with different variances
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Nichtparametrischer Test zum Vergleich von Lage

nichtparametrischer Test

Daten Die Daten haben keine angenommene Verteilung, müssen aber **unimodal** sein.

Ziel Testen der Eigenschaft Lage

Schätzwert durch die **Ränge** der Beobachtungen

Teststatistik Wir unterscheiden hier 2 Varianten:

- Wilcoxon Rangsummentest=Wilcoxon signed rank Test für 1 Stichprobe
Die Teststatistik lautet

$$W = \sum_{i=1}^{N_r} [\text{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]$$

- Wilcoxon-Mann-Whitney-Test=Mann-Whitney-U-Test für 2 Stichproben
Die Teststatistik lautet

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

Nullhypothese Unter der Nullhypothese wird bei nur 1 Stichprobe als Referenzwert der Mittelwert μ_0 angenommen
zweiseitig $H_0 : \text{Lage}_1 = \text{Lage}_2$
linksseitig $H_0 : \text{Lage}_1 \geq \text{Lage}_2$
rechtsseitig $H_0 : \text{Lage}_1 \leq \text{Lage}_2$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert
zweiseitig $H_A : \text{Lage}_1 \neq \text{Lage}_2$
linksseitig $H_A : \text{Lage}_1 < \text{Lage}_2$
rechtseitig $H_A : \text{Lage}_1 > \text{Lage}_2$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

```
wilcox.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
             mu = 0, paired = FALSE, exact = NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95)
```

Abhängige Stichproben t-Test

parametrischer Test

Daten Die Daten zweier Stichproben werden abhängig gezogen, etwa dieselben Messeinheiten zu unterschiedlichen Zeitpunkten. Anstatt der Originialdaten werden die Differenzen der Werte jeder Messeinheit $\Delta_i = x_{1;i} - x_{2;i}$

Die angenommene Verteilung der Differenzen ist die **Normalverteilung** $N(\mu_\Delta, \sigma_\Delta)$

Ziel Testen des **Parameters** μ_Δ = der **Mittelwert** von n Differenzen der Messungen derselben Messeinheit

Schätzwert arithmetische Mittelwert $\hat{\mu}_\Delta = \frac{1}{n} \sum_{i=1}^n \Delta_i$

Die Teststatistik lautet

$$t = \frac{\mu_\Delta - \hat{\mu}_\Delta}{\hat{s}_\Delta / \sqrt{n}}$$

Die angenommene Verteilung der Teststatistik ist die **student's t-Verteilung** mit $df = n - 1$ Freiheitsgraden.

Teststatistik

Nullhypothese Unter der Nullhypothese wird bei nur 1 Stichprobe als Referenzwert der Mittelwert Δ_0 angenommen
zweiseitig $H_0 : \mu_\Delta = \Delta_0$
linksseitig $H_0 : \mu_\Delta \geq \Delta_0$
rechtsseitig $H_0 : \mu_\Delta \leq \Delta_0$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert
zweiseitig $H_A : \mu_\Delta \neq \Delta_0$
linksseitig $H_A : \mu_\Delta < \Delta_0$
rechtseitig $H_A : \mu_\Delta > \Delta_0$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

Wir verwenden bei der Schätzung der Prädiktionsintervalle, Konfidenzintervalle, Akzeptanz- und Verwerfungsbereiche sowie für die Ermittlung des p-Werts jeweils die student's t-Verteilung. Das ist eine Anwendung der mathematischen Statistik, wonach die Verteilung des Mittelwerts von normalverteilten Daten eine t-Verteilung ist, wenn entsprechend viele Stichproben gezogen werden.

```
t.test(x = daten1, y=daten2, mu = 0, conf.level = 0.95, paired=TRUE,  
       alternative = c("two.sided", "less", "greater"))
```

Unterschied zwischen Tests von unabhängigen und abhängigen Daten Die folgenden Aussagen gelten für parametrische Tests unter der Voraussetzung der Normalverteilung.

- Bei **unabhängigen Daten** wird beim Testen der **Unterschied der Mittelwerte** zwischen zwei separaten Stichprobenszenarien getestet.
- Bei **abhängigen Daten** wird der **Mittelwert der Differenzen (Unterschiede)** zwischen zwei zusammenhängenden Stichprobenszenarien mit 0 verglichen.

Für nichtparametrische tests lässt sich analog verallgemeinern, dass

- Bei **unabhängigen Daten** wird beim Testen der **Unterschied der Lokation (Lage)** zwischen zwei separaten Stichprobenszenarien getestet.
- Bei **abhängigen Daten** wird die **Lage der Differenzen (Unterschiede)** zwischen zwei zusammenhängenden Stichprobenszenarien mit 0 verglichen.

R Beispiel zu Tests von 2 Stichproben

Beim Vergleichen der Lage/Mitte zweier Stichproben wird stets nach folgendem Schema vorgegangen, um die Eigenschaften der Daten und des Daten generierenden Prozesses zu berücksichtigen:

1. Wurden die beiden Stichproben **unabhängig** (d.h. unterschiedliche Personen oder experimentelle Einheiten werden beobachtet) oder **abhängig** (d.h. dieselben oder einander eindeutig zugeordnete Personen oder experimentelle Einheiten werden beobachtet) voneinander gezogen?
2. Sind die Daten **normalverteilt**?
Daraus ergibt sich, ob ein **parametrischer Test** oder ein **nichtparametrischer Test** gewählt werden muss.

Die Daten, die wir benutzen sind:

```
# For the examples we use the following data  
set.seed(23476864)
```

```
patienten1<-rnorm(100,mean=80,sd=10)  
patienten2abh<-patienten1-3 + rnorm(100)  
patienten2unabh<-rnorm(100,mean=87,sd=10)  
patienten3<-sapply(72+rgamma(100,shape = 10,rate = 2)+abs(rt(100,df=1)),FUN=function(x){min(x,115)+runif(1)})
```

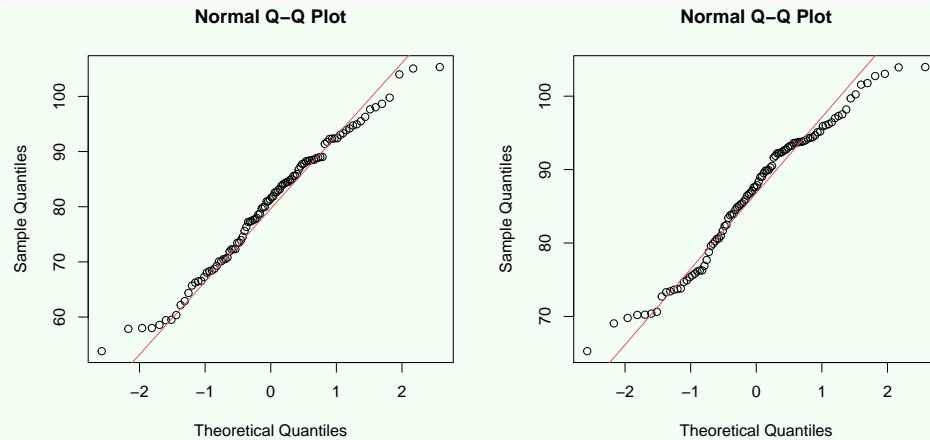
Im folgenden **Beispiel** wird ein Medikamententest durchgeführt, bei dem ein blutdrucksenkendes Mittel 100 Testpersonen aus Stichprobe **patienten1** verabreicht wird. Als Vergleichsgruppe in Stichprobe **patienten2unabh** wird 100 ein Placebo verabreicht.

Wir treffen die Überlegungen von zuvor:

1. Hier wurden die beiden Stichproben **unabhängig** voneinander gezogen, da andere Patienten das Originalpräparat und das Placebo erhalten.
2. Sind die Daten **normalverteilt**?
Das müssen wir erst feststellen durch Datenexploration von 2 unabhängigen Stichproben.

Überprüfung der Normalverteilung der Daten

```
par(mfrow=c(1,2))
qqnorm(patienten1); qqline(patienten1,col=2)
qqnorm(patienten2unabh); qqline(patienten2unabh,col=2)
```



```
par(mfrow=c(1,1))
```

Wir erkennen, dass beide Stichprobe annähernd normalverteilt sind und wir daher sinnvollerweise den t-Test zum Vergleich von 2 Stichproben durchführen dürfen. Wir können diese visuelle Exploration auch durch einen Test überprüfen, den Shapiro-Wilks-Test für Normalverteilung:

```
shapiro.test(patienten1)

##
##  Shapiro-Wilk normality test
##
## data: patienten1
## W = 0.98367, p-value = 0.2537
shapiro.test(patienten2unabh)

##
##  Shapiro-Wilk normality test
##
## data: patienten2unabh
## W = 0.96847, p-value = 0.01691
```

Beide sind nicht signifikant, wodurch die Nullhypothese der Normalverteilung nicht verworfen wird. Wir können also, wie auch aus der graphischen Exploration geschlossen, sinnvollerweise den t-Test zum Vergleich von 2 Stichproben durchführen.

t-Test für 2 unabhängige Stichproben

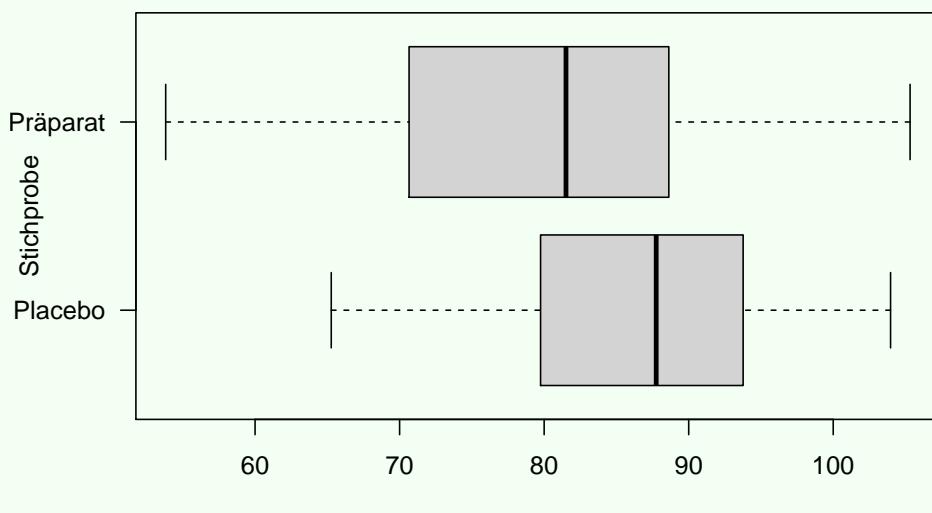
- Sind die Varianzen der Messungen der beiden Stichproben gleich groß oder verschieden?
Wir betrachten dafür den Boxplot als robuste Darstellung und können im Falle, dass wir uns unsicher sind, auch einen Test für Gleichheit der Varianzen durchführen.
 1. Die Varianzen sind annähernd gleich groß.
Exakter student t-Test für 2 unabhängige Stichproben
 2. Die Varianzen sind deutlich unterschiedlich
Welch student t-Test für 2 unabhängige Stichproben
Im Zweifelsfall wählen wir immer den Welch t-Test.

Überprüfung der Gleichheit der Varianzen

```
daten<-data.frame(Blutdruck=c(patienten1,patienten2unabh),
  Stichprobe=factor(c(rep("Präparat",length(patienten1)),rep("Placebo",length(patienten2unabh)))))
```

```
boxplot(Blutdruck~Stichprobe,data=daten,main="Vergleich der Stichproben",horizontal = TRUE,las=1)
```

Vergleich der Stichproben



Anhand der Boxplots erkennen wir, dass die Varianzen etwa gleich groß sind. Wir können dies auch noch mit einem Test für Varianzen überprüfen:

```
var.test(patienten1,patienten2unabh)

##
## F test to compare two variances
##
## data: patienten1 and patienten2unabh
## F = 1.6256, num df = 99, denom df = 99, p-value = 0.01642
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.093793 2.416070
## sample estimates:
## ratio of variances
##          1.625633
```

Dieser Test hat als Nullhypothese, dass die Varianzen gleich sind. Da die H_0 hier nicht verworfen wird, können wir davon ausgehen, dass wir mit gleichen Varianzen arbeiten.

Durchführung des 2-Stichproben-t-Test

```
t.test(patienten1,patienten2unabh,var.equal = TRUE,conf.level = 0.95)
##
## Two Sample t-test
##
## data: patienten1 and patienten2unabh
## t = -4.2398, df = 198, p-value = 3.43e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -9.432623 -3.443650
## sample estimates:
## mean of x mean of y
## 80.30851 86.74665
```

Die Teststatistik $t=-4.2398315$ hat eine t-Verteilung mit $df=198$ Freiheitsgraden, welche allerdings nur unter Normalverteilung der Daten Gültigkeit hat, was wir im Vorfeld überprüft haben.

Es wurde hier die zweiseitige Alterntivhypothese, dass die beiden Mittelwerte $\mu_1=80.3085122$ mmHG und $\mu_2=86.746649$ mmHG der beiden Stichproben verschieden sind, getestet. Der p-Wert des Tests ist 3.4299441×10^{-5} , was bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau verworfen wird, auf dem 1% Signifikanzniveau verworfen wird und auf dem 0.1% Signifikanzniveau verworfen wird. Der Testoutput beinhaltet auch das 95%-Konfidenzintervall, das durch den Code-Teil `conf.level = 0.95` eingestellt wurde. Dieses reicht von -9.4326234 bis -3.4436503 mmHg.

Durchführung des Welch' 2-Stichproben-t-Test

Alternativ hätten wir auch "auf Nummer Sicher" gehen können und den Welch t-Test wählen können, welcher die Standardeinstellung ist:

```
t.test(patienten1,patienten2unabh,conf.level = 0.95)
##
## Welch Two Sample t-test
##
## data: patienten1 and patienten2unabh
## t = -4.2398, df = 187.36, p-value = 3.509e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -9.433669 -3.442604
## sample estimates:
## mean of x mean of y
## 80.30851 86.74665
```

Die Teststatistik $t=-4.2398315$ hat eine t-Verteilung mit $df=187.3621778$ Freiheitsgraden, welche allerdings nur unter Normalverteilung der Daten Gültigkeit hat, was wir im Vorfeld überprüft haben. Es wurde hier die zweiseitige Alterntivhypothese, dass die beiden Mittelwerte $\mu_1=80.3085122$ mmHG und $\mu_2=86.746649$ mmHG der beiden Stichproben verschieden sind, getestet. Der p-Wert des Tests ist 3.4299441×10^{-5} , was bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau verworfen wird, auf dem 1% Signifikanzniveau verworfen wird und auf dem 0.1% Signifikanzniveau verworfen wird.

Der Testoutput beinhaltet auch das 95%-Konfidenzintervall, das durch den Code-Teil `conf.level = 0.95` eingestellt wurde. Dieses reicht von -9.4336692 bis -3.4426045 mmHg.

Beispiel mit 2 abhängigen Stichproben

Im folgenden Beispiel wird ein Medikamententest durchgeführt, bei dem ein blutdrucksenkendes Mittel 100 Testpersonen aus Stichprobe `patienten1` verabreicht wird. Als Vergleichswerte dienen die Messungen bei denselben 100 Patienten vor Verabreichung des Medikaments, zusammengefasst in Stichprobe `patienten2abh`.

Wir treffen die Überlegungen von zuvor:

1. Hier wurden die beiden Stichproben **abhängig** voneinander gezogen, da dieselben Patienten vor und nach Verabreichung des Medikaments verglichen werden.
2. Die Daten sind daher nicht die eigentlichen Messungen vor und nach Verabreichung, sondern die Differenzen dieser Werte bei jedem Patienten $\Delta_i = \text{Blutdruck}_{\text{vorher}} - \text{Blutdruck}_{\text{nachher}}$.
3. Sind die Daten **normalverteilt**?

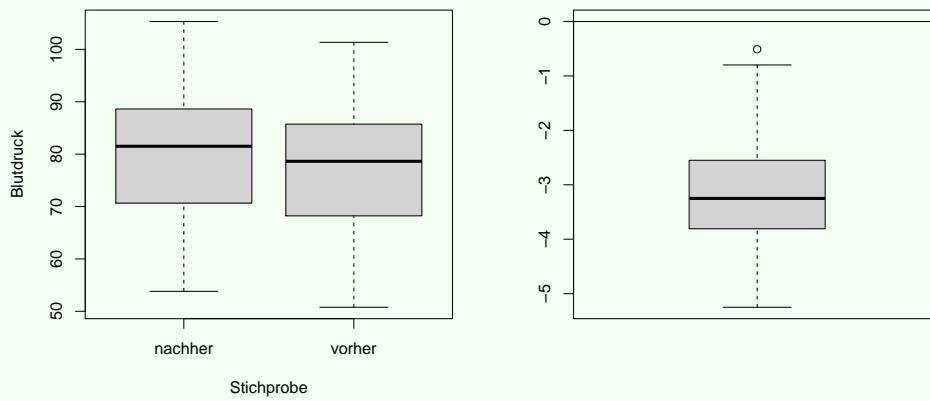
Das müssen wir erst feststellen durch Datenexploration der Stichprobe der Differenzen.

Testen der Differenzen anstatt Originalmesswerte

Es ist hier also wichtig, dass wir die Differenzen der Werte, nicht die Werte selbst auf Normalverteilung überprüfen. Die Werte selbst hier zu vergleichen, als wären sie unabhängig gezogen worden, könnte zu falschen Entscheidungen führen, wie wir durch den visuellen Vergleich der Originaldaten und der Differenzen erkennen.

```
datenabh<-data.frame(Blutdruck=c(patienten2abh, patienten1),
                      Stichprobe=factor(c(rep("vorher",length(patienten2abh)),rep("nachher",length(patienten1)))))

Differenzen<-patienten2abh-patienten1
par(mfrow=c(1,2))
boxplot(Blutdruck~Stichprobe, data=datenabh, main="Originaldaten")
boxplot(Differenzen, ylim=c(min(Differenzen),0)); abline(h=0)
```

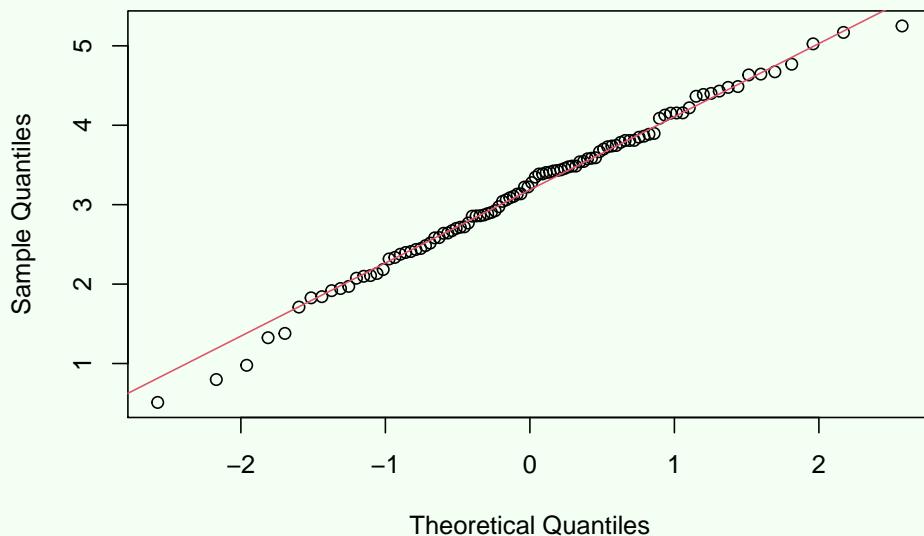


Während die Originaldaten starke Überschneidungen haben und nicht unterschiedlich hinsichtlich ihrer Lage wären, liegen die Differenzen deutlich unterhalb von 0 und der Effekt, den wir hier testen wollen, wird augenscheinlich.

Überprüfung der Normalverteilung der Differenzen

```
qqnorm(patienten1-patienten2abh,main="Überprüfung der Differenzen")
qqline(patienten1-patienten2abh,col=2)
```

Überprüfung der Differenzen



Wir erkennen, dass Stichprobe der Differenzwerte annähernd normalverteilt ist und wir daher sinnvollerweise den t-Test zum Vergleich von 2 abhängigen Stichproben durchführen dürfen. Wir können diese visuelle Exploration auch durch einen Test überprüfen, den Shapiro-Wilks-Test für Normalverteilung:

```
shapiro.test(patienten2abh-patienten1)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: patienten2abh - patienten1  
## W = 0.99104, p-value = 0.7479
```

Dieser ist nicht signifikant, wodurch die Nullhypothese der Normalverteilung nicht verworfen wird. Wir können also, wie auch aus der graphischen Exploration geschlossen, sinnvollerweise den t-Test zum Vergleich von 2 abhängigen Stichproben durchführen.

Durchführung des 2-Stichproben-t-Test für abhängige Stichproben

Wir wählen also den parametrischen student's t-Test mit der Option für gepaarte Stichproben durch `paired=TRUE`.

```
t.test(patienten2abh, patienten1, paired = TRUE, conf.level = 0.99)
##
##  Paired t-test
##
## data: patienten2abh and patienten1
## t = -33.185, df = 99, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -3.429793 -2.926708
## sample estimates:
## mean of the differences
##          -3.17825
```

Die Teststatistik $t=-33.1847464$ hat eine t-Verteilung mit $df=99$ Freiheitsgraden, welche allerdings nur unter Normalverteilung der Differenzen Gültigkeit hat, was wir im Vorfeld überprüft haben.

Es wurde hier die zweiseitige Alternativhypothese, dass die Differenz der Werte im Mittel $\Delta_\mu = -3.1782501$ mmHG beträgt, getestet. Der p-Wert des Tests ist $1.9130681 \times 10^{-55}$, was bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau verworfen wird, auf dem 1% Signifikanzniveau verworfen wird und auf dem 0.1% Signifikanzniveau verworfen wird.

Der Testoutput beinhaltet auch das 99%-Konfidenzintervall, das durch den Code-Teil `conf.level = 0.99` eingestellt wurde. Dieses bezieht sich wiederum auf die Differenz der Messwert und reicht von -3.4297926 bis -2.9267076 mmHg.

Äquivalenz des gepaarten 2-Stichproben-t-Test zum 1-Stichproben-t-Test für die Differenzen

Dieser 2-Stichproben-t-Test für gepaarte Stichproben ist identisch mit dem 1-Stichproben-t-Test für die Differenzen der Messwerte, wenn man die Messungen in der 2. Stichprobe von der 1. subtrahiert.

```
t.test(patienten2abh-patienten1, conf.level = 0.99)
##
##  One Sample t-test
##
## data: patienten2abh - patienten1
## t = -33.185, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -3.429793 -2.926708
## sample estimates:
## mean of x
##          -3.17825
```

Beispiel mit nicht-normalverteilten Daten

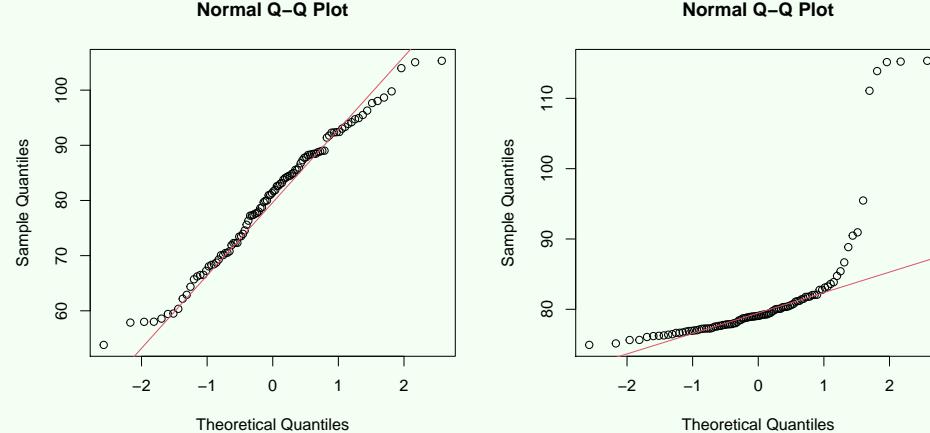
Wie bereits in unserem ersten Beispiel wird im folgenden Beispiel ein Medikamententest durchgeführt, bei dem ein blutdrucksenkendes Mittel 100 Testpersonen aus Stichprobe `patienten1` verabreicht wird. Als Vergleichsgruppe in Stichprobe `patienten3` wird 100 ein Placebo verabreicht.

Wir treffen die Überlegungen von zuvor:

1. Hier wurden die beiden Stichproben **unabhängig** voneinander gezogen, da andere Patienten das Originalpräparat und das Placebo erhalten.
2. Sind die Daten **normalverteilt**?
Das müssen wir erst feststellen durch Datenexploration von 2 unabhängigen Stichproben.

Überprüfung der Normalverteilung

```
par(mfrow=c(1, 2))
qqnorm(patienten1); qqline(patienten1, col=2)
qqnorm(patienten3); qqline(patienten3, col=2)
```



Wir erkennen, dass nicht beide Stichprobe annähernd normalverteilt sind und wir daher den t-Test zum Vergleich von 2 Stichproben nicht durchführen dürfen. Wir können diese visuelle Exploration auch durch einen Test überprüfen, den Shapiro-Wilks-Test für Normalverteilung:

```
shapiro.test(patienten1)

##
##  Shapiro-Wilk normality test
##
## data: patienten1
## W = 0.98367, p-value = 0.2537
shapiro.test(patienten3)

##
##  Shapiro-Wilk normality test
##
## data: patienten3
## W = 0.55358, p-value = 6.128e-16
```

Der erste Test ist nicht signifikant, wodurch die Nullhypothese der Normalverteilung nicht verworfen wird. Der zweite Test hingegen ist hochsignifikant mit einem p-Wert von $6.1279598 \times 10^{-16}$, der deutlich 0.05, 0.01 und 0.001 unterschreitet. Wir können also, wie auch aus der graphischen Exploration geschlossen, den t-Test zum Vergleich von 2 Stichproben nicht durchführen.

Durchführung des Wilcoxon-Rangsummentest

Statt für den t-Test entscheiden wir uns für den **nichtparametrischen Test für Lagevergleich**, den **Wilcoxon Rangsummentest** (Wilcoxon Rank Sum Test).

```
wilcox.test(patienten1, patienten3)
##
##  Wilcoxon rank sum test with continuity correction
##
## data: patienten1 and patienten3
## W = 5190, p-value = 0.6433
## alternative hypothesis: true location shift is not equal to 0
```

Die Teststatistik ist $W=5190$ und hat keine uns standardmäßig geläufige geschlossene Verteilung. Es wurde hier die zweiseitige Alterntivhypothese, dass die Lage der beiden Stichproben verschieden ist, getestet. Der p-Wert des Tests ist 0.6433485, was bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau beibehalten wird, auf dem 1% Signifikanzniveau beibehalten wird und auf dem 0.1% Signifikanzniveau beibehalten wird.

Wir sehen nun, was wir an Informationen verlieren, wenn wir den nichtparametrischen Test benutzen. Abgesehen von der geringeren Treue können wir hier weder einen Schätzer für die mittlere Lage der beiden Stichproben erhalten, noch bekommen wir ein Konfidenzintervall.

Beispiel mit 2 abhängigen Stichproben nichtnormalverteilter Daten

Im folgenden Beispiel wird ein Medikamententest durchgeführt, bei dem ein blutdrucksenkendes Mittel 100 Testpersonen aus Stichprobe **patienten1** verabreicht wird. Als Vergleichswerte dienen die Messungen bei denselben 100 Patienten vor Verabreichung des Medikaments, zusammengefasst in Stichprobe **patienten3**.

Wir treffen die Überlegungen von zuvor:

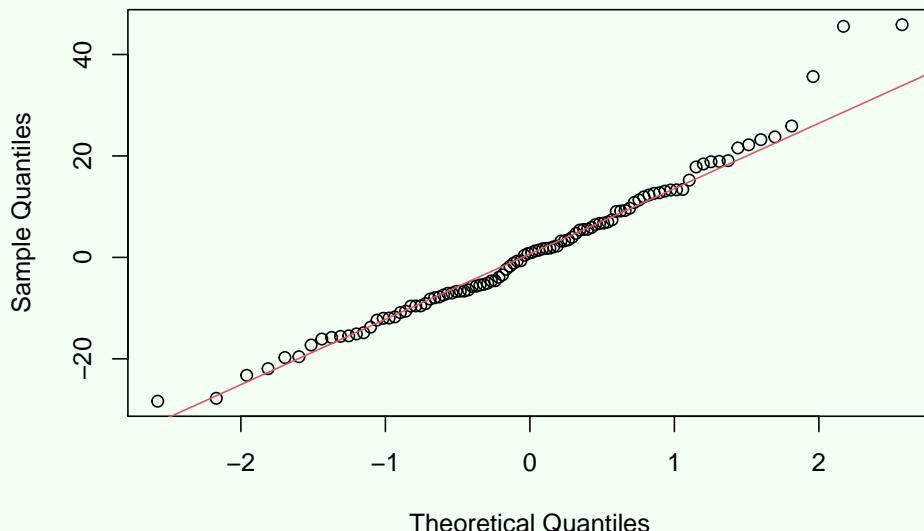
1. Hier wurden die beiden Stichproben **abhängig** voneinander gezogen, da dieselben Patienten vor und nach Verabreichung des Medikaments verglichen werden.
2. Die Daten sind daher nicht die eigentlichen Messungen vor und nach Verabreichung, sondern die Differenzen dieser Werte bei jedem Patienten $\Delta_i = \text{Blutdruck}_{\text{vorher}} - \text{Blutdruck}_{\text{nachher}}$.
3. Sind die Daten **normalverteilt**?
Das müssen wir erst feststellen durch Datenexploration der Stichprobe der Differenzen.

Überprüfung der Normalverteilung der Differenzen

Es ist hier also wichtig, dass wir die Differenzen der Werte, nicht die Werte selbst auf Normalverteilung überprüfen.

```
qqnorm(patienten3-patienten1,main="Überprüfung der Differenzen")
qqline(patienten3-patienten1,col=2)
```

Überprüfung der Differenzen



Wir erkennen, dass Stichprobe der Differenzwerte nicht annähernd normalverteilt ist und wir daher den t-Test zum Vergleich von 2 abhängigen Stichproben nicht durchführen dürfen. Wir können diese visuelle Exploration auch durch einen Test überprüfen, den Shapiro-Wilks-Test für Normalverteilung:

```
shapiro.test(patienten3-patienten1)
```

```
##
##  Shapiro-Wilk normality test
##
## data: patienten3 - patienten1
## W = 0.97371, p-value = 0.04281
```

Dieser ist signifikant unterhalb von 0.05, wodurch die Nullhypothese der Normalverteilung verworfen wird. Wir können also, wie auch aus der graphischen Exploration geschlossen, sinnvollerweise den t-Test zum Vergleich von 2 abhängigen Stichproben nicht durchführen.

Durchführung des Wilcoxon-Rangsummen-Test für gepaarte Stichproben

Wir entscheiden uns daher für den **nichtparametrischen Wilcoxon Rangsummentest mit der Option für gepaarte Stichproben paired=TRUE**.

```
wilcox.test(patienten1,patienten3,paired = TRUE)
##
##  Wilcoxon signed rank test with continuity correction
##
## data: patienten1 and patienten3
## V = 2428, p-value = 0.74
## alternative hypothesis: true location shift is not equal to 0
```

Schätzung und Test von Varianzen

1 Stichproben Varianz-Test

parametrischer Test

Daten Die angenommene Verteilung der Daten ist die **Normalverteilung** $N(\mu, \sigma)$

Ziel Testen des **Parameters** $\sigma =$ der **Standardabweichung** von n Messungen

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

Die Teststatistik lautet

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Die angenommene Verteilung der Teststatistik ist die **χ^2 -Verteilung** mit $df = n-1$ Freiheitsgraden.

Teststatistik

Nullhypothese Unter der Nullhypothese wird bei nur 1 Stichprobe als Referenzwert der Standardabweichung σ_0 angenommen

zweiseitig $H_0 : \sigma = \sigma_0$

linksseitig $H_0 : \sigma \geq \sigma_0$

rechtsseitig $H_0 : \sigma \leq \sigma_0$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert

zweiseitig $H_A : \sigma \neq \sigma_0$

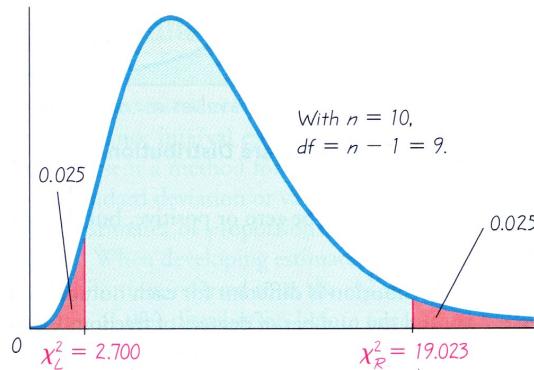
linksseitig $H_A : \sigma < \sigma_0$

rechtseitig $H_A : \sigma > \sigma_0$

Signifikanzniveau Gängig ist die Konfidenz $1-\alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1-\alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

Wir verwenden bei der Schätzung der Prädiktionsintervalle, Konfidenzintervalle, Akzeptanz- und Verwerfungsbereiche sowie für die Ermittlung des p-Werts jeweils die χ^2 -Verteilung. Das ist eine Anwendung der mathematischen Statistik, wonach die Verteilung der Varianz von normalverteilten Daten eine χ^2 -Verteilung ist, wenn entsprechend viele Stichproben gezogen werden.

Finding Critical Values of χ^2 A simple random sample is obtained. Construction of a confidence interval for the population variance σ^2 requires the left and right critical values of χ^2 corresponding to a confidence level of 95% and a sample size of $n = 10$. Thus, we are looking for a critical value χ_L^2 separating an area of 0.025 in the left tail, and a critical value χ_R^2 separating an area of 0.025 in the right tail.



- Objective: Construct a confidence interval used to estimate a population standard deviation or variance.
- Requirements:
 1. The sample is a simple random sample.
 2. The population must have normally distributed values (even if the sample is large).
- Confidence interval for σ^2 :

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

- Confidence interval for σ :

$$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}}$$

Example: Finding a CI for σ

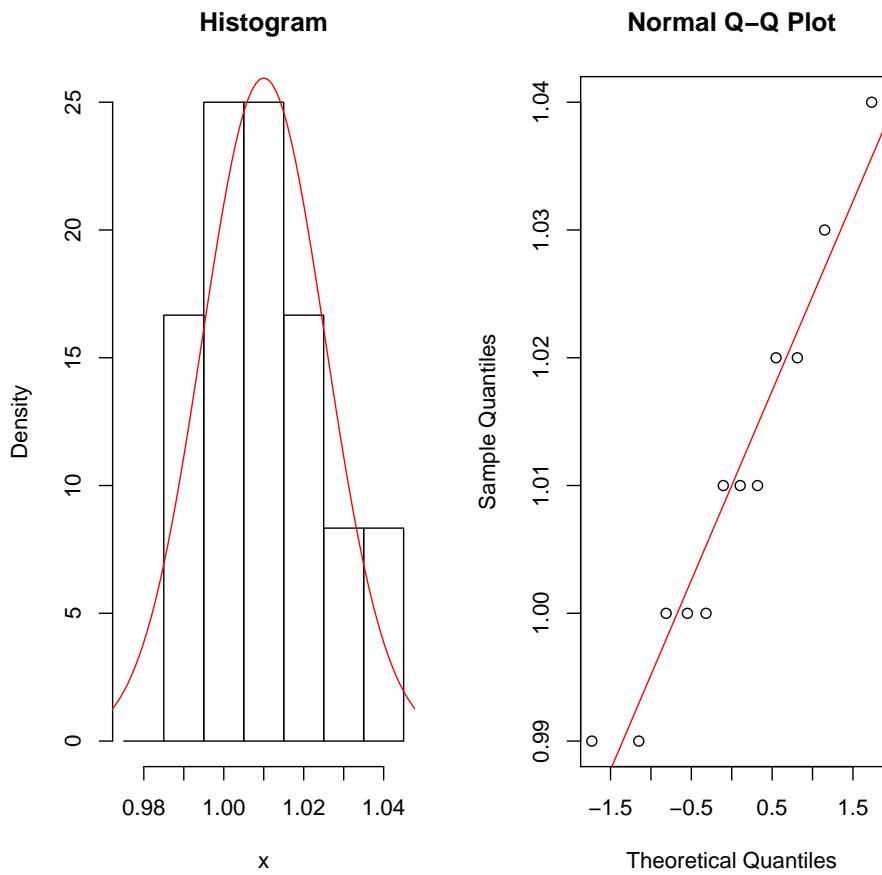
Confidence Interval for Bottle Fillings

Market conditions require fillings of bottles that do not vary much. Listed below are twelve fillings (in liters). Use the sample data to construct a 95% confidence interval estimate of the standard deviation of all fillings.

0.99 1.01 1.00 1.01 1.03 1.01 1.02 0.99 1.00 1.02 1.00 1.04

Requirement check:

1. Simple random sample \Rightarrow OK
2. Normality?



Normality OK!

- $s = 0.015374$
- $n = 12, df = 11 \Rightarrow \chi^2_L = 3.816, \chi^2_R = 21.920$
-

$$\sqrt{\frac{(n-1)s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$$

$$0.011 < \sigma < 0.026$$

Based on this result, we have 95% confidence that the limits of 0.011 liter and 0.026 liter contain the true value of σ . The confidence interval can also be expressed as $(0.011, 0.026)$ or $[0.011, 0.026]$, but the format of $s \pm E$ cannot be used because the confidence interval does not have s at its center.

F-Test = 2 Stichproben Varianz-Test

parametrischer Test

Daten Die angenommene Verteilung der Daten ist die **Normalverteilung** $N(\mu, \sigma)$

Ziel Testen des **Parameters** σ_1^2 = der **Varianz** von n_1 Messungen von Stichprobe 1 und σ_2^2 = der **Varianz** von n_2 Messungen von Stichprobe 2

Schätzwert Stichprobenstandardabweichung $\hat{\sigma}_i = \frac{1}{n_i-1} \sqrt{\sum_{k=1}^n (x_k - \bar{x}_i)^2}$ mit Stichprobe i=1,2

Die Teststatistik lautet

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

Die angenommene Verteilung der Teststatistik ist die **F-Verteilung** mit $df1 = n_1 - 1$ und $df2 = n_2 - 1$ Freiheitsgraden.

Teststatistik

Nullhypothese Unter der Nullhypothese wird folgender Vergleich zwischen den Stichproben angestellt:

zweiseitig $H_0 : \sigma_1 = \sigma_2$

linksseitig $H_0 : \sigma_1 \geq \sigma_2$

rechtsseitig $H_0 : \sigma_1 \leq \sigma_2$

Alternativhypothese Jeweils als Gegenteil der Nullhypothese formuliert

zweiseitig $H_A : \sigma_1 \neq \sigma_2$

linksseitig $H_A : \sigma_1 < \sigma_2$

rechtseitig $H_A : \sigma_1 > \sigma_2$

Signifikanzniveau Gängig ist die Konfidenz $1 - \alpha = 0.95$, also die Signifikanz $\alpha = 0.05$ zu setzen, genauso wie die Konfidenz $1 - \alpha = 0.99$, also die Signifikanz $\alpha = 0.01$ zu setzen.

```
var.test(x = daten, y=daten2, conf.level = 0.95,  
         alternative = c("two.sided", "less", "greater"))
```

Dieser Test besitzt zwei verbreitete Anwendungen:

1. Als Hilfsmittel, um vor dem 2-Stichproben-t-Test festzustellen, ob die Stichproben die gleiche Varainz haben.
2. Als Varianzanalyse (ANOVA) zum Vergleich des Mittelwerts mehrerer Stichproben. Die eigentliche ANOVA folgt aber bei den linearen Modellen.

Ergänzende Inhalte zu Hypothesentesten

Critics of classical hypothesis testing “The problem is simple, the researchers are disproving always false null hypotheses and taking this disproof as near proof that their theory is correct.”¹

‘Paradoxically, because the studies are usually underpowered to disprove this (easiest possible thing to disprove) hypotheses theyp-hack’ without even realizing it is a problem. Committees and reviewers demand p-values without understanding what they are.’² }

Multiple Testing Correction

¹(anonymous post from (<http://andrewgelman.com/>))[<http://andrewgelman.com/>])

²(<http://andrewgelman.com/2014/03/27/beyond-valley-trolls/#comment-156007>) [<http://andrewgelman.com/2014/03/27/beyond-valley-trolls/#comment-156007>]

Inference about two means: Independent samples

Independent and Dependent Samples

- Independent samples: The sample values from one population are not related to or somehow naturally paired or matched with the sample values from the other population. Example: Average number of words spoken per day, measured amongst 123 men and 134 women.
- Dependent samples: The sample values are *paired*, i.e. from the same subject (before/after) or from matched pairs (such as husband/wife). Example: “Freshman 15”.

Remark

We cover the case where σ_1 and σ_2 are unknown and not assumed to be equal, which is the most common case. Nevertheless, these assumptions can easily be modified (c.f. ~Triola 9-3).

Before conducting a hypothesis test, consider the context of the data, the source of the data, the sampling method, and explore the data with graphs and descriptive statistics. Be sure to verify that the requirements are satisfied.

Requirements

1. σ_1 and σ_2 are **unknown** and not necessarily equal.
2. The two samples are **independent**.
3. Both samples are **simple random samples**.
4. The two sample sizes are both **large** ($n_1 > 30$ and $n_2 > 30$) or both samples come from populations with normal distributions.

Example: Inference about two means: Independent samples

“Are Women Really More Talkative Than Men?” by Mehl et al., *Science*, Vol. 317, No. 5834

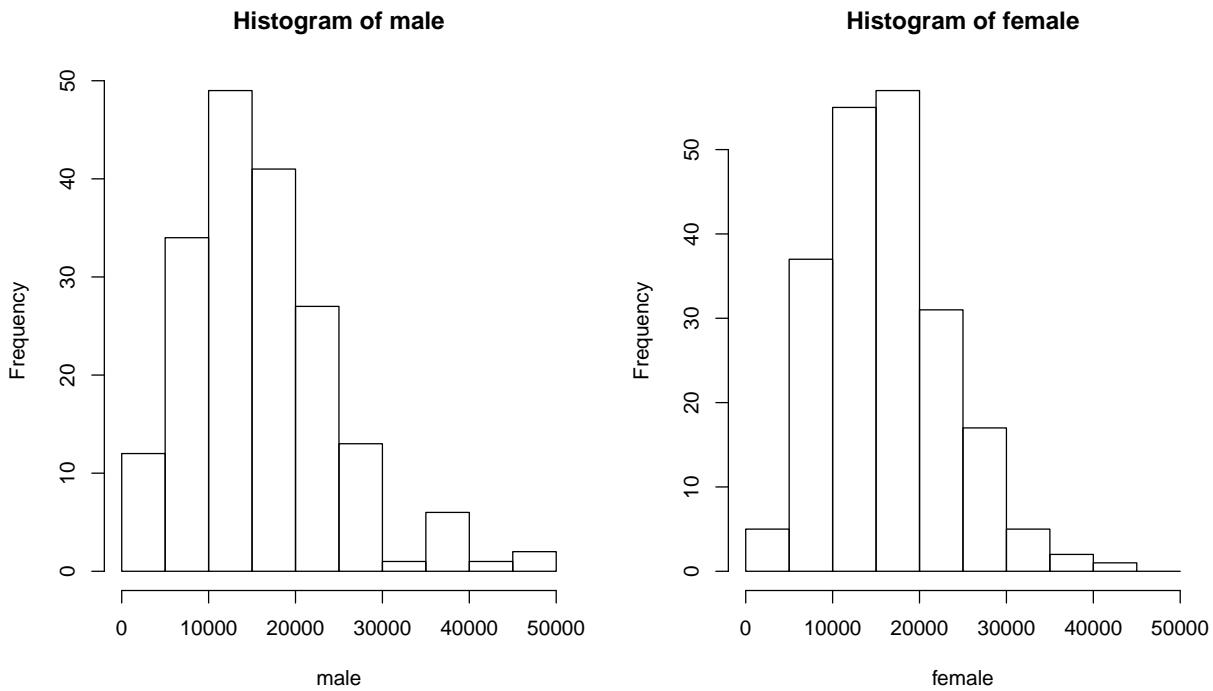
Number of Words Spoken in a Day	
Men	Women
$n_1 = 186.0$	$n_2 = 210.0$
$\bar{x}_1 = 15,668.5$	$\bar{x}_2 = 16,215.0$
$s_1 = 8632.5$	$s_2 = 7301.2$

Use a 0.05 significance level to test the claim that men and women speak the same mean number of words in a day.

Requirement check:

1. Population standard deviations are not known \Rightarrow OK!
2. The samples are independent \Rightarrow OK!
3. We assume that we have simple random samples \Rightarrow OK!
4. Both samples are large \Rightarrow OK!

Even though samples are large, let's look at our data:



Inference about two means: Independent samples

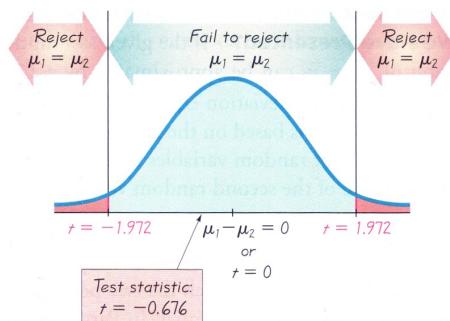
Hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Calculations:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \overbrace{(\mu_1 - \mu_2)}^0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -0.676, \quad df = \min(n_1, n_2) - 1 = 185$$

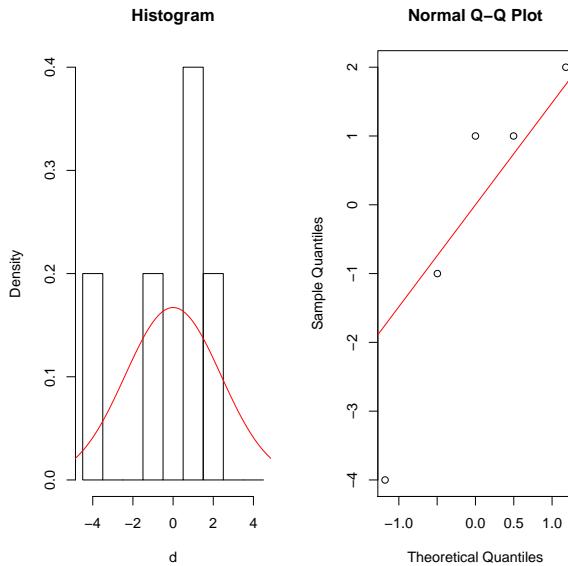


Beispiel zu sehr kleinen abhängigen Stichproben

Baby Data Set: “Freshman 15”

April weight (<i>after</i>)	66	52	68	69	71
September weight (<i>before</i>)	67	53	64	71	70
Difference d (<i>gain</i>)	-1	-1	4	-2	1

1. Dependent samples \Rightarrow OK!
2. Voluntary response, not simple random sample \Rightarrow FAIL! Let's continue anyway but be careful with interpreting the results
3. Sample size is small \Rightarrow Normality?



Hypotheses:

$$\begin{aligned} H_0 : \mu_d &= 0 \\ H_1 : \mu_d &\neq 0 \end{aligned}$$

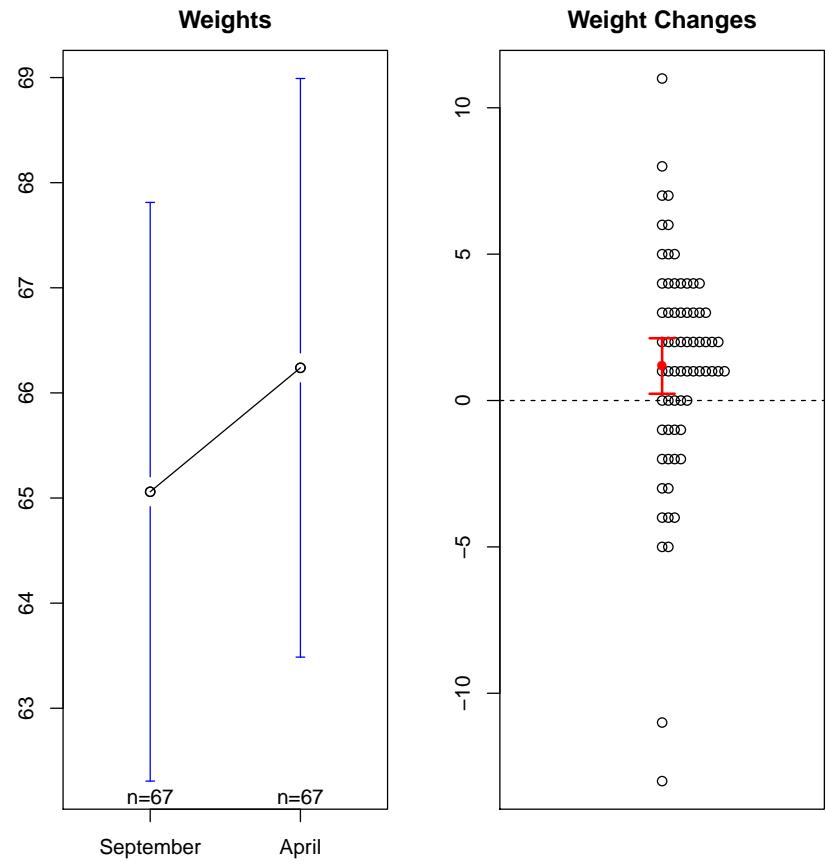
Calculations:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{0.2}{\frac{\sqrt{5.7}}{\sqrt{5}}} \approx 0.187, \quad df = n - 1 = 4.$$

Confidence interval:

$$E = t_{\alpha/2} \frac{s_d}{\sqrt{n}} = 2.776 \times \frac{\sqrt{5.7}}{\sqrt{5}} \approx 2.964 \Rightarrow -2.764 < \mu_d < 3.164$$

Not sufficient evidence to warrant rejection of the claim that the mean change in weight from September to April is equal to 0kg. Based on our sample, there does not appear to be a significant weight gain. Limitations: (1) only Rutgers students, (2) potential self-selection bias!



Abhangigkeit von 2 oder mehr Variablen

Um die statistische Abhangigkeit zweier (eindimensionaler) Merkmale X und Y festzustellen, nimmt man Daten, die aus simultanen Messungen (x_i, y_i) , $i = 1, \dots, n$, resultieren. Das bedeutet, dass unsere Beobachtungen im Gegensatz zu bisher betrachteten Szenarien zwei-dimensional sind. Anstatt nach Unterschieden zwischen 2 Stichproben zu suchen, interessiert uns hier der Zusammenhang zwischen den beiden Dimensionen, den Variablen X und Y.

Zusammenhange zwischen 2 kategorialen Variablen

Zusammenfassung und Visualisierung

Kontingenztafeln In unserer Messreihe von n Messungen nennen wir analog zum eindimensionalen Fall

$n_{i,j}$ die **absolute Häufigkeit** der Kombination (a_i, b_j) , und
 $h_{i,j} = \frac{n_{i,j}}{n}$ die **relative Häufigkeit** der Kombination (a_i, b_j) .

Diese relativen und absoluten Häufigkeiten basieren auf der zugrundeliegenden Tabellenzusammenfassung der Häufigkeiten der jeweiligen Kombinationen der Kategorien, welche **Kontingenztafel** heißt.

$Y \setminus X$	a_1	\dots	a_J	Zeilenhäufigkeiten
b_1	$n_{1,1}$	\dots	$n_{1,J}$	$n_{1,..}$
\vdots	\vdots	\ddots	\vdots	\vdots
b_I	$n_{I,1}$	\dots	$n_{I,J}$	$n_{I,..}$
Spaltenhäufigkeiten	$n_{.,1}$	\dots	$n_{.,J}$	n

Hier stehen in den Randsummen, also **absoluten Spaltenhäufigkeiten** $n_{.,j} = \sum_{i=1}^I n_{i,j}$ und **Zeilenhäufigkeiten** $n_{i,..} = \sum_{j=1}^J n_{i,j}$,

	I	II	Zeilenhäufigkeit
A	80.00	60.00	140.00
B	40.00	30.00	70.00
C	40.00	120.00	160.00
Spaltenhäufigkeit	160.00	210.00	370.00

oder die Randhäufigkeiten **relativen Spaltenhäufigkeiten** $h_{.,j} = \frac{n_{.,j}}{n}$ und **Zeilenhäufigkeiten** $h_{i,..} = \frac{n_{i,..}}{n}$, welche durch die Gesamtanzahl aller Beobachtungen n dividiert werden.

	I	II	Zeilenhäufigkeit
A	0.22	0.16	0.38
B	0.11	0.08	0.19
C	0.11	0.32	0.43
Spaltenhäufigkeit	0.43	0.57	1.00

Bei den relativen Häufigkeiten ist zu beachten, ob dass man auch relative Häufigkeiten nicht nur bezüglich der Gesamtanzahl ermitteln kann, sondern auch zeilenweise bzw. Spaltenweise auf 100% gegengerechnet, um innerhalb der Zeilenprozente und Spaltenprozente interpretieren zu können.

	I	II	Zeilenprozent
A	0.57	0.43	1.00
B	0.57	0.43	1.00
C	0.25	0.75	1.00
Anzahl	160.00	210.00	370.00

	I	II	Anzahl
A	0.50	0.29	140.00
B	0.25	0.14	70.00
C	0.25	0.57	160.00
Spaltenprozent	1.00	1.00	370.00

Kontingenztafeln stellen die grundsätzliche Zusammenfassung von Informationen simultan gemessener Ereignis und bedingter Ereignisse dar.

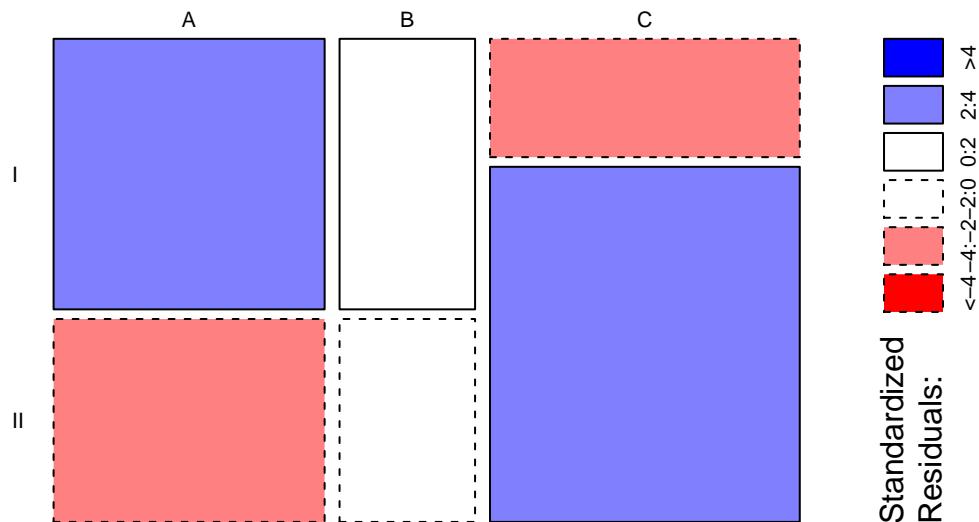
Visualisierungen

Visualisierungen von Kontingenztafeln sind großteils durch 2 Dimensionen limitiert und können daher nur beschränkt die zugrundeliegenden Szenarien graphisch umsetzen.

Mosaikplots

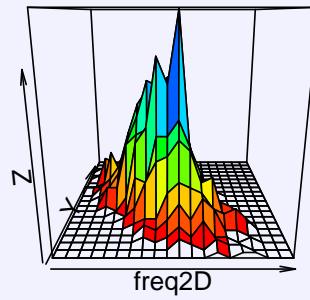
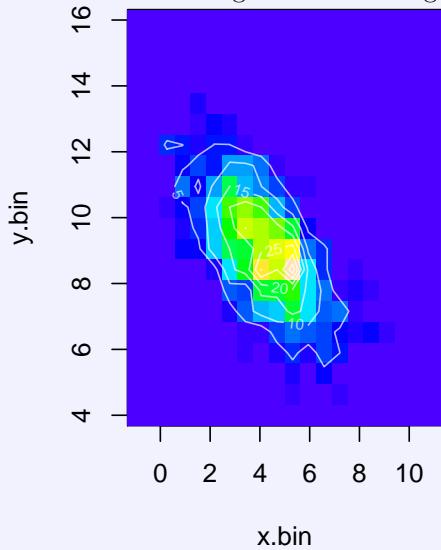
Mosaikplots sind die direkte graphische Umsetzung von Kontingenztafeln und den absoluten bzw. relativen Häufigkeiten, sowie den Randverteilungen. Die Längen und Breiten der Mosaikteile sind durch die Randhäufigkeiten definiert, die tatsächlichen absoluten bzw. relativen Häufigkeiten jeder Kombination von Kategorien wird dann durch die Fläche des Mosaikteils wiedergegeben.

Mosaikplot



2-dimensionale Histogramme

Analog zu 1-dimensionalen Histogrammen, geben **2-dimensionale Histogramme** die Häufigkeitsverteilung in 2 Dimensionen bei angeordneten Kategorien wieder.



Simpson's Paradox

Example: Skin color and death sentence

The following example may appear not politically correct at first sight, but it is a famous example for misinterpretations of tabulated data and meant as a warning regarding first impressions and misinterpretations.

The data used for this example come from the *New York Times Magazine*, March 11, 1979 which is the origin of some old-fashioned expressions used. Originally, they were published concerning the frequency of death sentences in Florida. They led to nationwide discussions and questioning of statistics. In the light of recent “Black Lives Matter” campaigns data analyses such as this one have gained a new relevance.

Watch out not to repeat such mistakes in your own research!

Case no.	Skin color of accused	Death sentence
1	b	0
2	b	0
3	w	0
4	b	1
:	:	:
4764	w	0

The data contain 4764 observations of two nominal categorical variables, skin colour of the accused and death sentence.

The summary as a contingency table with absolute frequencies looks like this:

	black skin	white skin	Σ
Death sentence	59	72	131
No death sentence	2448	2185	4633
Σ	2507	2257	4764
Proportion in %	2.4	3.3	2.8

This table was published and produced an uproar amongst those looking for racial discrimination. According to overall summaries more caucasian accused had been sentenced to death than African Americans. To learn why one should tread carefully when dealing with conditional probabilities will we extend the example by additional information.

Let us now also look at the skin-color of the *victim* and construct a three way cross tabulation:

Skin-color of victim		black		white	
Skin-color of accused		b	w	b	w
Death sentence		11	0	48	72
No death sentence		2209	111	239	2074
Sum		2220	111	287	2146
Proportion in %		0.5	0.0	20.1	3.5

This table changes the whole picture the previous table showed us. An African American accused of murder of a white person was sentenced to death more than 20 % of the time, not a single white accused of murder of an African American victim was sentenced to death. However, as most victims and accused share the same skin colour these extreme effects are covered by the majority of sentences. This effect is called **Simpson's paradox** which happens when subgroups are extremely unbalanced and show opposing effects to the majority.

Inferenz durch χ^2 -Test für Homogenität und Unabhängigkeit

Das Testen von Daten, welche in 2 Dimensionen 2 kategoriale Variablen enthalten, um Zusammenhänge zwischen diesen Variablen zu erkennen, erfolgt mithilfe eines χ^2 -Tests. Grundsätzlich können auch weitere Szenarien wie das Testen von Zusammenhängen zwischen 3 kategorialen Variablen mithilfe des Mantel-Haenszel-Tests oder zwischen einer abhängigen bivariaten oder multivariaten Variablen und beliebigen erklärenden Variablen mithilfe von logistischer Regression durchgeführt werden, aber diese sind nicht Teil dieser Lehrveranstaltung.

χ^2 -Test In der Kontingenztafel stehen die Wahrscheinlichkeiten für das Eintreten zweier Ereignisse gleichzeitig, also $\mathbb{P}[A \wedge B]$ in den jeweiligen Zellen der Tabelle, während an den Rändern die Randhäufigkeiten $\mathbb{P}[A]$ und $\mathbb{P}[B]$ zu finden sind. Ob Einzelereignisse also unabhängig sind oder nicht, erkennt man durch Bilden des Produkts der Randhäufigkeiten und Vergleich mit der Häufigkeit des Eintretens beider Ereignisse.

Im Rahmen der bedingten Wahrscheinlichkeiten haben wir das Prinzip besprochen, dass Ereignissen unabhängig sind, wenn ihre Eintrittswahrscheinlichkeiten miteinander multipliziert ein Eintrittswahrscheinlichkeit des Geschehens beider Ereignisse zum selben Zeitpunkt ist. Sind diese Werte ident (bei realen Beobachtungen annähernd gleich), spricht man von unabhängigen Ereignissen, anderenfalls von abhängigen Ereignissen. Dieses Prinzip ist die Basis des χ^2 -Tests, der genau diese Unabhängigkeit von in Kontingenztafeln erfassten Ereignissen überprüft.

Die Inferenz in Bezug auf Kontingenztafeln hat als zugrundeliegende Frage entweder

- Ist die Verteilung unabhängig von Teilkategorien?"(Unabhängigkeitstest) oder
- Ist die Verteilung dieselbe in allen Teilkategorien?"(Homogenitätstest).

Grundsätzlich sind beide Fragestellungen nur unterschiedliche Betrachtungsweisen desselben mathematischen Umstands, weshalb für beide derselbe Test durchgeführt wird, wobei sich lediglich die verbale Formulierung der Nullhypothese und Alternativhypothese unterscheidet.

Die getestete **Nullhypothese** lautet

H_0 : Die Verteilung ist unabhängig von der Teilkategorie.

und die **Alternativhypothese**

H_A : Die Verteilung ist abhängig von der Teilkategorie, also unterschiedlich zwischen den einzelnen Teilkategorien.

Würde die Nullhypothese also die Unabhängigkeit der Merkmale gelten, wären also ausschließlich die Randwahrscheinlichkeiten für die Berechnung der gemeinsamen Wahrscheinlichkeit des Auftretens zweier Ausprägungen relevant. Mathematisch kann man dadurch die erwartete Häufigkeit des Auftretens beider Merkmale a_i und b_j ermitteln, nämlich

$$\mathbb{P}[X = a_j \text{ UND } Y = b_i] = \mathbb{P}[X = a_j] \cdot \mathbb{P}[Y = b_i]$$

mit

$Y \setminus X$	a_1	\dots	a_J	Zeilenhäufigkeiten
b_1	$\mathbb{P}[X = a_1] \cdot \mathbb{P}[Y = b_1]$	\dots	$\mathbb{P}[X = a_J] \cdot \mathbb{P}[Y = b_1]$	$\mathbb{P}[Y = b_1]$
\vdots	\vdots	\ddots	\vdots	\vdots
b_I	$\mathbb{P}[X = a_1] \cdot \mathbb{P}[Y = b_I]$	\dots	$\mathbb{P}[X = a_J] \cdot \mathbb{P}[Y = b_I]$	$\mathbb{P}[Y = b_I]$
Spaltenhäufigkeiten	$\mathbb{P}[X = a_1]$	\dots	$\mathbb{P}[X = a_J]$	n

Die Teststatistik χ^2 misst die Unterschiede der Einträge der erwarteten Häufigkeiten gemäß der obigen Tabelle und der tatsächlich beobachteten Einträge der Kontingenztafel der Daten.

$$\chi^2 = n \times \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- o_{ij} ... beobachtete relative Häufigkeiten (**observed relative frequencies**)
- O_{ij} ... beobachtete absolute Häufigkeiten (**observed (absolute) frequencies**)
- e_{ij} ... erwartete relative Häufigkeiten unter Unabhängigkeit
(**expected relative frequencies (given independence)**)
- E_{ij} ... erwartete absolute Häufigkeiten unter Unabhängigkeit
(**expected (absolute) frequencies (given independence)**)
- n ... sample size

Diese χ^2 Statistik wird mit dem entsprechenden Quantil ihrer zugrundeliegenden Verteilung verglichen bzw. ermittelt man den p-Wert aus dieser zugrundeliegenden Verteilung: der χ^2 -Verteilung mit k Freiheitsgraden, wobei $k = (\text{Zeilenanzahl} - 1) \cdot (\text{Spaltenanzahl} - 1)$ ist.

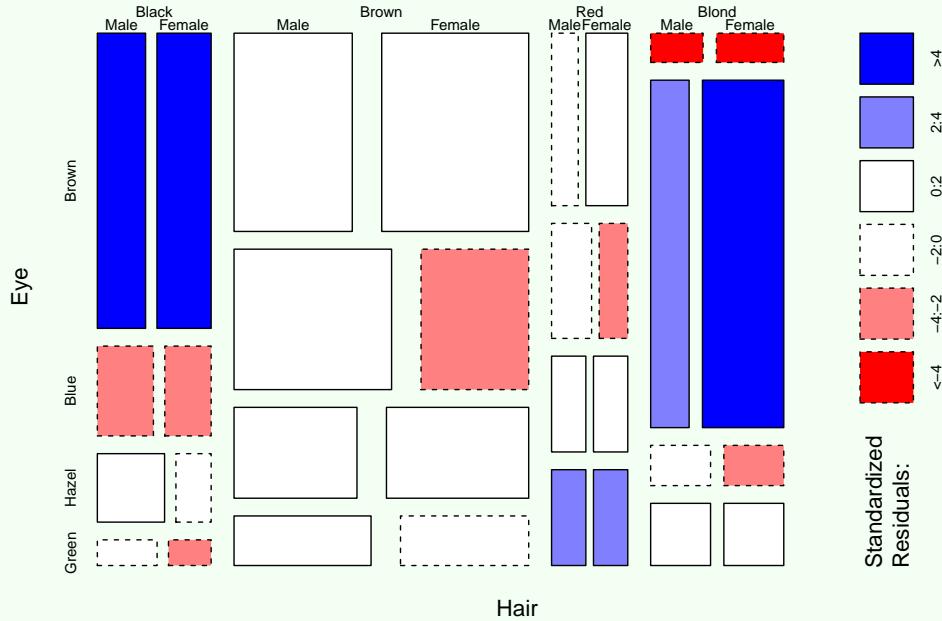
Es gilt grundsätzlich, dass diese χ^2 -Teststatistik stets ≥ 0 ist, da quadrierte Zahlenwerte kleinsteinstens 0 sein können. Exakt 0 ist sie genau dann, wenn alle beobachteten und erwarteten Tabelleneintragungen übereinstimmen.

Beispiel: Es soll ermittelt werden, ob die Verteilung von Haarfarben und Augenfarben bei Männern und Frauen unterschiedlich ist. Im Rahmen einer Studie in den USA unter 1216 kaukasischen Personen hat ergeben, dass diese Verteilung folgendermaßen aussieht:

```
##          Sex Male Female
## Hair Eye
## Black Brown      65     73
##      Blue       23     19
##      Hazel      21     11
##      Green       7      5
## Brown Brown     107    133
##      Blue      101     69
##      Hazel      51     59
##      Green      31     29
## Red  Brown      21     33
##      Blue      21     15
##      Hazel      15     15
##      Green      15     15
## Blond Brown      7      9
##      Blue      61    129
##      Hazel     11     11
##      Green      17     17
```

Die graphische Zusammenfassung dieser Daten mithilfe eines Mosakiplots sieht so aus:

Haar- und Augenfarben nach Geschlecht

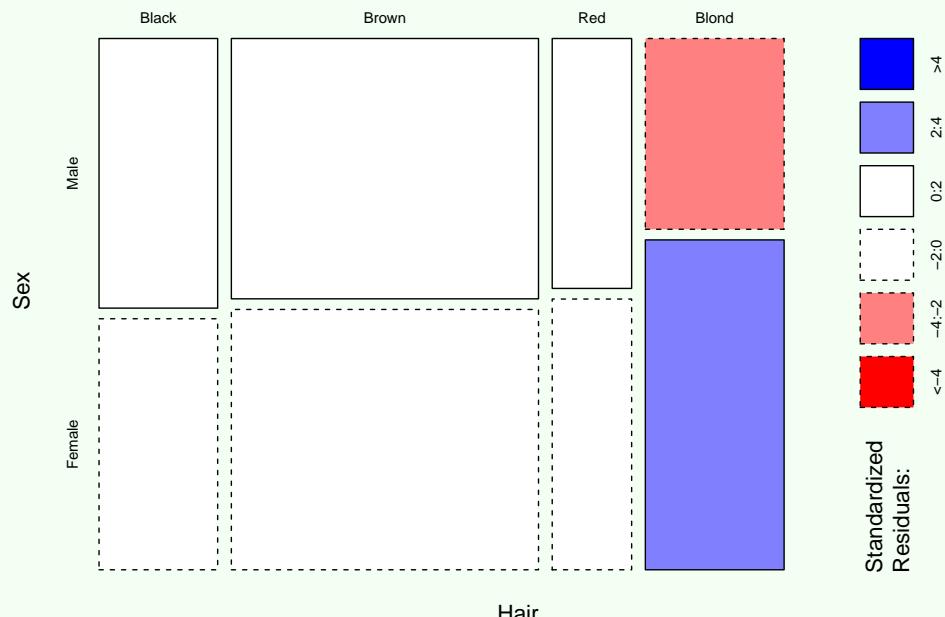


Durch die Farbschattierung eines zugrunde liegenden logistischen Modells erkennen wir, in welchen Bereichen "etwas abweicht". Es gibt etwa unter schwarzaarigen Personen, Männern und Frauen, überproportional viele mit braunen Augen, was uns die dunkelblaue Farbe anzeigen. Bei blonden Männern und Frauen hingegen gibt es weitaus weniger braunäugige Personen als dies bei einer gleichen Verteilung erwartet werden könnte.

ACHTUNG: Wir können mit dem χ^2 -Test nicht gleichzeitig Haar- und Augenfarbe berücksichtigen, sondern jeweils nur eines dieser Kriterien aufgeteilt nach Geschlecht.

	Male	Female
Black	116	108
Brown	290	290
Red	72	78
Blond	96	166

Haarfarben nach Geschlecht



Hier lässt sich erkennen, dass alle Haarfarben bei beiden Geschlechtern gleichermaßen verteilt scheinen, nur blond scheint bei Frauen häufiger als bei Männern vorzukommen. Diese Hypothese, dass die Verteilung der Haarfarben in mindestens einer Teilkategorie unterschiedlich verteilt ist, wollen wir durch den χ^2 -Test überprüfen.

```
chisq.test(Haarfarben)
```

```
## 
## Pearson's Chi-squared test
## 
## data: Haarfarben
## X-squared = 15.474, df = 3, p-value = 0.001453
```

Wir sehen hier deutlich, dass die Teststatistik $\chi^2 = 15.4737618$ beträgt.

Auch die Freiheitsgrade und ihre Berechnung kann man hier gut erkennen:

$df = (\text{Zeilenzahl} - 1) \cdot (\text{Spaltenanzahl} - 1) = (4 - 1) \cdot (2 - 1) = 3$.

Der p-Wert dieses Tests beträgt 0.0014534, was bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau verworfen wird, auf dem 1% Signifikanzniveau verworfen wird und auf dem 0.1% Signifikanzniveau beibehalten wird.

Zusammenhang zwischen einer metrischen Variable und kategorialen Variablen: Varianzanalyse (ANOVA = ANalysis Of VAriance)

Einfache Varianzanalyse (ANOVA)

ANOVA als Hypothesentest

Diese einfache Varianzanalyse wird auch als Test verwendet, um folgende Hypothesen zu testen:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots$$

H_1 : Wenigstens in einer Kategorie ist der Mittelwert unterschiedlich

Das ist die allgemeine Methode, die Unterschiede der mittleren Werte von bekannten Kategorien zu ermitteln.

Die Voraussetzung ist, dass die **Daten annähernd symmetrisch, unimodal und ohne Ausreißer** sind.

ANOVA als Modell für metrische und kategoriale Variablen

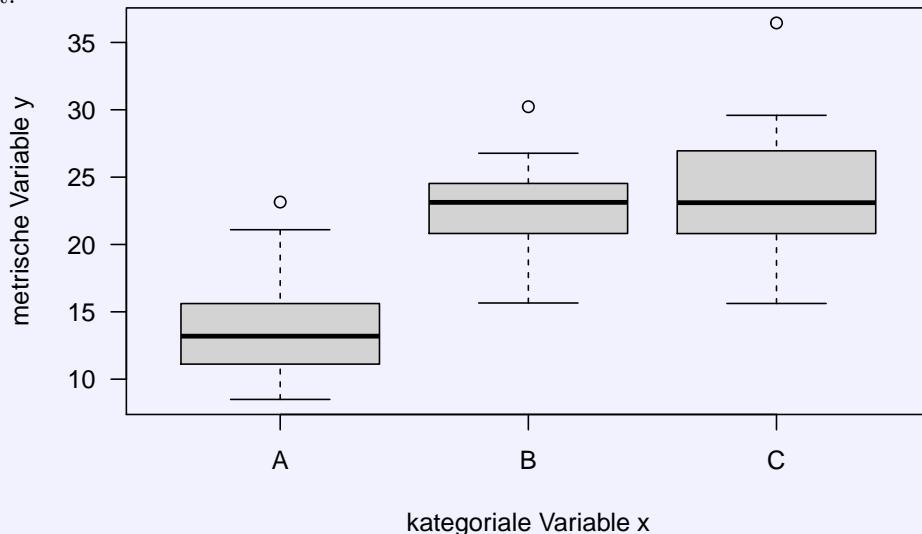
Wir gehen unseren ersten Schritt in Richtung der linearen Regressionsmodelle

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

mit dem einfachen Modell der Varianzanalyse.

Die Messungen y_i , die wir erklären wollen, sind metrische Zahlenwerte. Als Erklärungsvariablen werden bei der ANOVA aber immer nur qualitative Variablen x verwendet, was bedeutet, dass die Kategorien der Variablen x Einfluss auf den Kategoriemittelwert der Variable y haben.

In der explorativen Datenanalyse haben wir dieses Problem auch schon durch parallele Boxplots visualisiert.



Ansatz der ANOVA als lineares Modell

Das Modell wird also einfacher, nämlich im einfachsten Fall mit einer erklärenden Variable mit Kategorien $j = 1, \dots, K$

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

wobei wir hier die verschiedenen Kategorieausprägungen mit dem Index j versehen. α_j misst also die mittlere Abweichung des Mittelwertes der Kategorie j vom Mittelwert aller Daten y .

Das Modell für die einfache Varianzanalyse mit einer erklärenden Variable mit Kategorien $j = 1, \dots, K$

$$Y_{ij} = \underbrace{\mu}_{\text{Gesamtmittelwert}} + \underbrace{\alpha_i}_{\text{Abstand der Gruppenmittelwerte von } \mu} + \underbrace{\epsilon_{ij}}_{\text{Fehler=Residuen}}$$

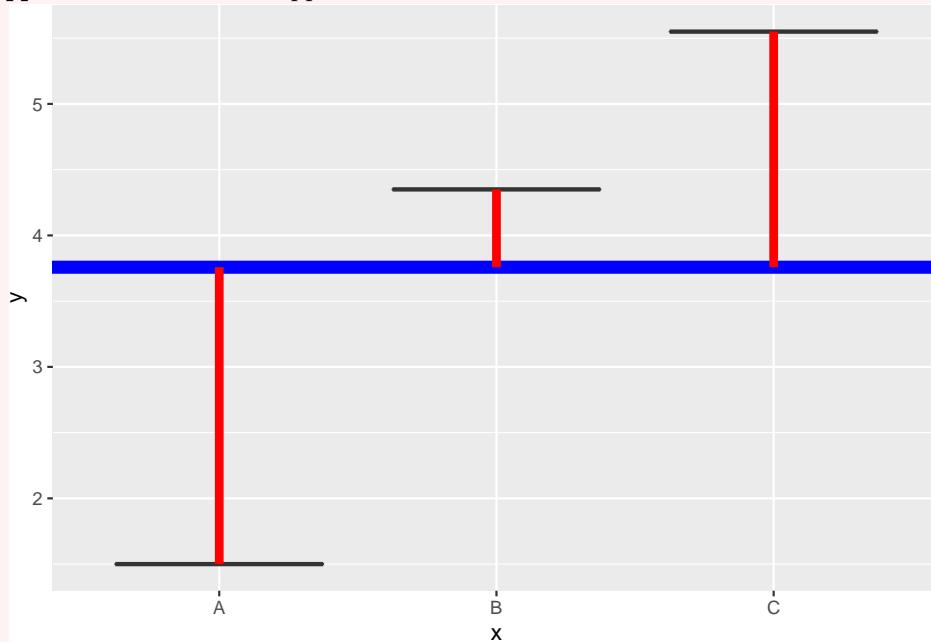
wobei wir hier die verschiedenen Kategorieausprägungen mit dem Index j versehen, bedeutet, dass α_j die mittlere Abweichung des Mittelwertes der Kategorie j vom Mittelwert aller Daten y misst.

Graphische Interpretation der Parameter:

$\mu = \bar{y}_{..}$ = **Gesamtmittelwert**

α_i = **Abstand der Gruppenmittelwerte von Gruppe i vom Gesamtmittelwert**

\bar{y}_i = **Gruppenmittelwert** von Gruppe i



ANOVA Designs

“Balance” eines Designs oder einer Studie bezieht sich darauf, wie häufig die unterschiedlichen Teilgruppen - also Kategorien der Faktorvariable bzw. verschiedenen Stichproben - beobachtet werden. Werden

- Werden alle Kategorien mit gleicher Häufigkeit beobachtet, spricht man von einem **balancierten Design**. Es gilt also, dass wir dieselbe Anzahl an Beobachtungen in jeder Stichprobe zur Verfügung haben.
- Werden die Kategorien mit unterschiedlicher Häufigkeit beobachtet, spricht man von einem **unbalancierten Design**. Es gilt also, dass wir verschiedene Anzahlen an Beobachtungen in jeder Stichprobe zur Verfügung haben, also in manchen den Mittelwert genauer ermitteln können als in anderen. Wir erinnern uns, dass die Trennschärfe stärker wird mit der Anzahl der Beobachtungen. Dies gilt hier in jeder Kategorie.

Balanciertes Design für 3 Kategorien 'A', 'B' und 'C'

A	C	B	A
C	A	C	B
B	C	B	A

Unbalanciertes Design für 3 Kategorien 'A', 'B' und 'C'

C	C	B	A
C	A	C	B
B	C	B	A

Why is ANOVA an analysis of variance?

We have a “sum of squares law”:

$$\underbrace{RSS_{total}}_{\text{total residual sum of squares}} = \underbrace{RSS_1}_{\text{within-sample-variance}} + \underbrace{RSS_0}_{\text{between-sample-variance}}$$

where RSS_1 and RSS_0 are uncorrelated.

Unterschiedliche Anpassungsqualitäten

ANOVA Modell mit klarer Überschneidung der Gruppen

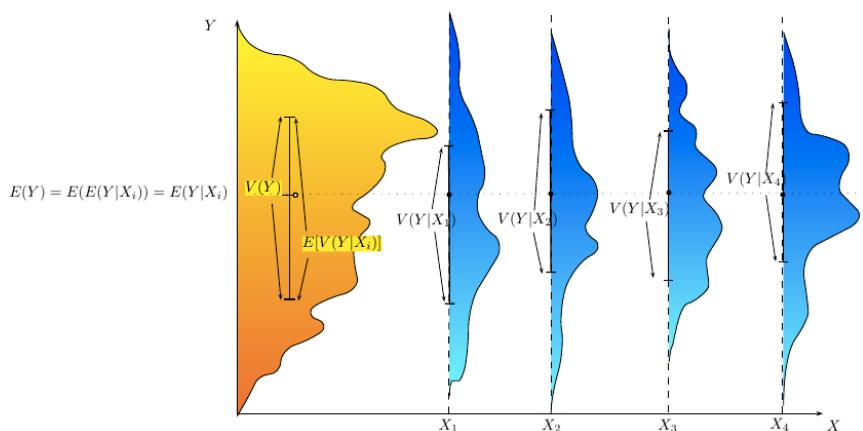


Figure 2: ANOVA : No fit

ANOVA Modell mit klarer Trennung der Gruppen

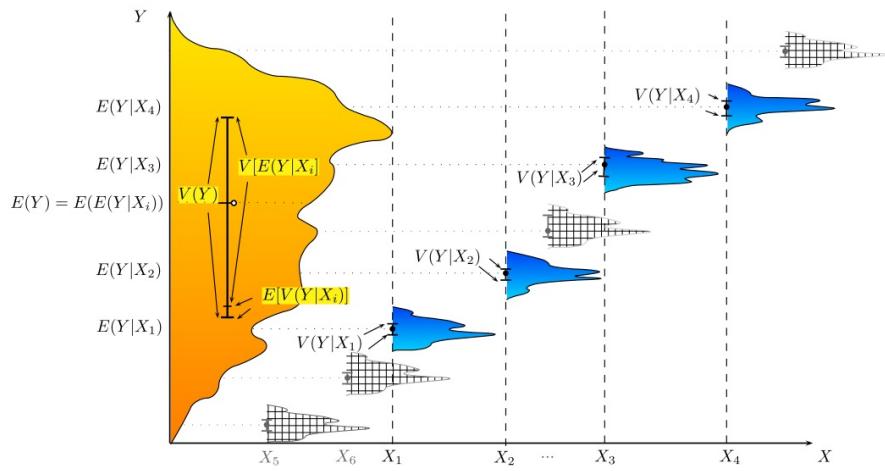


Figure 3: ANOVA : very good fit

ANOVA Modell mit unklarer Trennung der Gruppen

Hier ist der Test am sinnvollsten, da das Ergebnis nicht bereits beim Hinsehen offensichtlich ist.

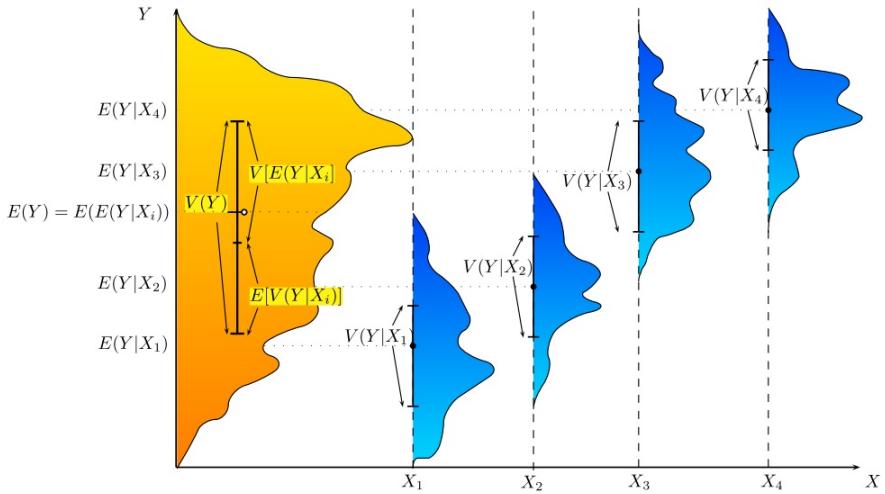


Figure 1: ANOVA : Fair fit

Zweiweg Varianzanalyse (ANOVA)

2-fache ANOVA als Modell

Das Modell wird erweitert um eine zweite erklärende kategoriale Variable ("Zweiweg") mit Kategorien, $m = 1, \dots, M$ zusätzlich zu den Kategorien der ersten erklärenden Variable $k = 1, \dots, K$

$$y_{i,k,m} = \mu + \alpha_k + \beta_m + \epsilon_{i,k,m}$$

wobei wir hier die verschiedenen Kategorieausprägungen mit dem Index k bzw. m für die 1. und 2. kategoriale Variable versehen.

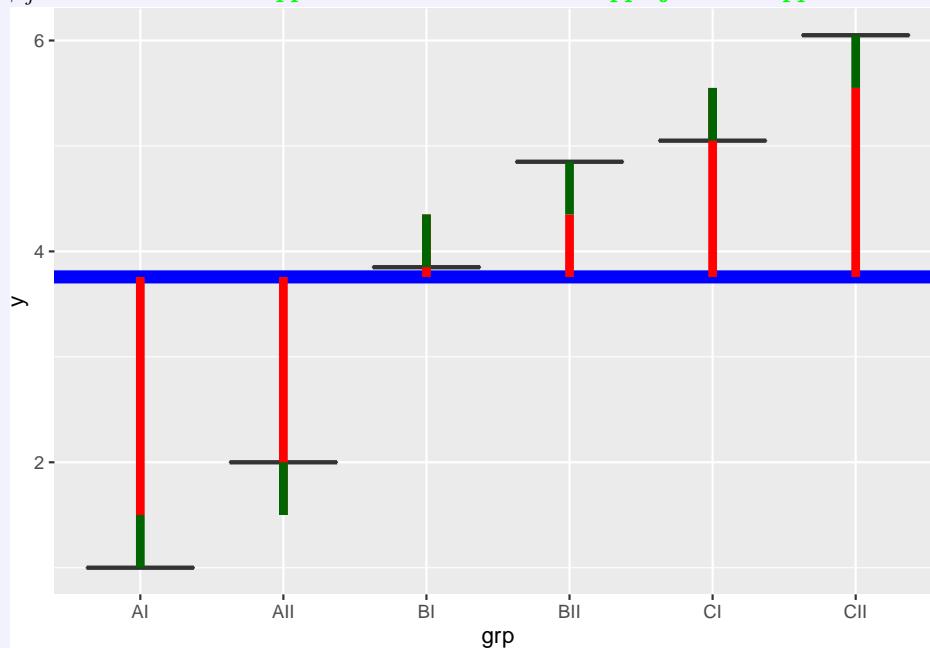
α_k misst also die mittlere Abweichung des Mittelwertes der Kategorie k vom Mittelwert aller Daten y .
 β_m misst also die mittlere Abweichung des Mittelwertes der Kategorie m von Mittelwerten der Unterkategorien $k = 1, \dots, K$ aller Daten y .

Graphische Interpretation der Parameter:

$\mu = \bar{y}_{..}$ = **Gesamtmittelwert**

α_i = **Abstand der Gruppenmittelwerte von x1-Gruppe i vom Gesamtmittelwert**

β_j = **Abstand der Gruppenmittelwerte von x2-Gruppe j vom Gruppenmittelwerte α_i**



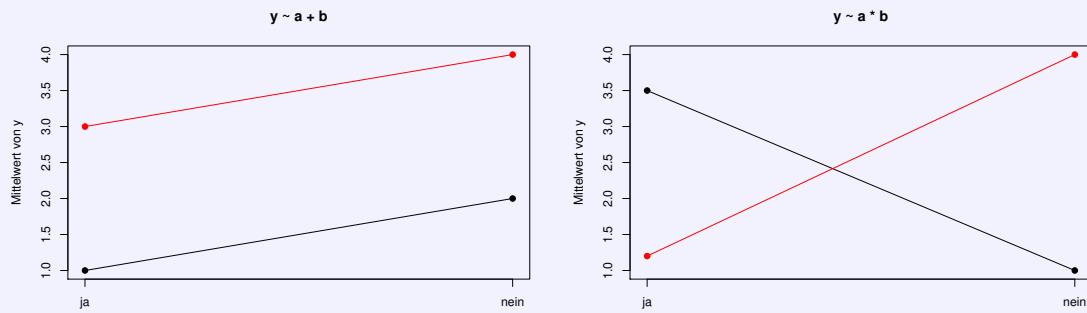
Modelloptionen für die Zweiweg-ANOVA

Bei zwei Variablen mit Unterkategorien, deren Zählungen wir in Kreuztabellen und mit Mosaicplots dargestellt haben, sind also mehrere Szenarien möglich:

- $y \sim 1$ Nur der Mittelwert der gesamten Daten wird als Mittelwert in allen Teilkategorien angenommen.
- $y \sim X_1$ Nur die erste kategoriale Variable X_1 führt zu einer Aufteilung der Mittelwerte, die zweite Variable hat keinen Effekt.
- $y \sim X_2$ Nur die zweite kategoriale Variable X_2 führt zu einer Aufteilung der Mittelwerte, die erste Variable hat keinen Effekt.
- $X_1 + X_2$ Beide kategoriale Variablen X_1 und X_2 führen zu einer Aufteilung der Mittelwerte in den unterschiedlichen Teilkategorien.
- $X_1 * X_2$ Beide kategoriale Variablen X_1 und X_2 führen zu einer Aufteilung der Mittelwerte in den unterschiedlichen Teilkategorien und zusätzlich addieren sich die Effekte nicht einfach auf sondern verstärken einander, schwächen einander gegenseitig ab oder drehen einander um (Interaktion).

Interaction Plot

Der **Interaction Plot** visualisiert das Verhalten der Mittelwerte über Kategorien hinweg. Dabei werden die Richtungen von einer Kategorie zur nächsten als Linien durchgezogen. Die Unterkategorien der 2. Variable werden durch unterschiedliche Linientypen visualisiert.



Example for motivating models: Star-Ratings of Cellular Phone Apps

Customer ratings based on certain features of products are basic data in marketing. We use a cell phone example, where one feature is the design as provided by Apple's iPhones or Samsung's products as an alternative. The second variable is the availability of Apps for the products. Hypothetical categories are defined for marketing experiments, trying to estimate consumer's behaviour.

	X_1 (Design)	
X_2 (Apps)	Apple	Samsung
App Store	10	10
Google Archives	10	10

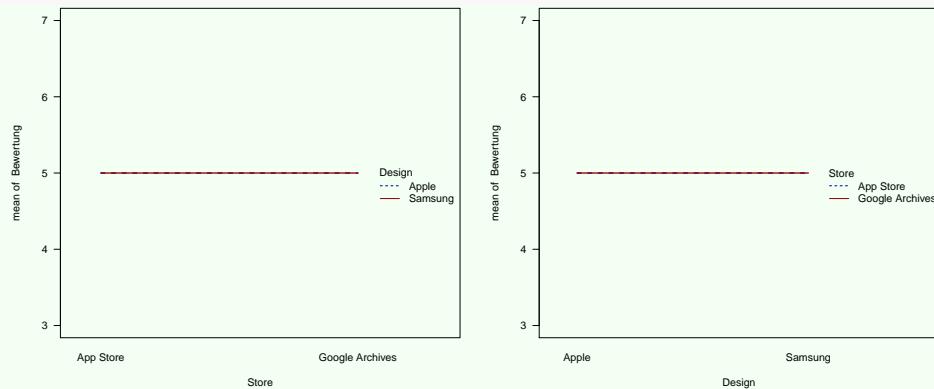
1st case: no effect of factors X_1 and X_2

X_2 (Apps)	X_1 (Design)	
	Apple	Samsung
App Store	5	5
Google Archives	5	5

All combinations yield the same outcome.

Thus, neither X_1 nor X_2 are required for the model (**null-model**).

`lm(Y ~ 1)`



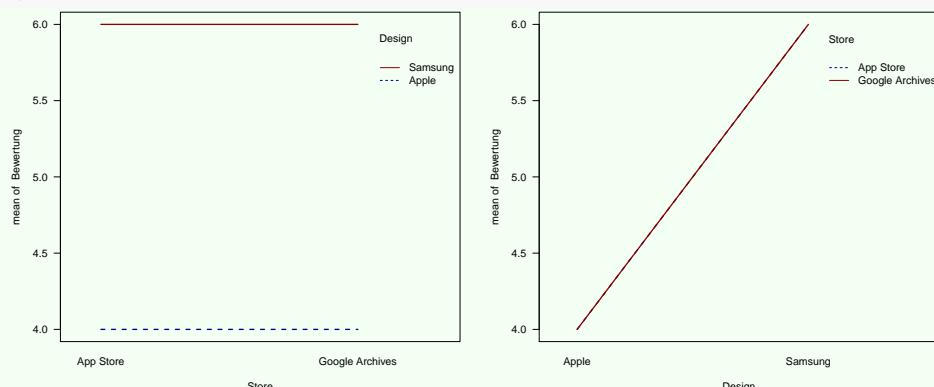
2nd case: effect of X_1 or X_2 only

X_2 (Apps)	X_1 (Design)	
	Apple	Samsung
App Store	4	6
Google Archives	4	6

Changes on means depend only on one of the categorical variables, not the other. The expected values thus do not change when the second factor is changed.

`lm(Y ~ X1)`

`lm(Y ~ X2)`



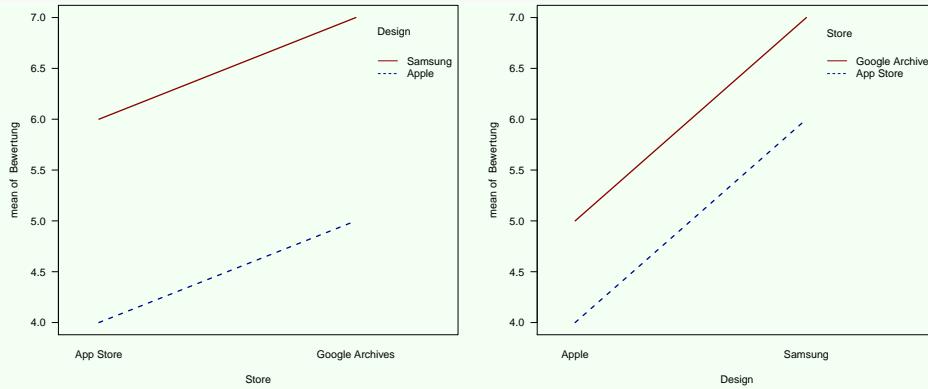
3rd case: additive effects $X_1 + X_2$

X_2 (Apps)	X_1 (Design)	
	Apple	Samsung
App Store	4	6
Google Archives	5	7

Mean of joint categories depend on both factor X_1 (design) and factor X_2 (Apps) independently.

Base model: independent additive effects of factors X_1 and X_2 .

`lm(Y ~ X1 + X2)`



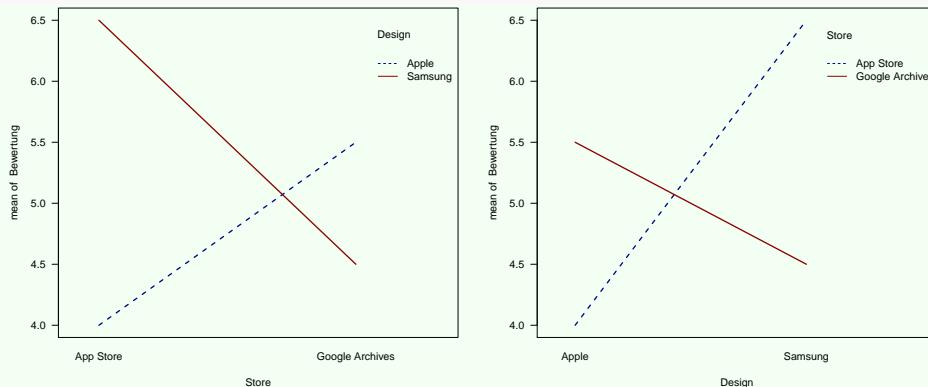
4th case: Interaction between X_1 and X_2

X_2 (Apps)	X_1 (Design)	
	Apple	Samsung
App Store	4	6.5
Google Archives	5.5	4.5

The change on group means depends on both factors X_1 and X_2 which do interact and thus yield different results in combination than each margin.

The corresponding model is the full model with interactions.

`lm(Y ~ X1 * X2)`



Example for motivating models: Plot Yield of Corn Fields

As Fisher's original application, we look at different crop species and fertilizers and their influence on plot yield.

We use an example, where one feature is the Aztec corn vs. Monsanto as an alternative. The second variable is the fertilizers for the products. The variable measure is the mean production in tons per squared-kilometer corn fields.

All data are hypothetical examples.

\begin{center}

		X_1 (Crop Type)	
X_2 (Fertilizer)		Aztec	Monsanto
Monsanto fertilizer	10	10	10
	10	10	10

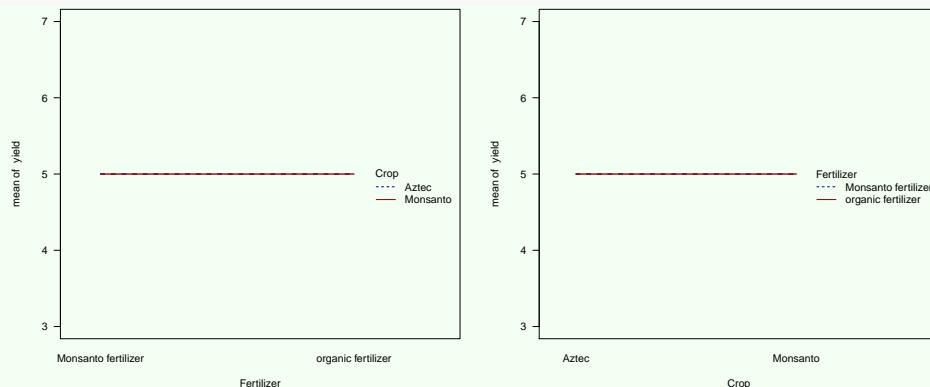
1st case: no effect of factors X_1 and X_2

		X_1 (Crop Type)	
X_2 (Fertilizer)		Aztec	Monsanto
Monsanto fertilizer	5	5	5
	5	5	5

All combinations yield the same outcome.

Thus, neither X_1 nor X_2 are required for the model (**null-model**).

lm(Y ~ 1)



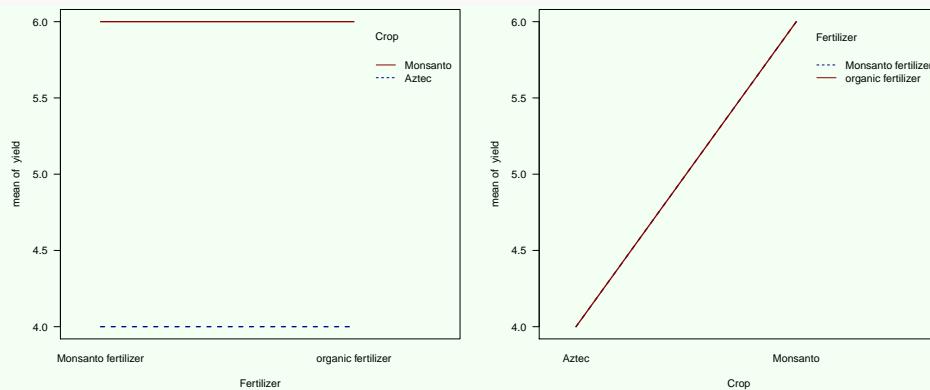
2nd case: effect of X_1 or X_2 only

X_2 (Fertilizer)	X_1 (Crop Type)	
	Aztec	Monsanto
Monsanto fertilizer	4	6
organic fertilizer	4	6

Changes on means depend only on one of the categorical variables, not the other. The expected values thus do not change when the second factor is changed.

`lm(Y ~ X1)`

`lm(Y ~ X2)`



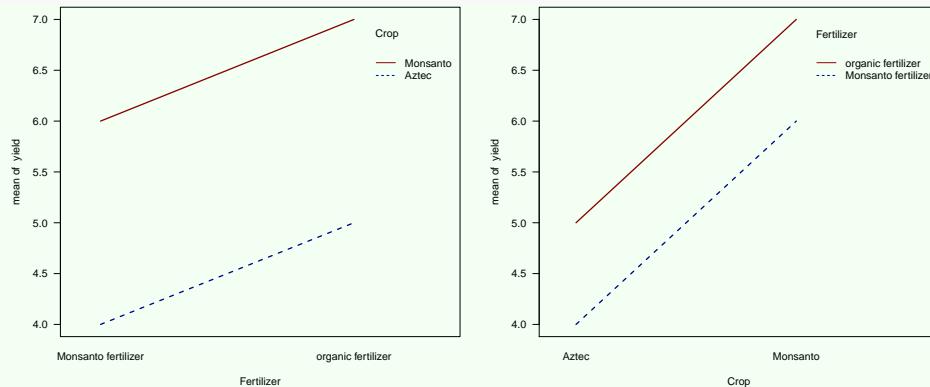
3rd case: additive effects $X_1 + X_2$

X_2 (Fertilizer)	X_1 (Crop Type)	
	Aztec	Monsanto
Monsanto fertilizer	4	6
organic fertilizer	5	7

Mean of joint categories depend on both factor X_1 (design) and factor X_2 (Apps) independently.

Base model: independent additive effects of factors X_1 and X_2 .

`lm(Y ~ X1 + X2)`



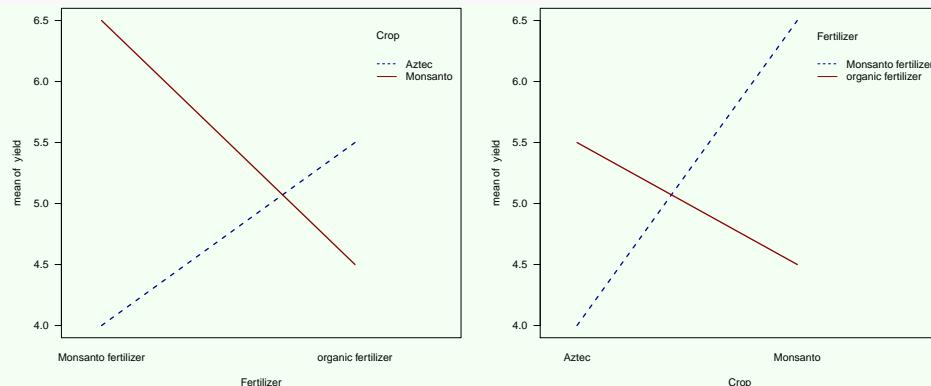
4th case: Interaction between X_1 and X_2

X_2 (Fertilizer)	X_1 (Crop Type)	
	Aztec	Monsanto
Monsanto fertilizer	4	6.5
organic fertilizer	5.5	4.5

The change on group means depends on both factors X_1 and X_2 which do interact and thus yield different results in combination than each margin.

The corresponding model is the full model with interactions.

`lm(Y ~ X1 * X2)`



ANOVA als Methode zur Modellselektion

Wir haben nun einige unterschiedliche Modelle kennengelernt, die für nur 2 erklärende Variablen möglich sind, anzupassen. Daher ist es sinnvoll, dass wir eine Methode zum Vergleichen von Modellen haben. Da die ANOVA Residuenquadratsummen vergleicht, also genau die Summen, die von Regressionsmodellen und ANOVA-Modellen minimiert werden, kann diese Methode auch 1:1 zum Vergleichen unterschiedlicher Modelle miteinander genutzt werden. Die notwendige Bedingung dafür ist, dass die Modelle in einander vollständig enthalten, englisch "nested", sind.

Geschachtelte Modelle für 2 erklärende Variablen

Model	Formula	degrees of freedom df
RSS_0	$Y \sim 1$	$\sum_{i=1}^I n_i - 1$
RSS_1	$Y \sim X_1$	$I_1 - 1$
RSS_2	$Y \sim X_2$	$I_2 - 1$
RSS_{1+2}	$Y \sim X_1 + X_2$	$I_1 + I_2 - 2$
RSS_{1*2}	$Y \sim X_1 * X_2$	$I_1 + I_2 - 4$

Beispiel: Das Modell $y \sim x_1 + x_2 + x_3$ enthält das Modell $y \sim x_1 + x_2$ vollständig, daher kann man sie miteinander vergleichen. Jedoch das Modell $y \sim x_1 + x_4$ ist nicht geschachtelt oder enthalten in $y \sim x_1 + x_2 + x_3$ und daher nicht mithilfe dieser Methode vergleichbar!

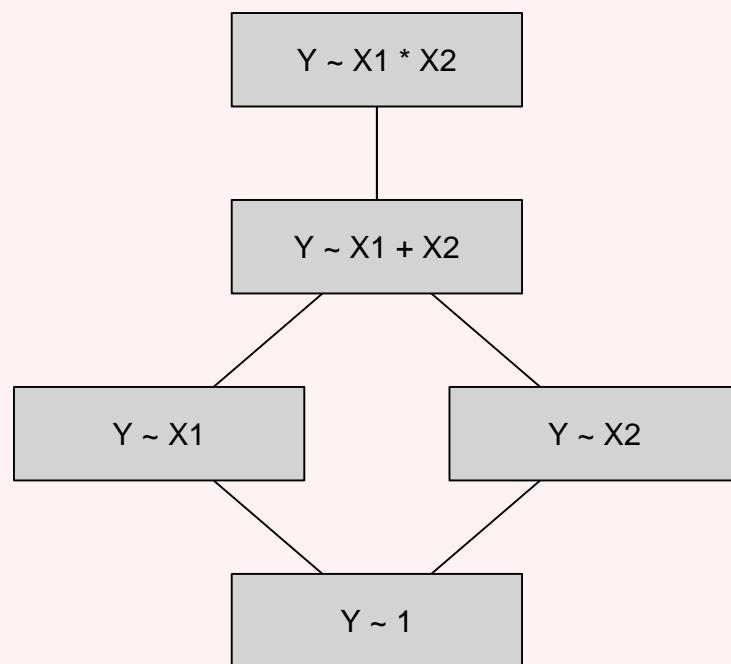
Modellselektion für geschachtelte Modelle

Der F-Test für **Modellselektion** kann durch den folgenden Hypothesentest 2 geschachtelte Modelle miteinander vergleichen.

H_0 : das einfachere Modell genügt

H_1 : mindestens 1 Parameter des komplexeren Modells ist erforderlich

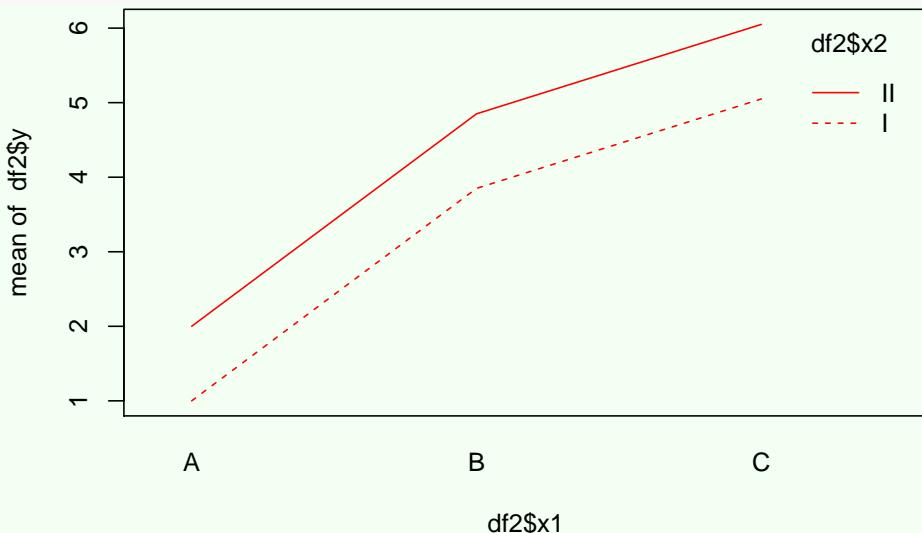
Die Grafik gibt visuell den Zugang zur Modellselektion durch ANOVA wieder. Hier geht es um eine Top-Down-Selektion, die generell für ANOVA als Werkzeug empfohlen wird, was bedeutet, dass wir beim komplexesten Modell mit den meisten erklärenden Variablen und Interaktionen starten und uns von oben nach unten schrittweise durch den Graphen arbeiten. Wenn wir Parameter oder Interaktionen einzeln selektieren wollen, wird also schrittweise vom komplexeren zum nächst weniger komplexen Modell der F-Test durchgeführt. Das ist auch der Grund, weshalb die Modelle geschachtelt sein müssen, da nicht mit einem Modell verglichen werden kann, dass den Parameter nicht, aber dafür andere erklärende Faktoren enthält. Salopp formuliert bedeutet das: "Wir vergleichen nicht Äpfel mit Birnen."



ANOVA Models in R

First, we take a look at the interaction plot

```
interaction.plot(x.factor = df2$x1, trace.factor = df2$x2, response = df2$y)
```



ANOVA fit in R for additive model

ANOVA models are special cases of linear models (see linear regression) and fitted with the same command

```
# fitting additive ANOVA model
anovamodel_additive <- lm(y~x1+x2, data = df2)
anovamodel_additive
##
## Call:
## lm(formula = y ~ x1 + x2, data = df2)
##
## Coefficients:
## (Intercept)      x1B      x1C      x2II
##       1.00       2.85       4.05       1.00
```

ANOVA fit in R for interaction model

```
# fitting ANOVA model with interaction
anovamodel_interaction <- lm(y~x1*x2, data = df2)
anovamodel_interaction
##
## Call:
## lm(formula = y ~ x1 * x2, data = df2)
##
## Coefficients:
## (Intercept)      x1B      x1C      x2II   x1B:x2II   x1C:x2II
##       1.000e+00    2.850e+00    4.050e+00   1.000e+00 -2.663e-15 -1.392e-15
```

ANOVA model comparison in R

```
# comparing the ANOVA models
# with ANOVA for model comparison
anova(anovamodel_additive, anovamodel_interaction)
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 * x2
##   Res.Df      RSS Df  Sum of Sq    F Pr(>F)
## 1     8 8.6104e-30
## 2     6 5.0619e-30  2 3.5485e-30 2.1031 0.2032
# the more complex model is NOT
# significantly better than the simple one
# we therefore keep the simpler model
```

Zusammenhänge zwischen 2 metrischen Variablen

Für kontinuierliche Variablen X und Y kann man sich grundsätzlich über viele Arten des Zusammenhangs Gedanken machen. Um eine Vorstellung von der Art des Zusammenhangs zu bekommen,

Der Einfachheit der Überlegungen und der mathematischen Schätzung halber, wollen wir uns aber im Folgenden auf lineare Abhängigkeit konzentrieren. Die dabei zur Anwendung kommende Methode der linearen Regression wird in der chemischen Analytik auch als Kalibration und die Regressionsgerade als Kalibrationskurve bezeichnet. Wir verbleiben im folgenden bei der in der Statistik und allen wissenschaftlichen Fachbereichen üblichen Bezeichnung der Regression.

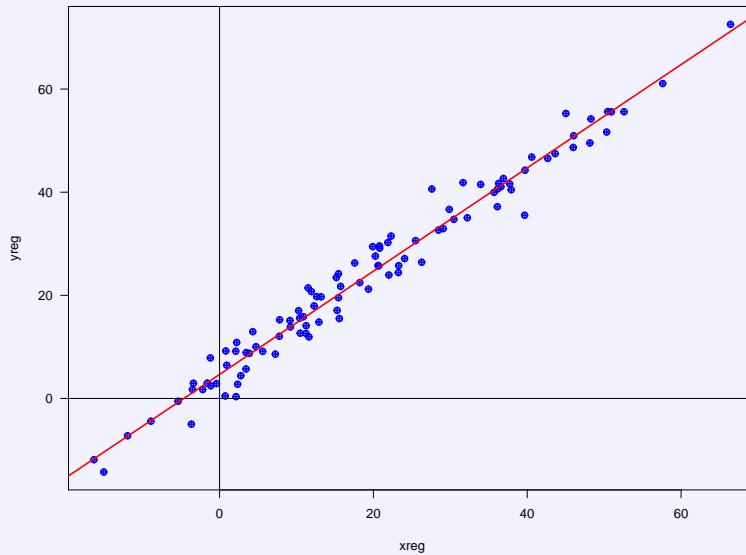
Inferenz durch lineare Regression

Lineare Regression

Hierbei ist die Frage, inwiefern die simultane Messungen von zwei metrischen Merkmalen (x_i, y_i) um eine Gerade mit der *Modellgleichung*

$$y = \alpha + \beta \cdot x,$$

streuen. Hierbei wird aus dem Kontext x als die **unabhängige Variable** bezeichnet, was damit zusammenhängt, dass diese bei Experimenten bewusst eingestellt werden kann, während y als die **abhängige Variable** bezeichnet wird, weil sie sich durch die eingestellten Bedingungen ergibt und gemessen, aber nicht gezielt eingestellt wird.



Allgemein legt man einer solchen Anpassung, welche Modellierung mittels **linearer Regression** genannt wird, eine *Regressionsgleichung*

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

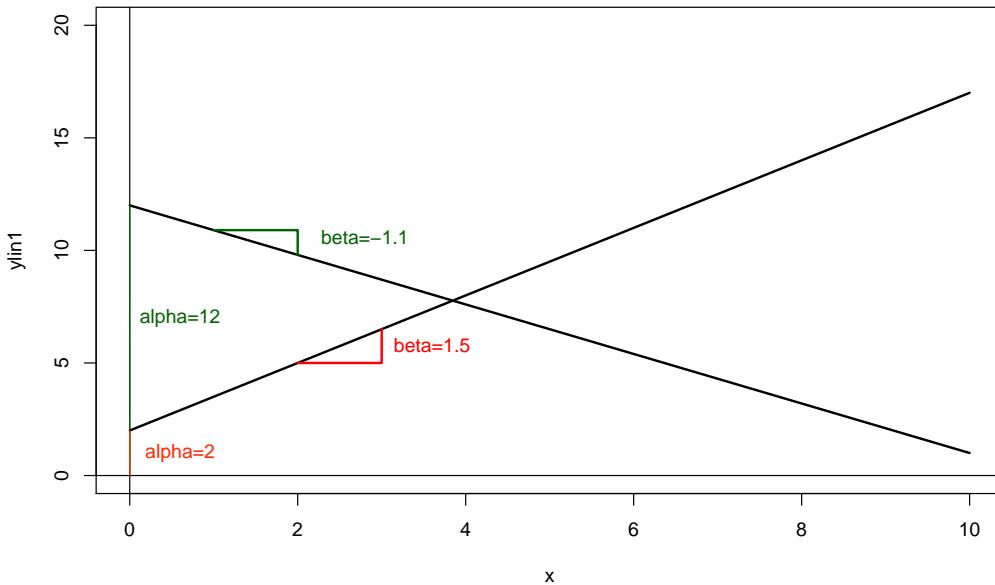
für die Messungen (x_i, y_i) zugrunde. Die Parameter haben dabei mathematische und kontextuelle Bedeutung:

- α ist der Achsenabschnitt der Gerade (auf der y-Achse). Dies bedeutet, dass es der erwartete Wert der y-Variable ist, wenn die x-Variable den Wert 0 annimmt.
- β ist die Steigung der Gerade. Dies bedeutet, dass der Wert der y-Variable um β Einheiten zunimmt, wenn die x-Variable um 1 Einheit vergrößert wird.
- ε_i ist der Residuenfehler der Punkte. Dies bedeutet, dass der Punkt (x_i, y_i) von der Geraden ε_i Einheiten in vertikaler Richtung entfernt ist.

Zur mathematischen Bedeutung der Parameter:

- α ist der Wert auf der y-Achse, an dem die Gerade die y-Achse schneidet, oder auf der Startwert für $x=0$
- β ist die Steigung der Gerade.
 - für $\beta > 0$ ist die gerade steigend
 - für $\beta < 0$ ist die gerade fallend

Lineares Wachstum und Zerfall



Beispiel: Messungen der Dichte eines Gases in mg/m^3 in Abhängigkeit von der Temperatur in $^{\circ}C$ werden durchgeführt.

Temperatur	Dichte
1	-1.11
2	1.85
3	1.82
4	-1.12
5	-2.43
6	-0.24
7	4.86
8	-3.01
9	-0.15
10	0.53

Dafür wird eine Regressionsgleichung der Form

$$\text{Dichte}_i = \alpha + \beta \cdot \text{Temperatur}_i + \varepsilon_i$$

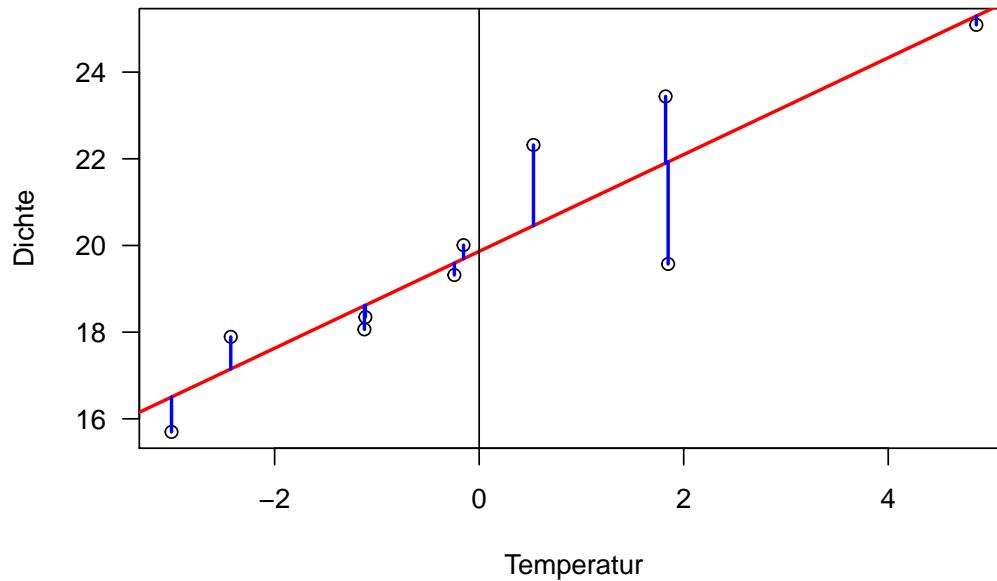
Die aus der Regressionsschätzung resultierende Modellgleichung lautet

$$\text{Dichte} = 19.86 + 1.12 \cdot \text{Temperatur}$$

Das bedeutet, dass der erwartete Wert der Dichte bei einer Temperatur von $0^{\circ}C$ $19.86\ mg/m^3$ beträgt. Pro $1^{\circ}C$ um das die Temperatur ansteigt, steigt die Dichte um $1.12\ mg/m^3$.

Graphisch umgesetzt, sieht dieses Modell folgendermaßen aus, wobei die Gerade gemäß der Modellgleichung in rot und die Residuen in blau dargestellt sind:

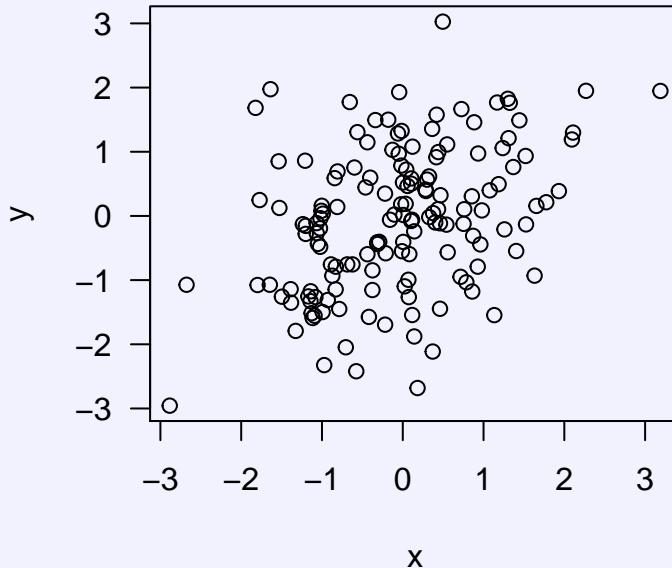
Regressionsgerade und Residuen



Visualisierung und Korrelation als Maß für linearen Zusammenhang

Die graphische Darstellung der Zusammenhänge zweier metrischer Variablen erfolgt durch ein **Streudiagramm** (engl. scatterplot). Dabei wird die unabhängige Variable auf der x-Achse und die abhängige Variable auf der y-Achse aufgetragen und die einzelnen simultanen Messungen als Punkte im Koordinatensystem dargestellt.

Streudiagramm



So wie man Lage und Streuung von Daten mit unterschiedlichen Maßzahlen wie Mittelwert und Varianz oder Median und Interquartilsdistanz messen kann, misst man den Grad des linearen Zusammenhangs mithilfe der Maße **Kovarianz** und **Korrelation**.

Die **Kovarianz** σ_{xy} und **Korrelation** ρ_{xy} bilden daher als Maß der linearen Abhängigkeit zweier metrischer Variablen eine der wichtigsten Grundlagen für lineare Regression.

Kovarianz und Korrelation

Die Kovarianz zwischen 2 metrischen Merkmalen X und Y wird geschätzt durch

$$\widehat{\text{cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}).$$

Sie ist das Äquivalent der Varianz in Bezug auf quadratische Abweichungen in 2 Raumrichtungen x und y simultan.

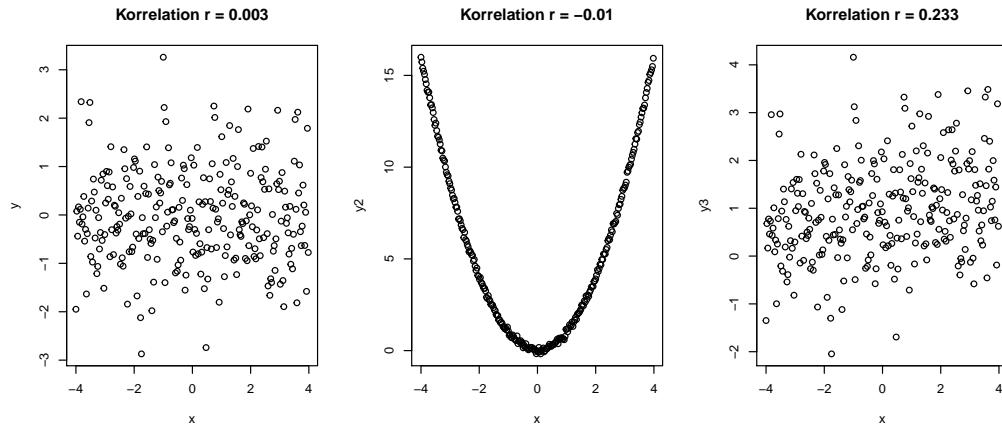
Als standardisiertes Maß für die lineare Abhängigkeit wird der **Pearson Korrelationskoeffizient** definiert durch

$$r = r(X, Y) = \frac{\widehat{\text{cov}}(X, Y)}{s(X) \cdot s(Y)}.$$

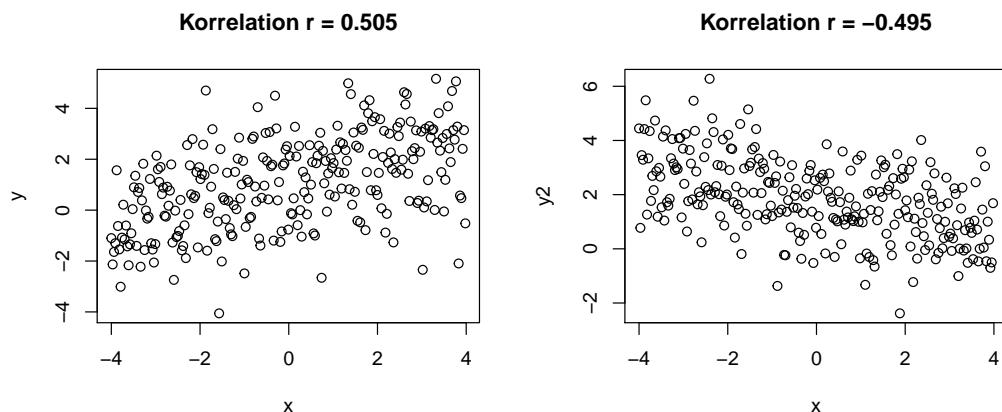
Bei diesem Koeffizient wird die Achsenkalierung herausgerechnet und dadurch nimmt er ausschließlich Wert zwischen -1 und 1 an.

Die Interpretation der Größenordnung der Zahlenwerte der Korrelation ist wichtig, um die Sinnhaftigkeit der linearen Regressionsschätzung bereits im Vorfeld einschätzen zu können. Eine Korrelation von etwa 0, also

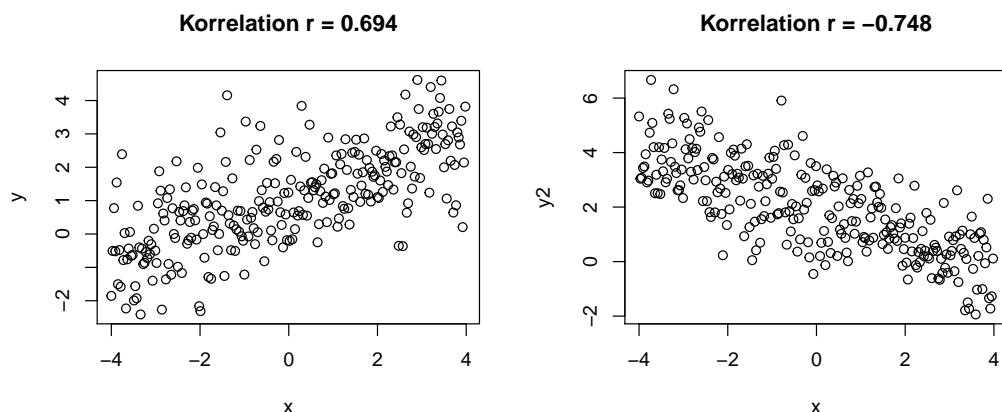
Werte zwischen ungefähr -0.35 und +0.35, bedeutet, dass kein deutlicher linearer Zusammenhang zwischen den Variablen besteht. Vorsicht ist geboten, da das nicht bedeuten muss, dass gar kein Zusammenhang besteht, sondern es bedeutet nur, dass kein linearer Zusammenhang besteht.



Ein Korrelation im Bereich von -0.55 bis -0.35 und 0.35 bis 0.55 spricht für einen mäßigen linearen Zusammenhang, fallend bei negativen Korellationswerten, steigend bei positiven Korrelationswerten. Ein solcher kann in der Modellierung durchaus relevant bzw. signifikant sein, aber graphisch ist er nicht sehr deutlich merkbar.

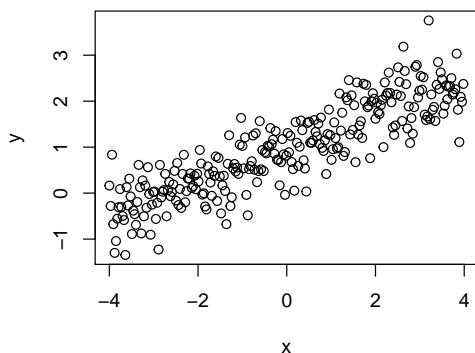


Im Bereich von 0.55 bis 0.75 ist der lineare Zusammenhang bereits optisch deutlich erkennbar, aber die Werte streuen noch sehr stark um die Gerade, was bedeutet, dass die Residuen sehr groß sein werden.

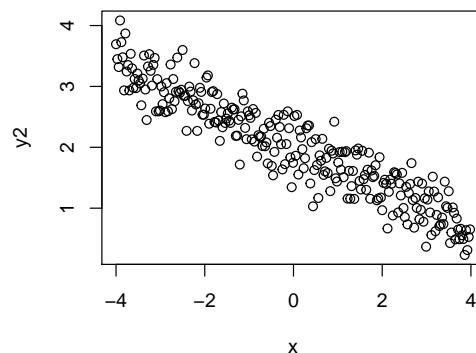


Im Bereich von 0.75 bis 0.95 ist der lineare Zusammenhang sehr deutlich erkennbar und die Werte streuen nicht mehr stark um die Gerade, was bedeutet, dass die Residuen klein im Vergleich mit den anderen Szenarien sein werden.

Korrelation $r = 0.872$

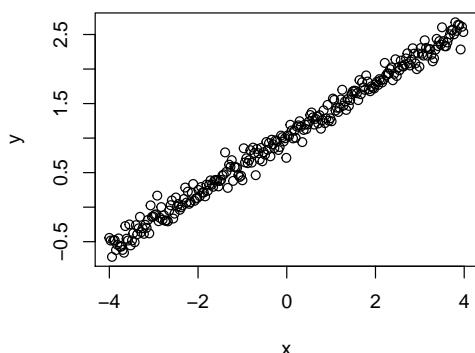


Korrelation $r = -0.932$

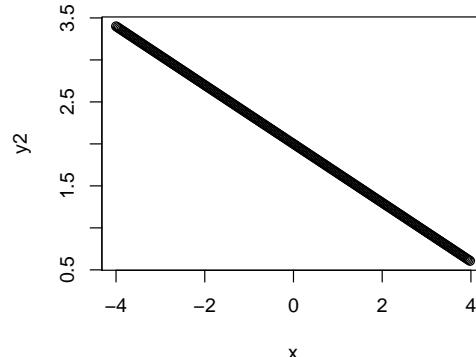


Korrelationswerte über 0.95 deuten auf nahezu perfekte lineare Zusammenhänge hin. Hierbei sollte man vorsichtig sein, zu perfekte Daten sollten Misstrauen erregen. Wenn die Korrelation exakt 1 oder -1 beträgt, ist es viel wahrscheinlicher, dass jemandem ein Fehler beim Kopieren einer Tabellenspalte passiert ist, als dass einem das perfekte Experiment gelungen wäre.

Korrelation $r = 0.994$



Korrelation $r = -1$



Spearman's Rangkorrelation

Ränge sind uns bereits von der Lage- und Streuungsschätzung her bekannt und wir wissen, dass sie robust sind und außerdem nicht nur für metrische sondern auch für ordinale Daten verwendbar sind. Dieses Maß ist außerdem nicht-parametrisch, da es ohne Annahme annähernder Normalverteilung auskommt.

Daher definiert der Spearman Korrelationskoeffizient die entsprechend rangbasierte Messung der Korrelation:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

bei der d_i die Differenzen zwischen den Rängen zweier Werte innerhalb des i-ten Paars "without ties", ohne doppelt auftretende Ränge, sind.

Umsetzung in R erfolgt durch den Befehl

```
cor(x, method = c('pearson', 'spearman'))
```

Hier gibt es die Möglichkeit zwischen Pearson-Korrelation und Spearman-Korrelation zu wechseln.

Da Regression grundsätzlich nicht nur auf die Modellierung der Zusammenhänge zwischen 2 Variablen beschränkt ist, sondern auch für größere Modelle verwendet werden kann, bei denen mehrere unabhängige Variablen vorkommen.

Korrelationsmatrix

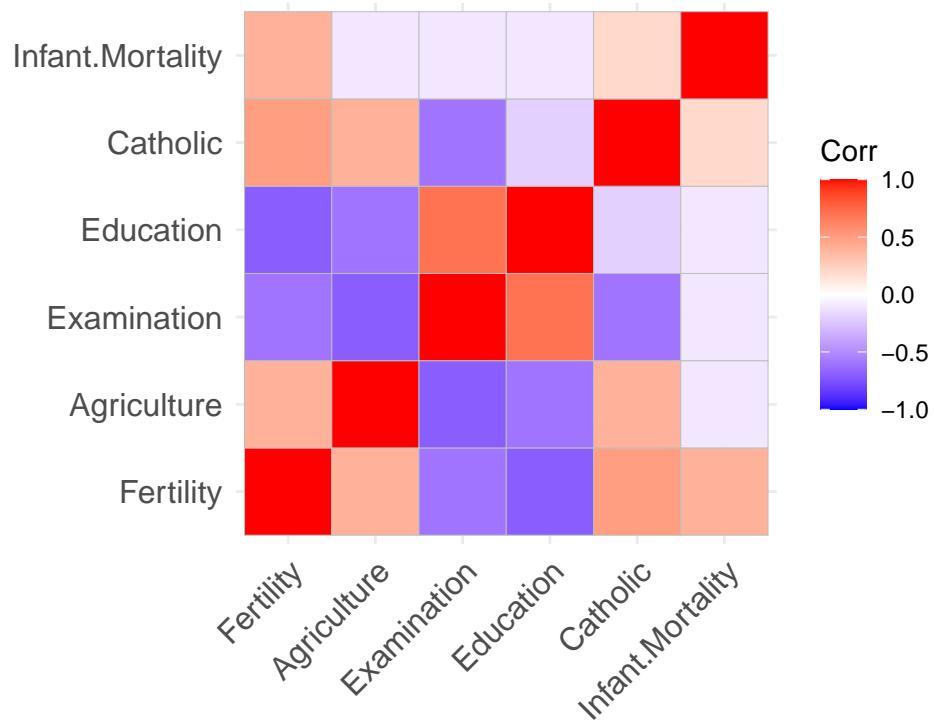
Die hochdimensionale Verallgemeinerung der Kovarianz ist die **Kovarianzmatrix** $Cov(\mathbf{X})$ und die der Korrelation die **Korrelationsmatrix** $Cor(\mathbf{X})$.

$$Cov(\mathbf{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_{x_1 x_2} & \dots & \sigma_{x_1 x_n} \\ \sigma_{x_2 x_1} & \sigma_2^2 & \dots & \sigma_{x_2 x_n} \\ \vdots & & & \\ \sigma_{x_n x_1} & \sigma_{x_n x_2} & \dots & \sigma_n^2 \end{pmatrix},$$

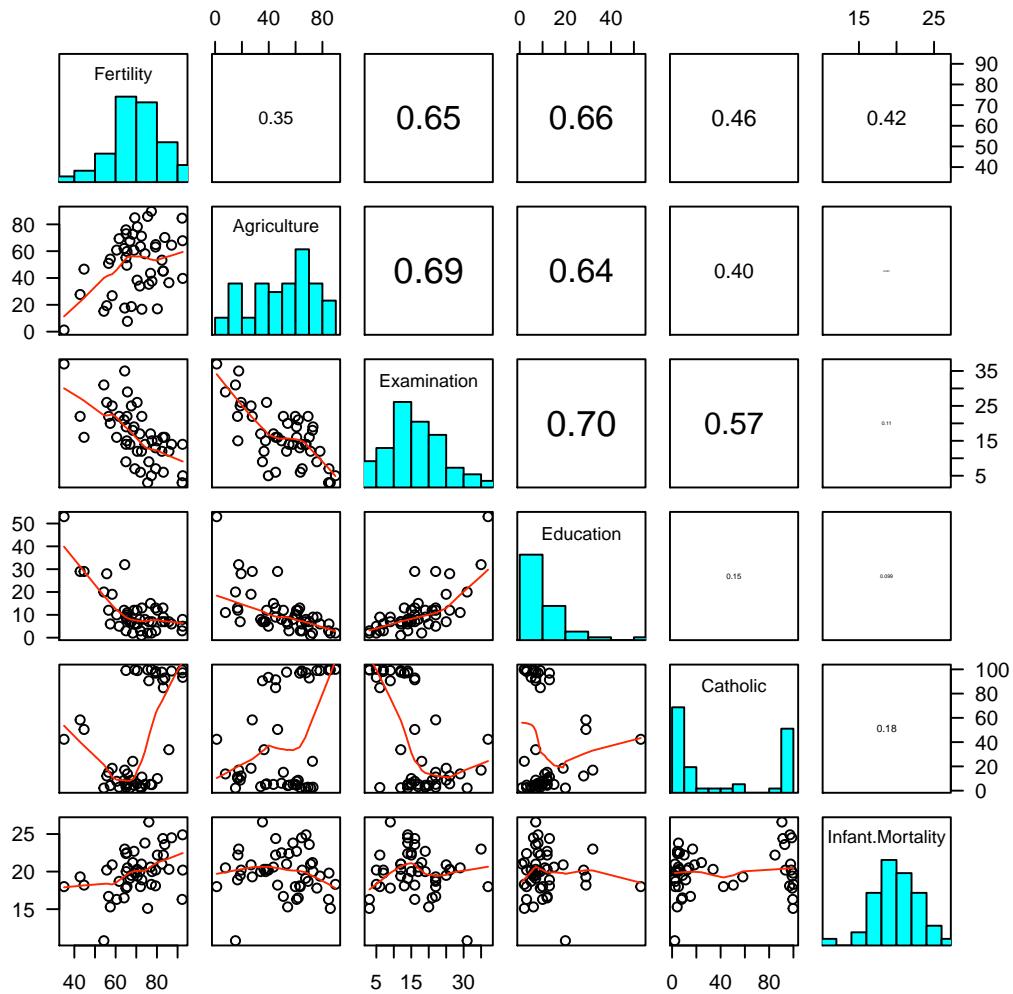
$$Cor(\mathbf{X}) = \begin{pmatrix} 1 & \rho_{x_1 x_2} & \dots & \rho_{x_1 x_n} \\ \rho_{x_2 x_1} & 1 & \dots & \rho_{x_2 x_n} \\ \vdots & & & \\ \rho_{x_n x_1} & \rho_{x_n x_2} & \dots & 1 \end{pmatrix}$$

Die graphische Darstellung dieser Zusammenhänge erfolgt auf zweierlei Weise:

- Ein **Korrelogramm** stellt direkt die Größenordnung der Korrelation gemäß einer ähnlichen wie der obigen Größenordnungseinschätzung. Dabei sind codiert eine Farbskala das Vorzeichen und eine Hell-Dunkel-Skala die Größenordnung.



- Ein **paarweises Streudiagramm** (engl. pairwise scatterplot) stellt die Verläufe der linearen Zusammenhänge in Streudiagrammen von je 2 Variablen dar.



Der dazu gehörend R Code ist

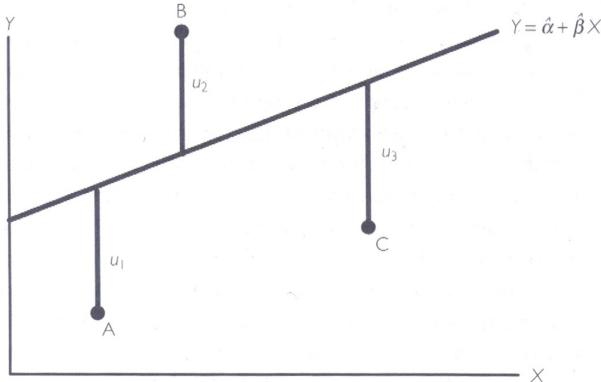
```
panel.hist <- function(x, ...)  
{  
  usr <- par("usr"); on.exit(par(usr))  
  par(usr = c(usr[1:2], 0, 1.5) )  
  h <- hist(x, plot = FALSE)  
  breaks <- h$breaks; nB <- length(breaks)  
  y <- h$counts; y <- y/max(y)  
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)  
}  
  
puts histograms in the diagonal  
  
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)  
{  
  usr <- par("usr"); on.exit(par(usr))  
  par(usr = c(0, 1, 0, 1))  
  r <- abs(cor(x, y))  
  txt <- format(c(r, 0.123456789), digits = digits)[1]  
  txt <- paste0(prefix, txt)  
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)  
  text(0.5, 0.5, txt, cex = cex.cor * r)  
}  
  
puts (absolute) correlations on the upper panels, with size proportional to the correlations.  
  
pairs(swiss, lower.panel = panel.smooth, upper.panel = panel.cor,  
      diag.panel = panel.hist, las=1)
```

generiert den Plot mit allen Details.

Kleinste Quadrate Methode

Das optimale Modell wird bestimmt, indem die Fehler so klein wie möglich gemacht werden sollen.

Allerdings sind manche Residuenfehler $e_i = (y_i - \hat{\alpha} - \hat{\beta}x_i)$ negativ und andere positiv, weshalb wir, um alle Fehler zu berücksichtigen und nicht die positiven durch die negativen zu reduzieren, alle Residuen 'positiv bekommen' wollen.



Die erste Idee wäre also, die Beträge der Residuen zu wählen anstatt der Residuen selbst. Das erscheint ja auch ganz sinnvoll, doch nun erinnern wir uns daran, was wir eigentlich vorhatten, nämlich die Fehler so kleine wie möglich zu bekommen, also das Minimum der Summe ihrer Beträge zu ermitteln, mathematisch

$$\sum_{i=1}^n |e_i| = \sum_{i=1}^n |y_i - \hat{\alpha} - \hat{\beta}x_i| \rightarrow \min.$$

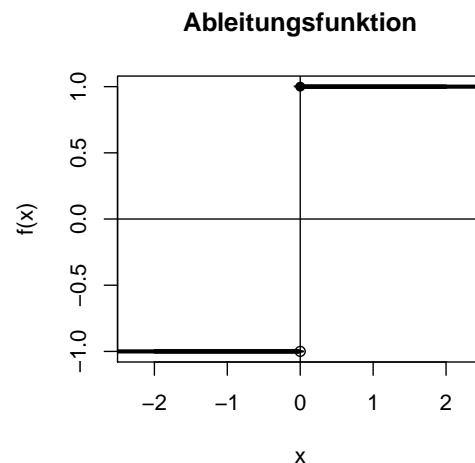
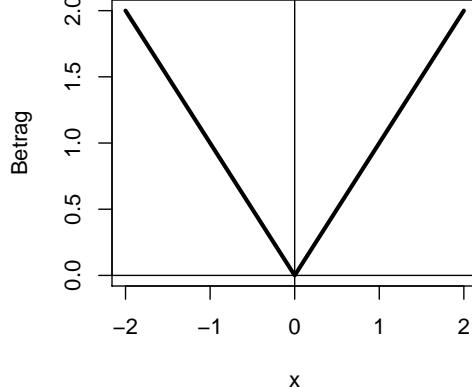
Aus der Differentialrechnung, von Kurvendiskussionen und Optimierung wissen wir, wie die Ermittlung eines Minimums erfolgt: Man bildet die 1. Ableitung und ermittelt deren Nullstellen.

Betrachten wir dafür die Betragsfunktion

$$f(x) = |x| = \begin{cases} -x & x < 0 \\ x & x \geq 0 \end{cases}$$

und ihre Ableitung

$$f'(x) = \frac{d}{dx}|x| = \begin{cases} -1 & x < 0 \\ 1 & x \geq 0 \end{cases}$$



Da die Betragsfunktion über die x-Achse springt, gibt es keine Nullstelle, auch wenn wir optisch ein Minimum im rechten Plot erkennen können. Das ist das Problem mit einer nicht stetig differenzierbaren Fehlerfunktion - wir können kein Minimum numerisch ermitteln.

Also gehen wir zu einer neuen Idee über, um die Residuen "positiv zu bekommen", wir **quadrieren** sie.

Wie zuvor bilden wir die 1. Ableitung und ermitteln deren Nullstellen. Die quadratische Funktion

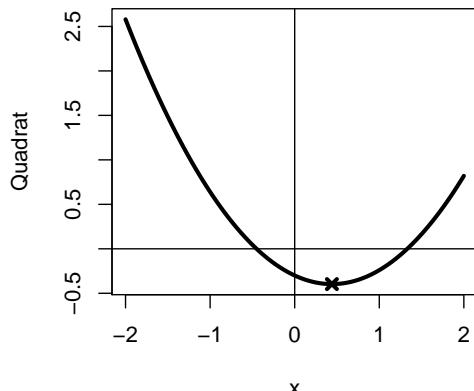
$$f(x) = a \cdot x^2 + b \cdot x + c$$

und ihre Ableitung

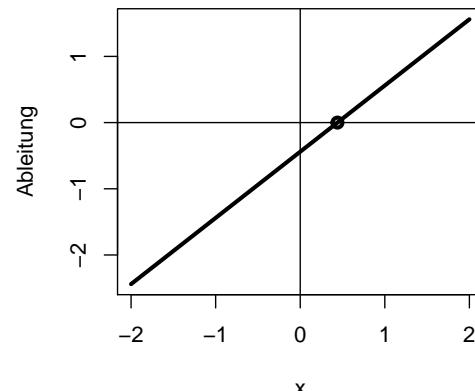
$$f'(x) = 2 \cdot a \cdot x + b$$

sind stetig. Die Ermittlung der Nullstelle ist also kein Problem.

Quadratische Funktion



Ableitungsfunktion

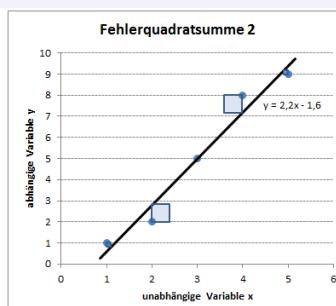
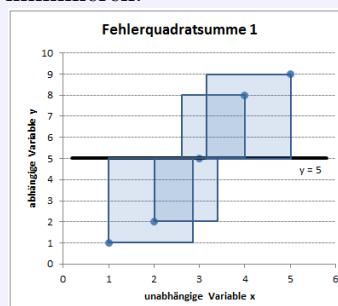


Methode der kleinsten Quadrate (Ordinary Least Squares = OLS)

Das optimale Modell wird ermittelt, indem wir die **Summe der quadrierten Residuen (Residuenquadratsumme)** (englisch residual sum of squares (RSS))

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2, i = 1, 2, \dots, N.$$

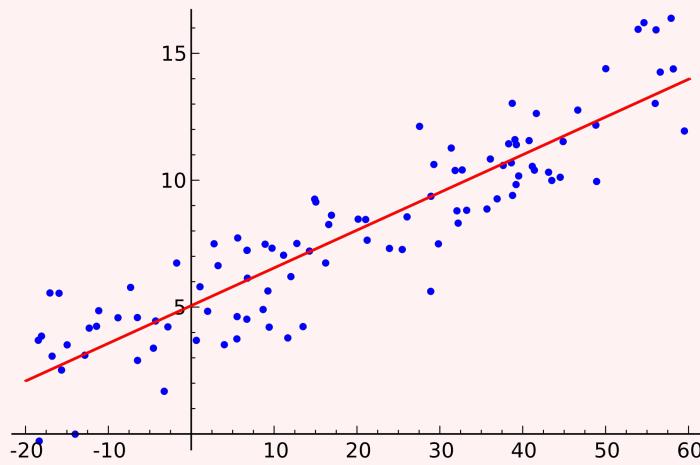
minimieren. Optisch bedeutet das, dass wir die Summe der Fläche der Quadrate, deren Seitenlängen die Residuen sind, minimieren.



Für das einfache lineare Regressionsmodell ergeben sich als Modellparameter

$$\begin{aligned}\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= r_{xy} \frac{s_y}{s_x} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum x_i^2 - \bar{x}^2}\end{aligned}$$

Insbesondere fällt uns hier auf, dass der Regressionskoeffizient $\hat{\beta}$ dasselbe Vorzeichen wie der Korrelationskoeffizient r_{xy} hat.



Multiple lineare Regression

Wenn wir anstatt nur einer unabhängigen erklärenden Variablen mehrere benutzen, gehen wir von simplen linearen Regressionsmodell zum multiplen linearen Regressionsmodell über:

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} + \varepsilon_i$$

In der Notation mittels Matrizen und Vektoren wird es auch angeschrieben als

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

und sieht damit dem einfachen linearen Modell wieder so ähnlich, dass man den Zusammenhang erkennen kann. Dabei sind die Bestandteile

\mathbf{y} ... n-dimensionaler Vektor der Beobachtungen

\mathbf{X} ... $n \times p$ Matrix der p Beobachtungsspalten

$\boldsymbol{\beta}$... p -dimensionaler Vektor der Regressionskoeffizienten
der p verschiedenen Variablen

$\boldsymbol{\varepsilon}$... n-dimensionaler Vektor der Fehler (Residuen)

Annahmen und Voraussetzungen für multiple Regression

Die Schätzung kann nur funktionieren, wenn die Daten bestimmte Annahmen erfüllen. Diese Annahmen betreffen insbesonders den Modellfehler ε , aber auch die Regressoren X_j .

(A1) Das Modell hat keinen systematischen Fehler.

$$E(\varepsilon_i) = 0$$

(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß (homoskedastisch).

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

(A3) Die Komponenten des Fehlerterms sind nicht korreliert.

$$\text{COV}(\varepsilon_i, \varepsilon_j) = 0$$

(A4) Der Modellfehler sei normalverteilt.

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

(A5) Es gibt keine lin. Abhängigkeiten zwischen den Regressoren.

$$\text{rank}(X) = p$$

ANOVA als lineares Modell

Bisher haben wir im linearen Regressionsmodell

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} + \varepsilon_i$$

als Erklärungsvariablen immer nur quantitative Variablen x_{ij} verwendet. Aber natürlich können auch qualitative Erklärungsvariablen einen Einfluß auf y_i haben.

In der explorativen Datenanalyse haben wir dies Problem auch schon durch parallele Boxplots visualisiert.

Modellselektion

Wenn mehrere Variablen als erklärend erwogen werden, müssen noch lange nicht alle erklärend sein. Man kann daher auf unterschiedliche Weise selektieren, welche relevant sind.

Es gilt: "so klein wie möglich, so groß wie notwendig soll ein Model sein"

- **t-Tests** für die Koeffizienten β_i
einfacher und direkter Weg, auf linearen Zusammenhang zu testen. (inkludiert in summary(linearesModell))
- **ANOVA** für geschachtelte Modelle
ANOVA kann geschachtelte Modelle miteinander vergleichen
- allgemeine schrittweise Modellselection mittels "goodness-of-fit" Maßen
nutzt Maßzahlen wie **Akaike Information Criterion (AIC)** oder **Bayesian Information Criterion (BIC)** zum Vergleich von Modellen und wählt das am besten Passende aus.

Erweiterte Konzepte: Voraussetzungen für die Anpassung eines Regressionmodells

Die Schätzung kann nur funktionieren, wenn die Daten bestimmte Annahmen erfüllen.

Diese Annahmen betreffen insbesondere beim einfachen linearen Regressionsmodell den Modellfehler ε , nicht notwendigerweise die Daten selbst!.

(A1) Das Modell hat keinen systematischen Fehler.

$$E(\varepsilon_i) = 0$$

(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß (homoskedastisch).

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

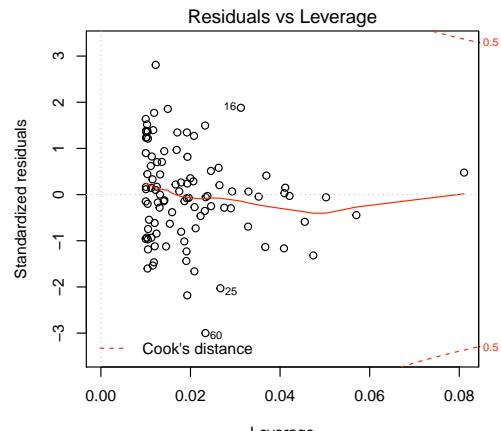
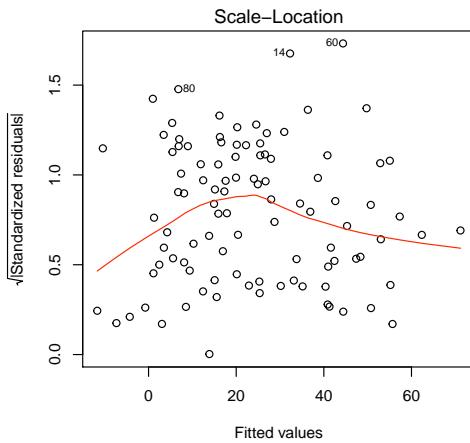
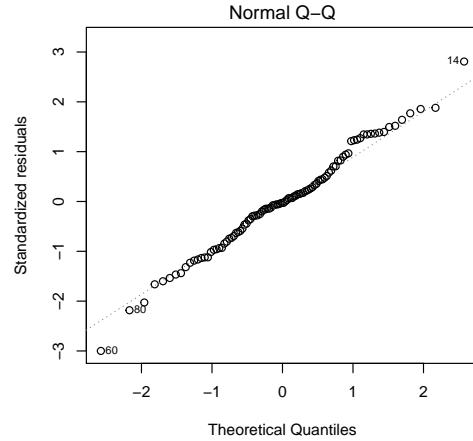
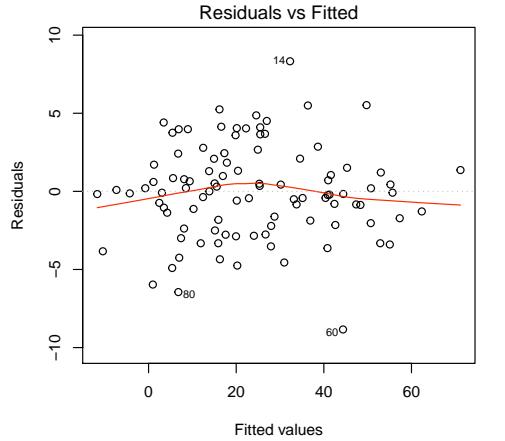
(A3) Die Komponenten des Fehlerterms sind nicht korreliert.

$$\text{COV}(\varepsilon_i, \varepsilon_j) = 0$$

(A4) Der Modellfehler sei normalverteilt.

$$\varepsilon_i \sim N(0, \sigma^2)$$

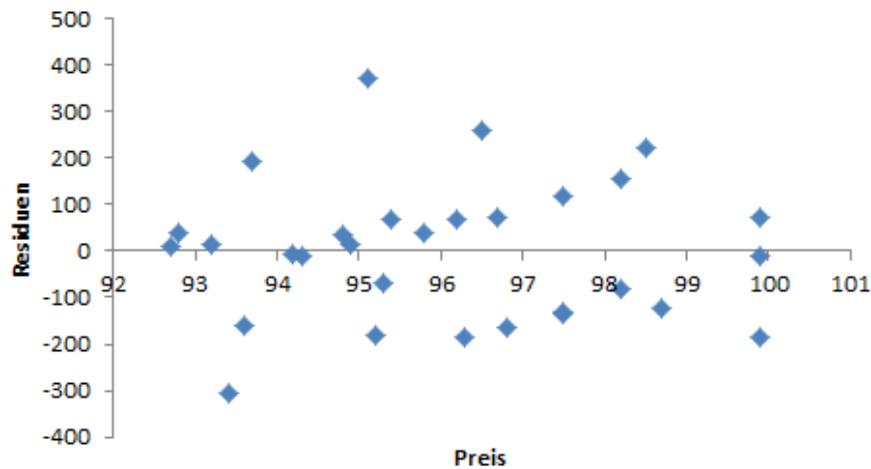
Die Überprüfung der Annahmen geschieht mithilfe der Residuenplots.



Visualisierung unkorrelierter und korrelierter Fehler

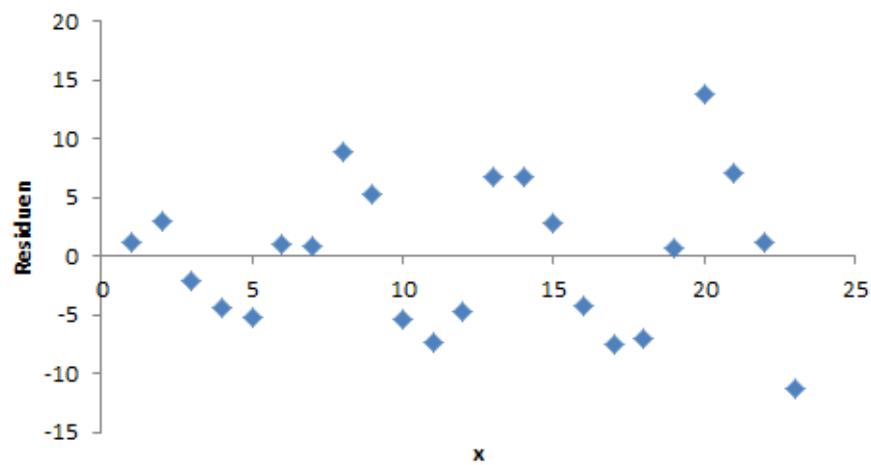
Unkorrelierte Fehler zeigen keine Struktur.

Residuenplot: unkorrelierte Residuen



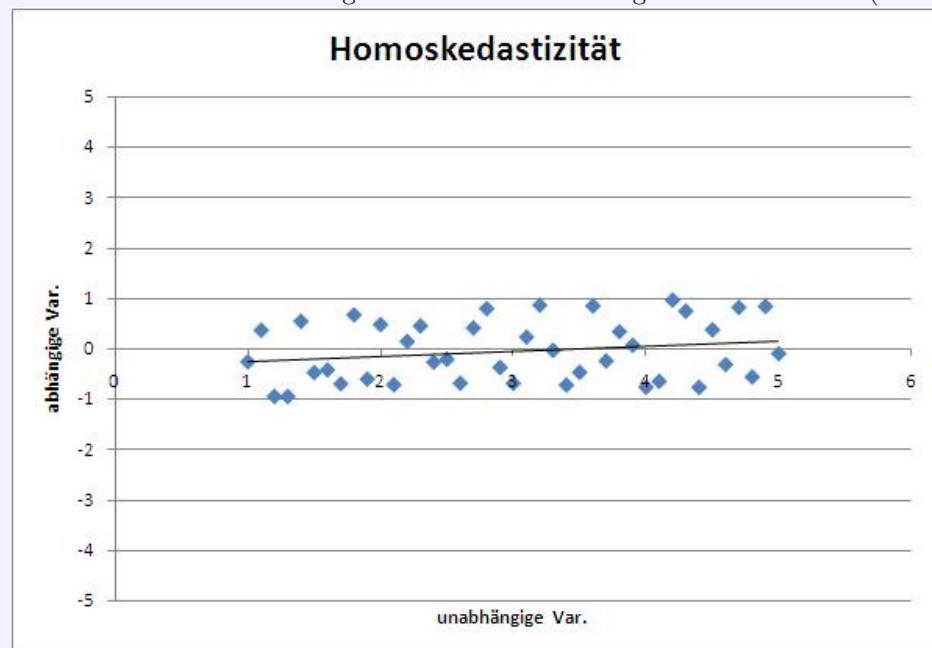
Bei korrelierten Fehlern lässt sich eine Struktur im Verlauf der Residuen erkennen, wie hier eine sinusartige Funktion.

Residuenplot: korrelierte Residuen

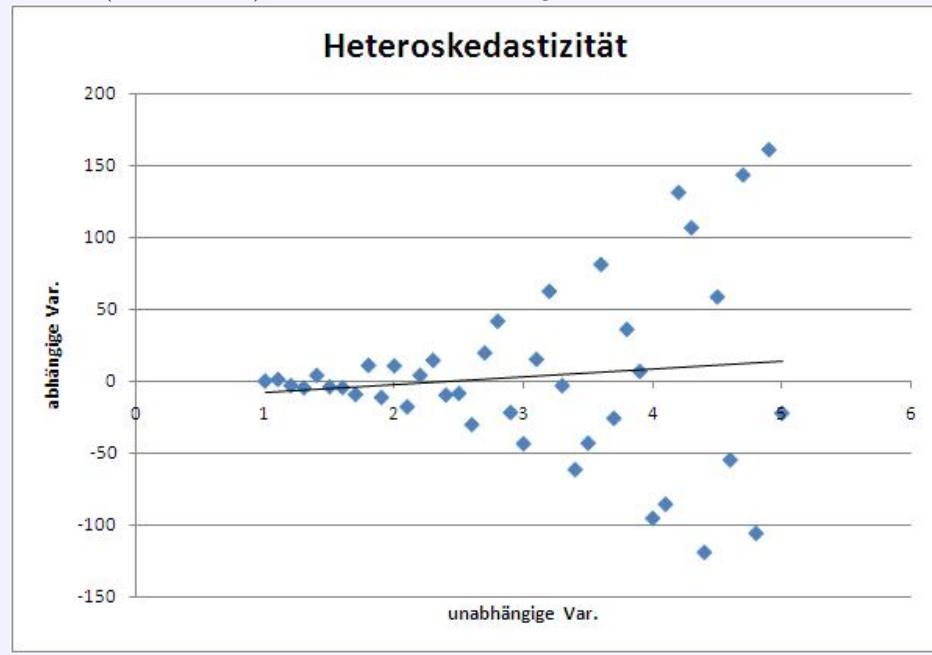


Visualisierung der Fehlervarianzen

Homoskedastizität bedeutet gleiche Varianz bei allen geschätzten Werten (fitted values).



Heteroskedastizität bedeutet unterschiedliche Varianz bei kleinen, mittleren und großen geschätzten Werten (fitted values), etwa ein 'trichterförmiges Auseinanderlaufen'.



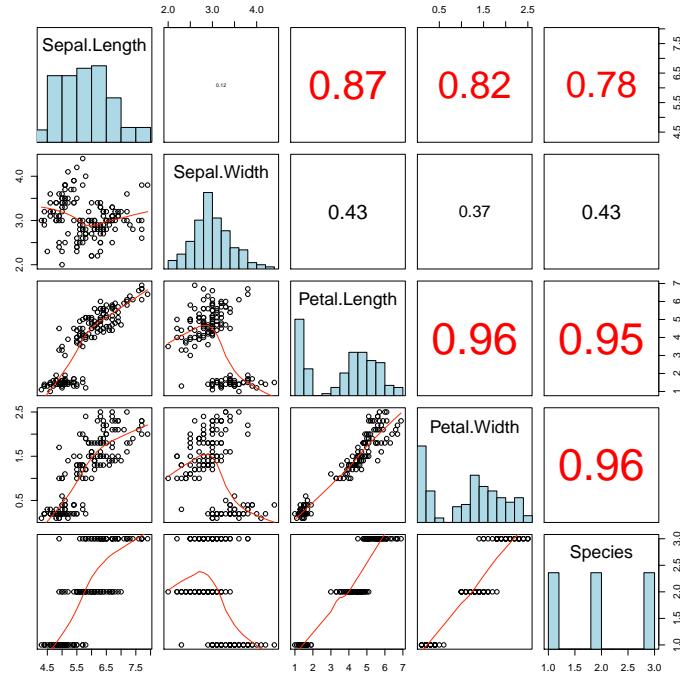
Zusätzlich existiert für multiple lineare Regressionmodelle mit mehr als einer erklärenden Variable x die Bedingung:

- (A5) Es gibt keine lin. Abhängigkeiten zwischen den Regressoren.

$$\text{rank}(X) = k$$

Diese wird VOR der Modellanpassung mithilfe einer pairwise Scatterplotmatrix geprüft. Alle hierbei im

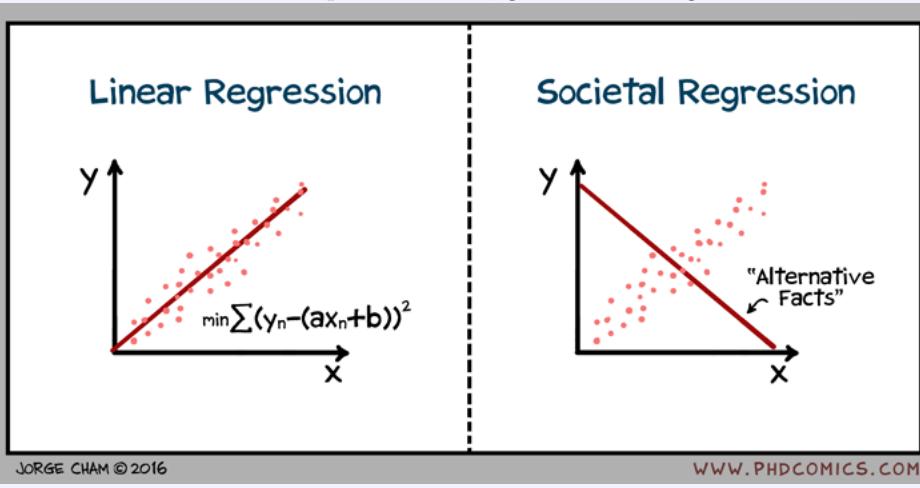
Vorfeld als korreliert erkannten Regressoren müssen selektiert werden, sodass nur eine solche Variable im Modell verbleibt.



Hebelwirkung = Leverage

Hebelpunkte (Engl. Leverage points) sind Beobachtungen, die sehr starken Einfluss auf den Verlauf der Gerade haben, da sie extreme oder ausreißende Werte in Bezug auf die unabhängige(n) Variable(n) X oder weit abseits der Punkte nahe der Regressionsfunktion liegen. Sowohl Steigung β als auch Intercept α sind davon betroffen.

Der "comichafte" Einfluss eines solchen Hebelpunkts wird durch die "social regression" wiedergegeben. Weshalb diese Punkte Hebelpunkte heißen, gibt die Grafik gut wieder: Sie "hebeln die Gerade aus".



Schätzung der Gerade und Konfidenzbereiche

Mithilfe des Konfidenzintervalls der einzelnen Regressionskoeffizienten erfolgt die Modellselektion, die sich auf p-Werte stützt.

Die Standardfehler = Breite des Konfidenzintervalls hängen von unterschiedlichen Faktoren ab:

- **Stichprobengröße:** Je mehr Beobachtungen gemacht wurden, desto schmäler ist das Konfidenzintervall.
- **Messfehler:** Je geringer der Messfehler der Daten, desto schmäler ist das Konfidenzintervall.
- Der Messfehler entspricht der Varianz der Residuen ε , die möglichst klein sein sollte, also dass die Punkte eng um die Gerade liegen.
- **x-Range:** Je breiter die Spannweite der unabhängigen Variablen, desto schmäler ist das Konfidenzintervall.

Daher sind "gute Heelpunkte", die oft weit weg von den anderen x-Werten liegen und dadurch die Spannweite vergrößern, zusätzlich wichtig für das Modell.

Vertrauensbereiche und Vorhersagebereiche der Regression

- μ Bei einem bekannten Wert der Lage μ lässt sich vor der Durchführung der Beobachtung für den arithmetischen Mittelwert \bar{x} vorhersagen, dass dieser mit Wahrscheinlichkeit 95% im Intervall

$$\left(\mu - 1,96 \cdot \frac{\text{standardabweichung}}{\sqrt{n}}; \mu + 1,96 \cdot \frac{\text{standardabweichung}}{\sqrt{n}} \right)$$

liegen wird, wenn man diese Standardabweichung und Stichprobengröße n vorher kennt.

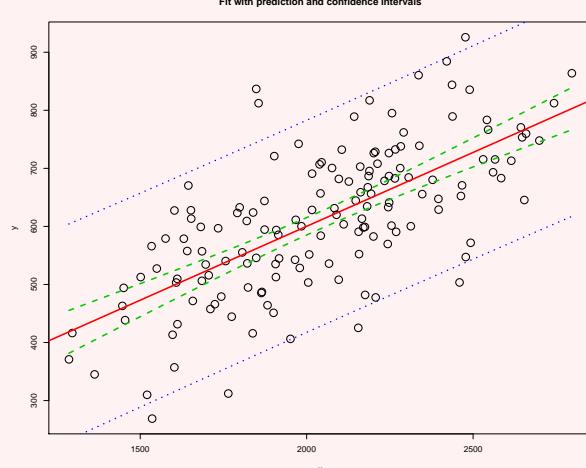
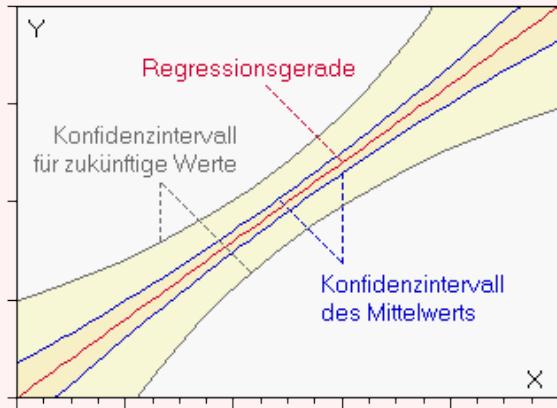
- p Bei einem bekannten Wert der Erfolgswahrscheinlichkeit p lässt sich vor der Durchführung der Beobachtung für die relative Häufigkeit \hat{p} vorhersagen, dass diese mit Wahrscheinlichkeit 95% im Intervall

$$\left(p - 1,96 \cdot \frac{p \cdot (1-p)}{\sqrt{n}}; p + 1,96 \cdot \frac{p \cdot (1-p)}{\sqrt{n}} \right)$$

liegen wird, wenn man die Stichprobengröße n vorher kennt.

- Der **Vorhersagebereich einer Prädiktion aus dem Modell** wird an einer Stelle x_i als Prädiktionsbereich mit dem Mittelwert gleich dem Wert der Regressionsgerade und mit Konfidenzniveau α berechnet. An jeder Stelle wird dieser Vorhersagebereich "punktweise" ermittelt und bezieht sich auf zukünftige y-Werte, die vorhergesagt werden.
- Der **Vertrauensbereich des Modells** wird so ermittelt, dass nicht nur an jeder Stelle, sondern über die gesamte Länge der Geraden hinweg ein Konfidenzniveau α gültig bleibt. Der Vertrauensbereich des Modells bezieht sich auf den Verlauf der Regressionsgeraden bzw. -(hyper)ebene und NICHT auf die y-Werte.

Vertrauensbereiche und Vorhersagebereiche der Regression



Die Umsetzung erfolgt durch den R Befehl

```
predict(linearesmodell,neuedaten,level = 1-alpha,interval = c('confidence', 'prediction'))
```

Modellzusammenfassung

```
Modellzusammenfassung
summary(linearesModell)

##
## Call:
## lm(formula = yreg ~ xreg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -8.8417 -1.9122 -0.0851  1.7403  8.3269 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.64956   0.44002  10.57   <2e-16 ***
## xreg        1.00158   0.01684  59.47   <2e-16 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 2.983 on 98 degrees of freedom
## Multiple R-squared:  0.973, Adjusted R-squared:  0.9728 
## F-statistic: 3537 on 1 and 98 DF, p-value: < 2.2e-16
Die Zusammenfassung enthält die wesentlichen Informationen und Modellschätzungen des linearen Regressionsmodells. Es beginnt mit der Anzeige des geschätzten Modells lm, yreg ~ xreg. Das ist die Kurzschreibweise für das Regressionsmodell
```

$$yreg_i = a + b \cdot xreg_i + \varepsilon_i$$

Als nächstes wird die 5-number-summary der Modellfehler, der Residuen ε_i , angezeigt.

```
summary(linmodsummary$residuals)

##      Min. 1st Qu. Median Mean 3rd Qu. Max. 
## -8.84167 -1.91219 -0.08511 0.00000 1.74033 8.32685
Das Modell wird stets so geschätzt, dass der Mittelwert der Residuen = 0 ist, wen ein Intercept in der Schätzung enthalten ist. Dadurch ist die Voraussetzung ( $A_1$ )  $\sum_{i=1}^n \varepsilon_i = 0$  automatisch erfüllt. Weitere Eigenschaften der Residuen können wir hier aus der Zusammenfassung nicht direkt ablesen. Dafür stehen uns die Residuenplots des linearen Regressionsmodells zur Verfügung, die wir separat besprechen, wenn es um die Überprüfung der Voraussetzungen des Regressionsmodells geht.
```

Modellzusammenfassung - Koeffizienten des Regressionsmodells

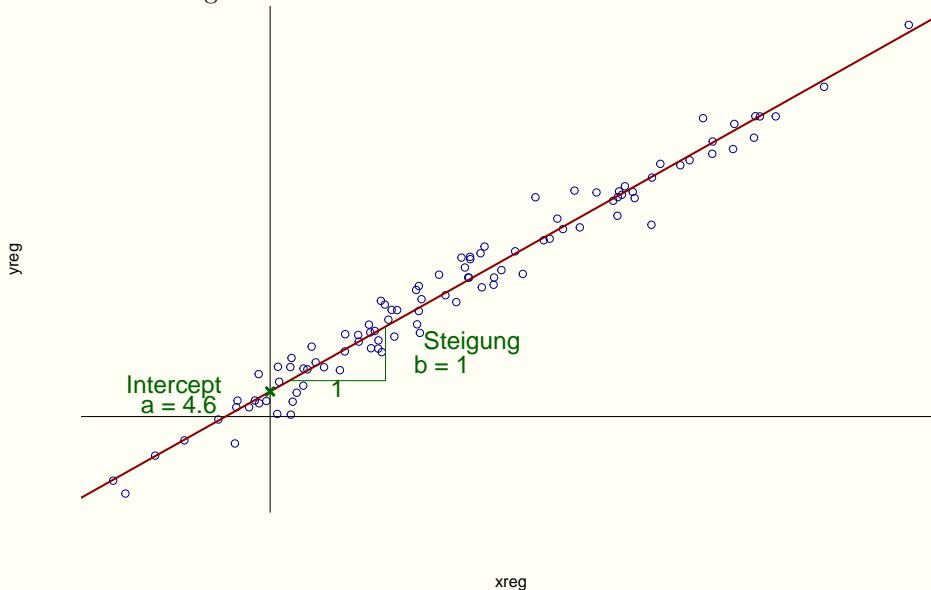
Als nächstes werden die Schätzungen der Modellkoeffizienten a = Intercept und b = Steigung der Variablen `xreg`, angezeigt.

```
print(linmodsummary$coefficients)
##             Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) 4.649559 0.44001996 10.56670 7.154942e-18
## xreg        1.001583 0.01684114 59.47236 1.039006e-78
```

Hieraus können wir direkt die geschätzte Regressionsgeradengleichung ablesen:

$$y_{\text{reg}} = 4.6495588 + 1.0015825 \cdot x_{\text{reg}}$$

Graphisch bedeutet das folgendes Modell für unsere Daten:



Dabei gilt, dass das Intercept einen mittleren Wert $\hat{a} = 4.6495588$ von y für $x=0$ hat. Um diesen mittleren geschätzten Wert lässt sich das Konfidenzintervall mithilfe von $\hat{a} \pm t_{1-\frac{\alpha}{2}; df=n-1} \cdot \text{Standard Error}$ bauen, wobei das entsprechende Quantil der student's t Verteilung mit $n-1$ Freiheitsgraden eingesetzt wird, wie wir das von der Konstruktion der Konfidenzintervalle für Mittelwerte kennen.

Das 95%-Konfidenzintervall bei einer Stichprobengröße von $n=100$ für den Interceptwert ist also $[4.6495588 - 1.984217 \cdot 0.44002; 4.6495588 + 1.984217 \cdot 0.44002] = [3.7764638; 5.5226539]$.

Die Steigung der Regressionsgerade in Richtung der unabhängigen Variable `xreg` hat einen mittleren Wert $\hat{b} = 1.0015825$ hat, y steigt also um 1.0015825 Einheiten, wenn x um 1 Einheit zunimmt. Um diesen mittleren geschätzten Wert lässt sich das Konfidenzintervall mithilfe von $\hat{b} \pm t_{1-\frac{\alpha}{2}; df=n-1} \cdot \text{Standard Error}$ bauen, wobei das entsprechende Quantil der student's t Verteilung mit $n-1$ Freiheitsgraden eingesetzt wird, wie wir das von der Konstruktion der Konfidenzintervalle für Mittelwerte kennen.

Das 95%-Konfidenzintervall bei einer Stichprobengröße von $n=100$ für die Steigung ist also $[1.0015825 - 1.984217 \cdot 0.0168411; 1.0015825 + 1.984217 \cdot 0.0168411] = [0.968166; 1.034999]$.

```

Modellzusammenfassung - Hypothesentests zur Modellselektion
##
## Call:
## lm(formula = yreg ~ xreg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.8417 -1.9122 -0.0851  1.7403  8.3269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.64956   0.44002 10.57   <2e-16 ***
## xreg        1.00158   0.01684 59.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.983 on 98 degrees of freedom
## Multiple R-squared:  0.973, Adjusted R-squared:  0.9728
## F-statistic: 3537 on 1 and 98 DF, p-value: < 2.2e-16

```

Wir erinnern uns auch daran, dass die Konstruktion von Vertrauensbereichen und Zufallsstreibereichen eng mit Hypothesentesten zusammenhängt. Hier wird jeweils ein t-Test zur Modellselektion der Parameterwerte durchgeführt. Jeder t-Test hat als Nullhypothese H_0 : Parameter = 0 und als Alternativhypothese H_A : Parameter $\neq 0$. Damit diese t-Tests durchgeführt werden können, ist die Normalverteilung der Residuen erforderlich.

Für das Intercept wird also der Test $H_0 : a = 0$ gegen $H_A : a \neq 0$ durchgeführt. Die dazu gehörende Teststatistik lautet $t = 10.5666999$ und der p-Wert dieses Tests beträgt $p=7.1549418 \times 10^{-18}$, wobei er in der 4. Spalte mit seiner ursprünglichen Definition “die Wahrscheinlichkeit, unter dem Nullhypothesenszenario einen solchen Wert oder einen noch extremeren zu beobachten” angeschrieben ist, was der Irrtumswahrscheinlichkeit, wenn wir die Nullhypothese verwerfen, entspricht. Das bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau verworfen wird, auf dem 1% Signifikanzniveau verworfen wird und auf dem 0.1% Signifikanzniveau verworfen wird.

Für die Steigung wird entsprechend der Test $H_0 : b = 0$ gegen $H_A : b \neq 0$ durchgeführt. Die dazu gehörende Teststatistik lautet $t = 59.4723622$ und der p-Wert dieses Tests beträgt $p=1.0390065 \times 10^{-78}$, wobei er in der 4. Spalte mit seiner ursprünglichen Definition “die Wahrscheinlichkeit, unter dem Nullhypothesenszenario einen solchen Wert oder einen noch extremeren zu beobachten” angeschrieben ist, was der Irrtumswahrscheinlichkeit, wenn wir die Nullhypothese verwerfen, entspricht. Das bedeutet, dass die Nullhypothese auf dem 5% Signifikanzniveau verworfen wird, auf dem 1% Signifikanzniveau verworfen wird und auf dem 0.1% Signifikanzniveau verworfen wird.

Erweiterte Aspekte: Modellselektion

Üblicherweise sind nicht alle erklärende Variablen, die in einem multiplen Regressionsmodell zur Modellierung verwendet werden, auch tatsächlich relevant bzw. signifikant an der Modellierung der erklärten Variable beteiligt. Ziel ist also von allen k Regressoren nur eine Teilmenge zu bestimmen, die die tatsächlich relevanten Variablen enthält. Dafür gibt es unterschiedliche Herangehensweisen:

- Jeder Modellkoeffizient wird unabhängig von den anderen einem **t-Test** unterzogen.

Dabei wird jeweils der Nullhypothese $H_0 : \beta_i = 0$, also dass die erklärende Variable x_i keinen linearen Zusammenhang mit der erklären Variable y hat, die Alternativhypothese $H_A : \beta_i \neq 0$ gegenübergestellt. Anhand des p-Werts wird dann eine Entscheidung getroffen. Dieser Student's t Test ist automatisch Teil des R summary Outputs.

Der Nachteil dieser Methode ist, dass der Einfluss des Entfernens der am wenigsten relevanten Variable(n) nicht ersichtlich ist, bevor ein neues Modell ohne die Variable(n) angepasst wurde. Auch der Zusammenhang zwischen Variablen wird bei der separaten Modellselektion nicht berücksichtigt. Das ist ein besonderes Problem, wenn nicht vor der Modellanpassung überprüft wurde, dass die erklärenden Variablen von einander unabhängig sind. Dabei kann es zu "Maskierungseffekten" kommen, bei denen beide Variablen als nicht signifikant relevant für das Modell evaluiert werden. Nach Entfernen einer der beiden hochkorrelierten Regressoren wird der andere dann als hochsignifikant eingestuft, was zuvor vom korrelierten Regressor maskiert wurde.

- **Information Criteria** erlauben eine schrittweise Modellselektion anhand von Modellkriterien, die dabei optimiert werden sollen.

- Akaike Information Criterion (AIC)

$$AIC = -2 \ln(\hat{L}(\theta)) + 2k$$

Because this is the negative log-likelihood function penalised by the number of model parameters q minimising the AIC is equivalent to maximising the likelihood function which is the statistical basis for construction of many tests and models.

For linear regression AIC is equivalent to Mallows's C_p .

- Bayesian Information Criterion (BIC)

$$BIC = -2 \ln(\hat{L}(\theta)) + \ln(n)k$$

Assumes a prior probability of each candidate model $\frac{1}{\#\text{candidate models}}$ which is combined with the likelihood function which comprises the information of the data. BIC puts a larger penalty on model size (for $n > 7$) than AIC and thus selects smaller models.

Beispiel für Modellselektion

Die Modellselektion mithilfe von t-Tests haben wir bereits bei der Besprechung des Regressionsoutputs in R besprochen. Sie automatisch in der R summary mitgeliefert.

```
slm<-summary(lm(Education~.-Examination,data=swiss))
(slms$coefficients)

##              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) 49.99303171 6.18640648 8.0811101 4.312200e-10
## Fertility    -0.52070092 0.07868790 -6.6172931 5.139985e-08
## Agriculture  -0.22879685 0.03906287 -5.8571441 6.374386e-07
## Catholic      0.08333381 0.02178681  3.8249658 4.275083e-04
## Infant.Mortality 0.28436994 0.30040325  0.9466274 3.492434e-01
```

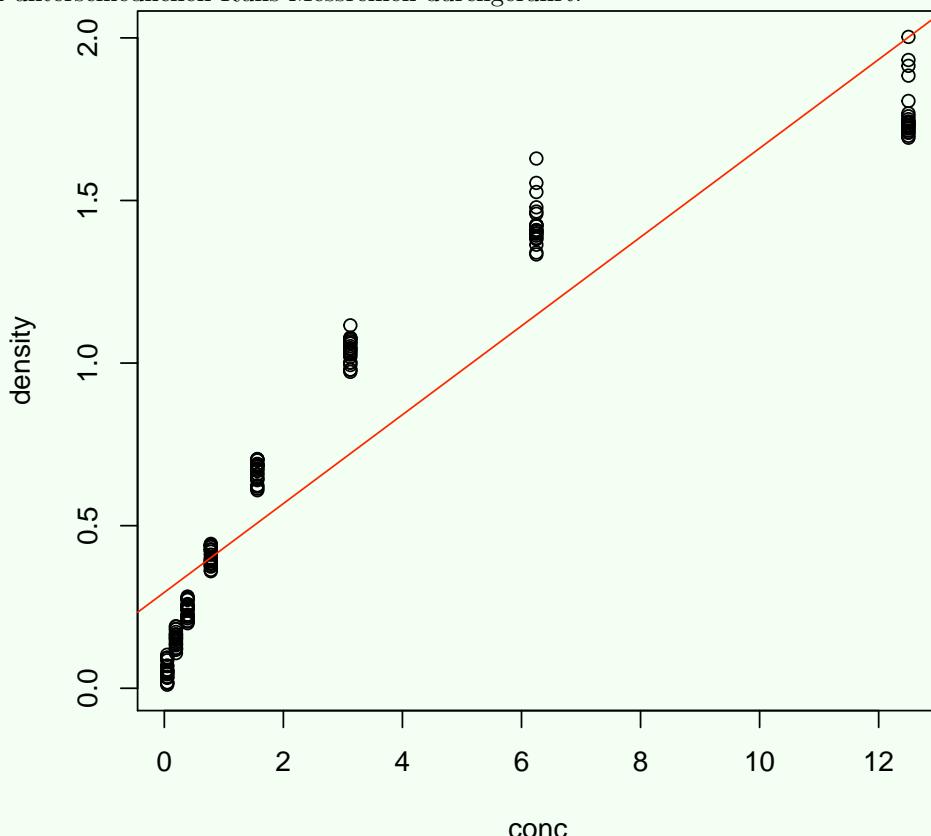
Information Criteria basierte schrittweise (stepwise) Modellselektion erfolgt mithilfe der Funktion step.

```
step(lm(Education~.-Examination,data=swiss))

## Start: AIC=160.13
## Education ~ (Fertility + Agriculture + Examination + Catholic +
##               Infant.Mortality) - Examination
##
##              Df Sum of Sq    RSS    AIC
## - Infant.Mortality 1     24.46 1170.8 159.12
## <none>                      1146.3 160.13
## - Catholic          1     399.32 1545.7 172.17
## - Agriculture        1     936.34 2082.7 186.19
## - Fertility          1    1195.15 2341.5 191.69
##
## Step: AIC=159.12
## Education ~ Fertility + Agriculture + Catholic
##
##              Df Sum of Sq    RSS    AIC
## <none>                      1170.8 159.12
## - Catholic          1     410.71 1581.5 171.25
## - Agriculture        1    1079.89 2250.7 187.84
## - Fertility          1    1289.36 2460.2 192.02
##
## Call:
## lm(formula = Education ~ Fertility + Agriculture + Catholic,
##     data = swiss)
##
## Coefficients:
## (Intercept)    Fertility  Agriculture    Catholic
##      53.8505     -0.4888     -0.2380      0.0844
```

Beispiele zur Regression in R

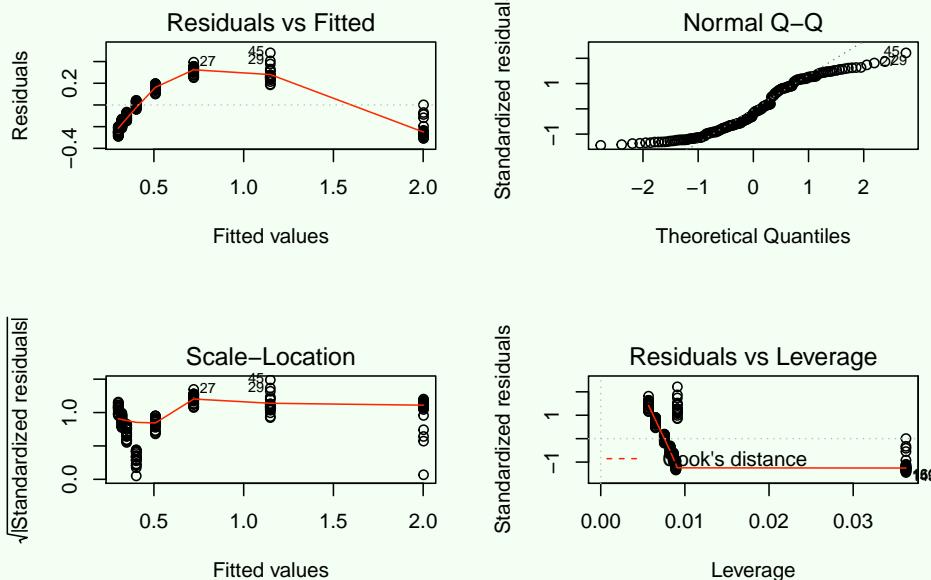
Beispiel: Bei der Bestimmung der optischen Dichte für verschiedene Konzentrationen von DNase werden in unterschiedlichen Runs Messreihen durchgeführt.



Wir erkennen zwar schon optisch, dass hier ein Abflachen der optischen Dichte eintritt, passen aber im ersten Schritt dennoch eine Gerade an.

```
linearesModell<-lm(density~conc,data = DNase)
summary(linearesModell)

##
## Call:
## lm(formula = density ~ conc, data = DNase)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.30901 -0.19640 -0.03957  0.19498  0.48056 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.29488   0.02072 14.23 <2e-16 ***
## conc        0.13657   0.00406 33.63 <2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2181 on 174 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8659 
## F-statistic: 1131 on 1 and 174 DF,  p-value: < 2.2e-16
```



Probleme mit Modell

- Heteroskedastische Fehler - Schwankungen wird größer (A2 nicht erfüllt)
- ”Batch-Effekt“ Punkt zu Gruppen zusammenclustern (A3 nicht erfüllt)
- Systematischer Bogen oberhalb und unterhalb der 0-Linie im 1. Plot (A3 nicht erfüllt)
- 2. Modus in den Fehlern - großen Fehlerwerten (A4 nicht erfüllt)
- keine Punkt ist Hebelpunkt
- eigentlich ist hier die Gerade nicht die bestmögliche Anpassung. Eine nichtlineare Funktion würde besser passen.

Modell passt gut ($R^2 = 0.8667001$), aber nicht gültig, weil Voraussetzungen nicht erfüllt sind!

Erweiterte Konzepte: Transformationen

Wenn Variablen nicht direkt miteinander linear zusammenhängen, gibt es die Möglichkeit, diese zu transformieren und einen linearen Zusammenhang zwischen transformierten Variablen zu bekommen. Die wichtigsten dieser Transformationen sind:

- **Lineare Transformationen**

Lineare Datentransformationen bedeuten

- **Translation = Verschiebung** der Daten um einen konstanten Wert k

$$T_i = X_i + k$$

Diese Transformation geschieht oft (aber in sehr geringem Maß), wenn dieselbe Probe auf unterschiedliche geeichten Messgeräten gemessen wird.

- **Skalierung = Streckung/Stauchung** der Daten um einen Streckungs- bzw. Stauchungsfaktor s

$$S_i = s \cdot X_i$$

Diese Transformation geschieht so gut wie immer, wenn zwischen Maßeinheiten l zu mm^3 , g zu kg umgerechnet wird.

- Eine spezielle Kombination von Translation und Skalierung, ist die **Normalisierung** oder **Standardisierung** von Daten

$$Z_i = (X_i - \bar{X})/sd(X)$$

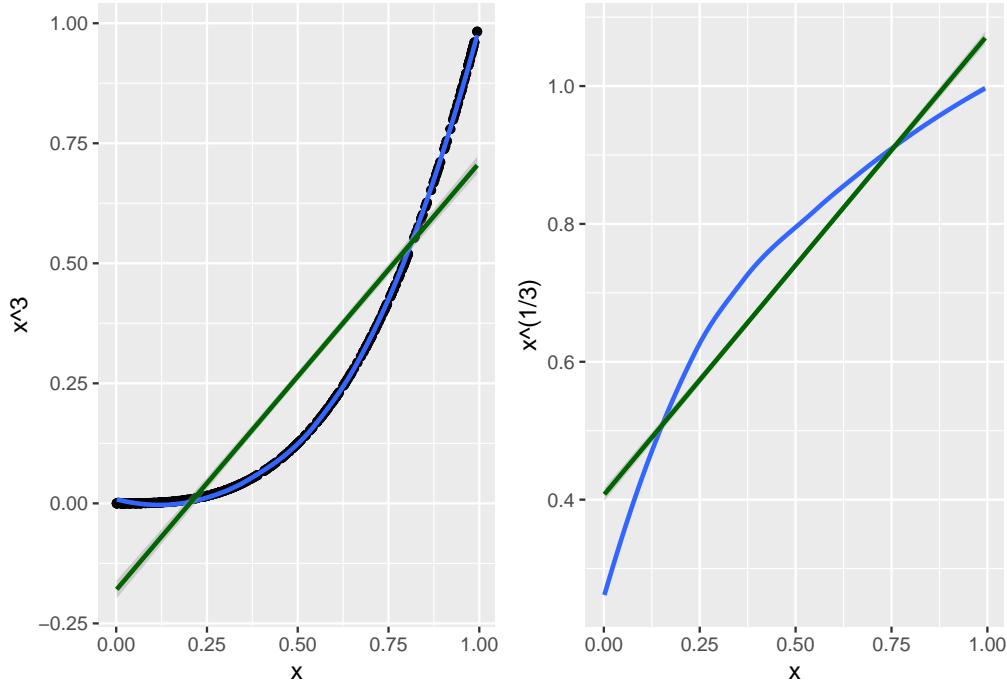
Durch diese Transformation haben standardisierte Daten Z_i immer Mittelwert 0 und Standardabweichung 1. Für bestimmte Algorithmen und Verfahren ist diese Standardisierung eine notwendige Vorbedingung

- **Potenzen und Wurzeln** (engl. Power-Transformation)

Transformationen vom Typus

$$P_i = X_i^k$$

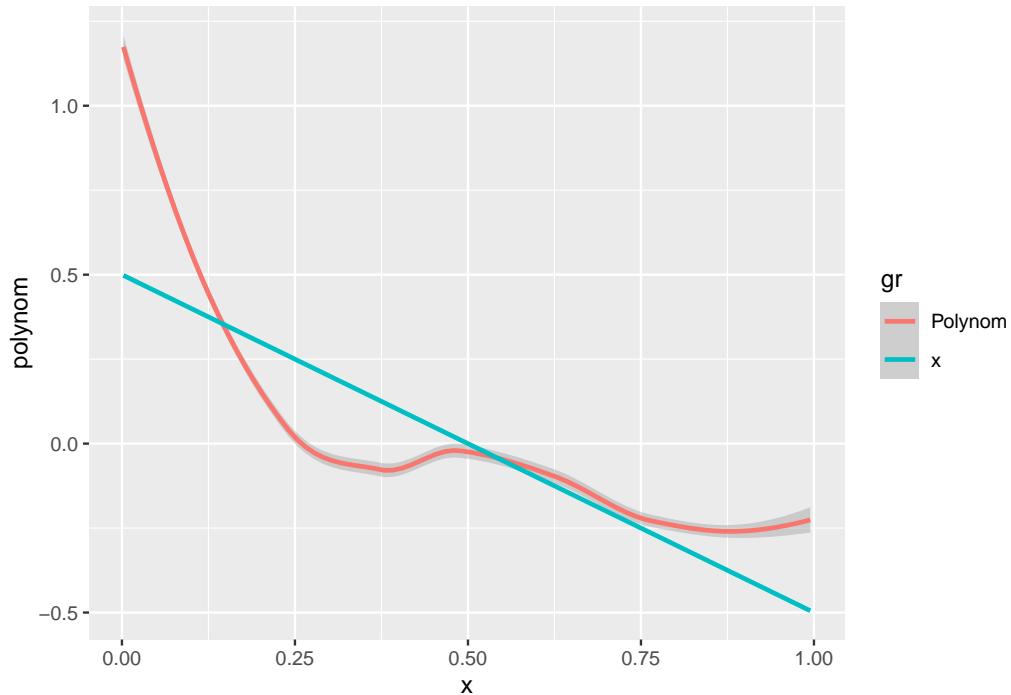
heißen Power-Transformation, wenn eine bestimmte Potenz für alle Werte angewendet wird. Hier ist Vorsicht geboten, da etwa geradzahlige Wurzeln nicht für negative Zahlenwerte anwendbar sind.



Zu den wichtigsten Transformationen dieses Typus zählen die *Quadratur* $Q_i = X_i^2$ und *Kubizierung* $K_i = X_i^3$ und die *Quadratwurzel* $QW_i = \sqrt{X_i}$ und *Kubikwurzel* $KW_i = \sqrt[3]{X_i}$.

- Polynome

Polynome $y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \dots + a_n \cdot x^n$ sind mathematisch flexibel genug, um beinahe jede Funktion wenigstens in einem Teilbereich hinreichend gut annähern zu können, was etwa in der Mathematik das Prinzip hinter Taylorreihenschätzung ist.



- **Logarithmus und Exponentialfunktion**

Die **Exponentialfunktion** mit natürlicher Basis (Eulerschen Zahl e) oder dekadischer Basis (10)

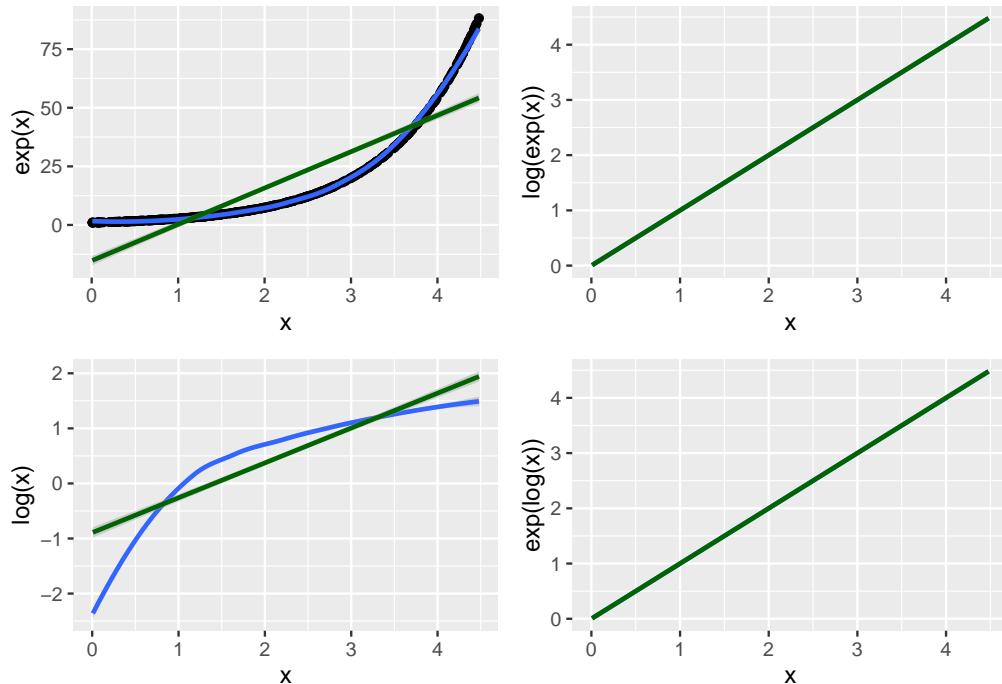
$$E_i = \exp(X_i) \text{ oder } D_i = 10^{X_i}$$

kommt in vielen natürlichen Prozessen vor. Sollten Daten vom gegenteiligen logarithmischen Verlauf sein, weil sie so erhoben oder durch Preprocessing transformiert wurden, dann können Sie mithilfe der Exponentialfunktion zurücktransformiert werden.

Öfter sind die Daten aber aus einem exponentiell verlaufenden Prozess stammen und müssen mithilfe der Logarithmustransformation

$$L_i = \log(X_i)$$

zu einem linearen Verlauf transformiert werden.



Wir modellieren durch Logarithmus-Transformation das **exponentielle Modell**

$$y(t) = a \cdot b^t = a \cdot e^{\lambda \cdot t}$$

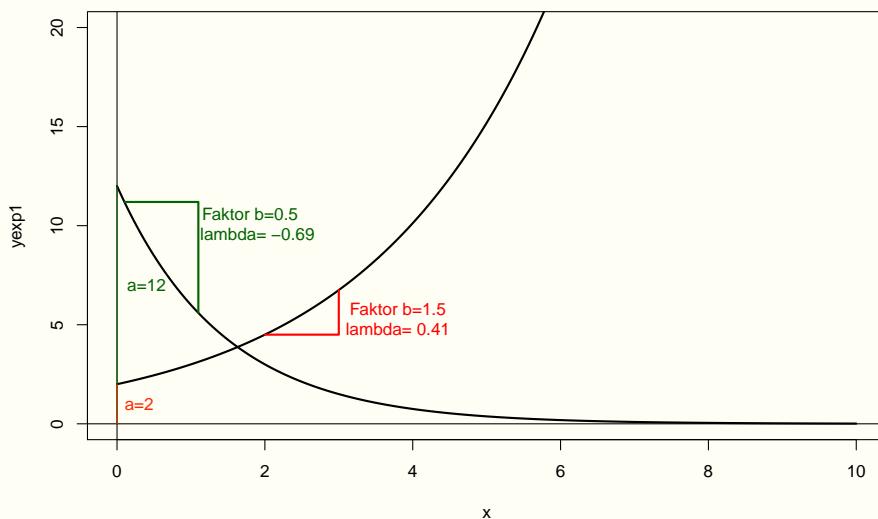
mit der Parameterbedeutung

$$y(t) = a \cdot \underbrace{b}_\text{Wachstumsfaktor}^t = a \cdot e^{\lambda \cdot t}$$

durch das lineare Modell

$$\log(y(t)) = \log(a) + \log(b) \cdot t = \log(a) + \lambda \cdot t$$

Exponentielles Wachstum und Zerfall



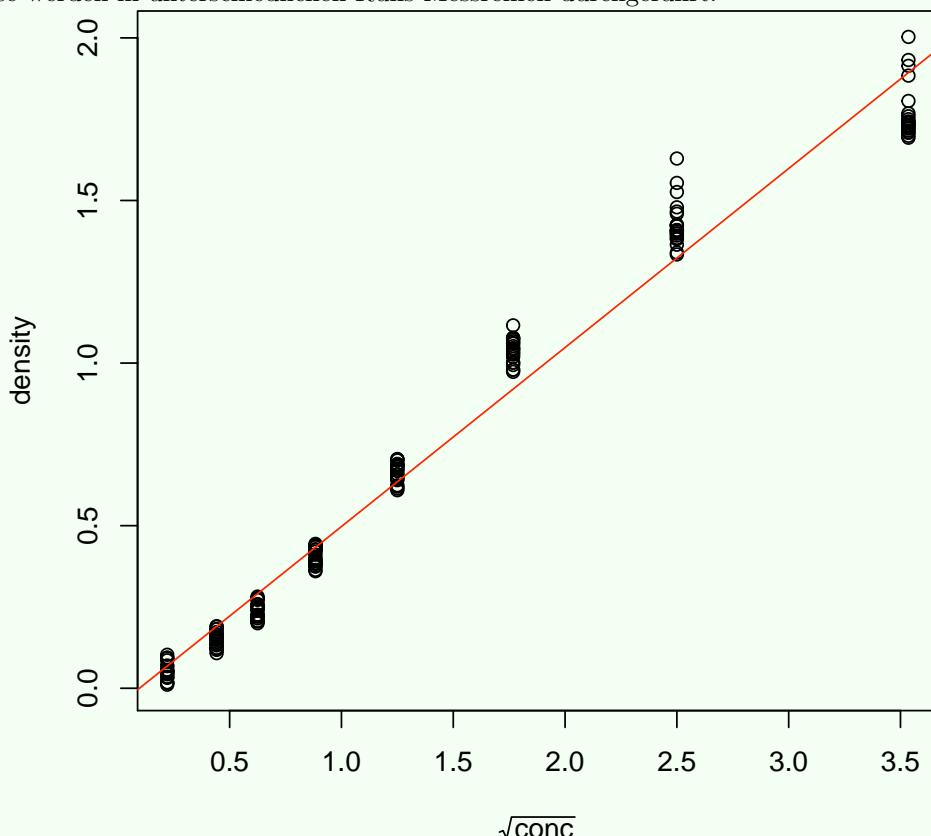
Dazu gehört als Interpretation der Modellparameter:

a ... $y(0)=a$... Startwert von y zum Zeitpunkt $t=0$ (mit Einheit!)

b ... $t=1$ wird mit b multipliziert ... in einem Zeitschritt wächst y um $(1-b) \cdot 100\%$ ($b > 1$ Wachstum, $b < 1$ Zerfall) (keine Einheit)

λ ... konstante Wachstum pro Zeiteinheit (Einheit = 1/Zeiteinheit)

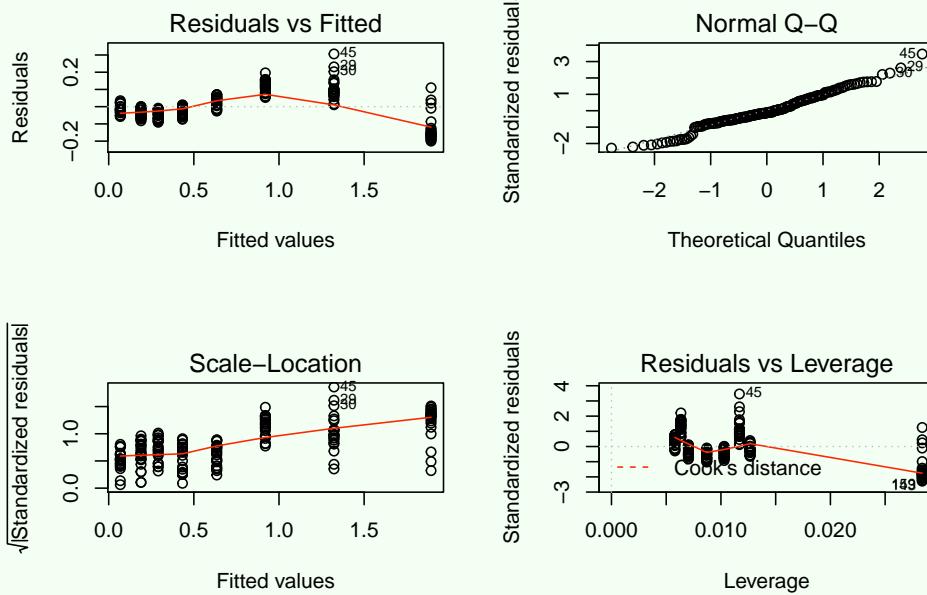
Rückkehr zum **Beispiel**: Bei der Bestimmung der optischen Dichte für verschiedene Konzentrationen von DNase werden in unterschiedlichen Runs Messreihen durchgeführt.



Wir wissen von zuvor, dass hier ein Abflachen der optischen Dichte eintritt, und eine lineare Funktion nicht direkt passt. Nun wurde auf der x-Achse die Wurzel aus den Konzentrationswerten gezogen und das Modell dafür angepasst. Das Ziehen der Quadratwurzel wird mit der mathematischen Potenz $(\cdot)^{\frac{1}{2}}$ durchgeführt.

```
linearesModell<-lm(density~I((conc)^(0.5)), data = DNase)
summary(linearesModell)

##
## Call:
## lm(formula = density ~ I((conc)^(0.5)), data = DNase)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.20009 -0.05060 -0.01110  0.05827  0.30600 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.053297  0.011081  -4.81 3.26e-06 ***
## I((conc)^(0.5)) 0.550521  0.006287  87.57 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08897 on 174 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9777 
## F-statistic: 7668 on 1 and 174 DF,  p-value: < 2.2e-16
```



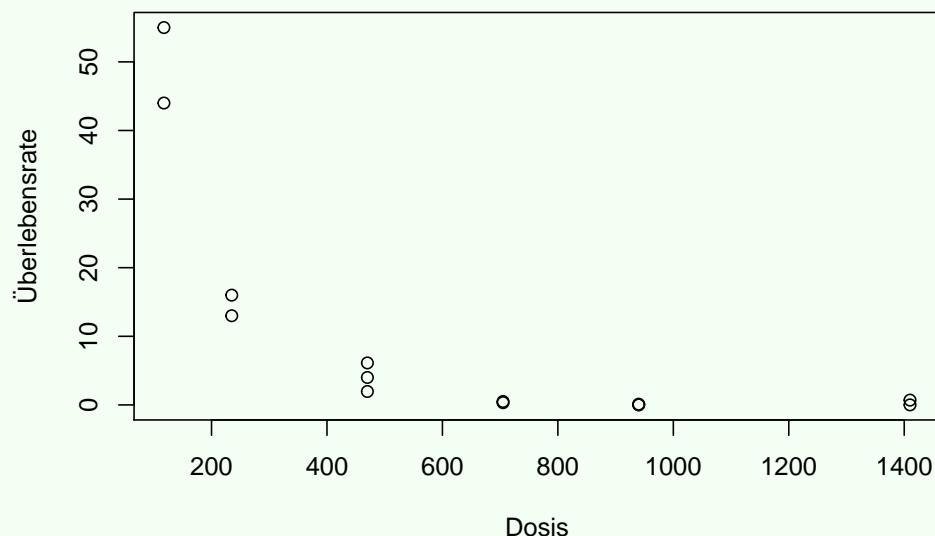
Probleme mit Modell und was wir verbessern konnten

- Heteroskedastische Fehler - Schwankungen werden größer, aber weniger deutlich als zuvor (A2 nicht erfüllt)
- "Batch-Effekt" Punkt zu Gruppen zusammenclustern (A3 nicht erfüllt)
- Der Bogen um die 0-Linie im 1. Plot ist verschwunden
- Normalverteilung ist nun annähernd erreicht
- kein Punkt ist Heelpunkt
- Die Wurzelfunktion passt hier gut zu den Daten, hat aber immer noch manche Probleme nicht ausgleichen können.

Modell passt sehr gut ($R^2 = 0.9778113$), aber nicht gültig, weil Voraussetzungen nicht erfüllt sind!

Beispiel: Die Überlebensdauern von Bakterien abhängig von der Dosierung eines Antibiotikums.

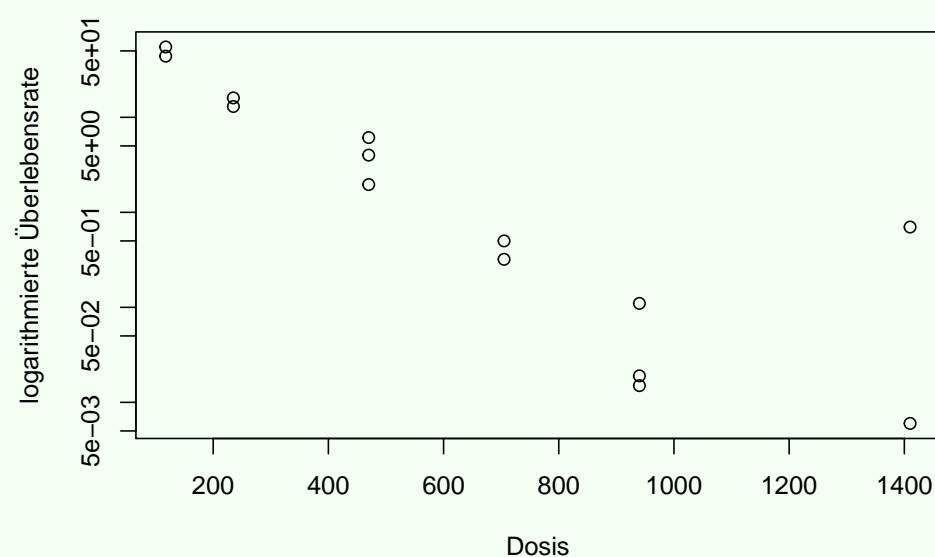
```
library(boot); data(survival)
with(survival, plot(dose, surv, main="Überleben von Bakterien", xlab="Dosis", ylab="Überlebensrate"))
Überleben von Bakterien
```



Log-transformation

Wir werden hier die y-Werte log-transformieren.

```
with(survival, plot(dose, surv, log="y", main="Überleben von Bakterien", xlab="Dosis", ylab="logarithmierte Überlebensrate"))
Überleben von Bakterien
```



```

fit <- lm(log(surv) ~ dose, data=survival)
summary(fit)

##
## Call:
## lm(formula = log(surv) ~ dose, data = survival)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.4637 -0.5679 -0.1079  0.5772  4.1592 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.823648   0.811788   4.710 0.000505 ***
## dose        -0.005915   0.001047  -5.651 0.000107 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 1.629 on 12 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.7041 
## F-statistic: 31.93 on 1 and 12 DF,  p-value: 0.0001071

```

Lineare Modell für Logarithmen

$$\log(\text{surv}) = \underbrace{3.82}_{\text{mittlere logarithmierte Überlebensdauer surv ist, wenn Dosis=0 ist}} - 0.006 * \text{dose}$$

Wenn die Dosis um 1 Einheit steigt, fällt logarithmierte Überlebensdauer um 0.006 Einheiten.

Exponentielles Modell für Originaldaten

$$\text{surv} = \exp(3.82 - 0.006 * \text{dose}) = \exp(3.82) * \exp\left(\underbrace{-0.006}_{Wachstumskonstante} * \text{dose}\right)$$

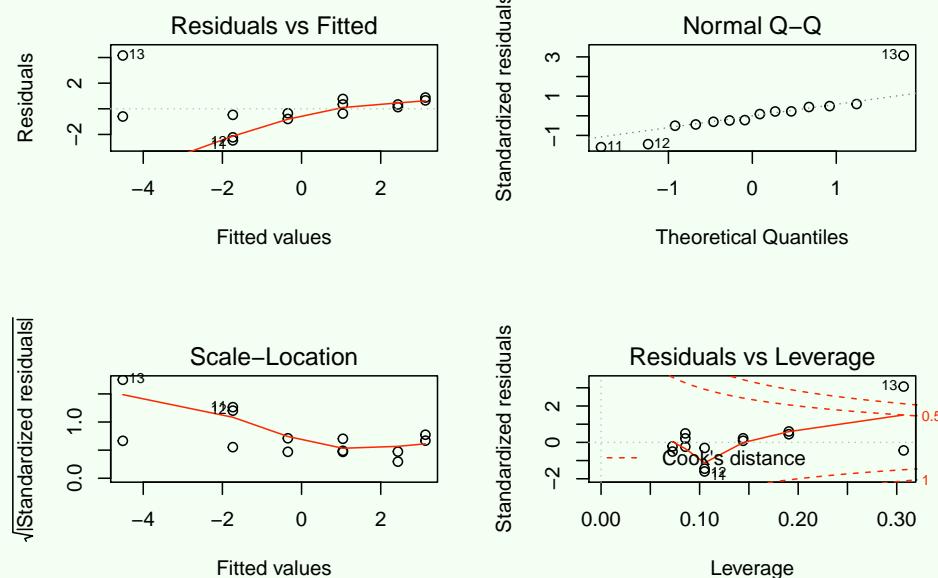
$$\text{surv} = \exp(3.82) * \exp(-0.006)^{\text{dose}} = 45.60 * \underbrace{0.994018}_{Wachstumsfaktor}^{\text{dose}}$$

Überlebensdauer von 45.6 Einheiten, wenn Dosis = 0.

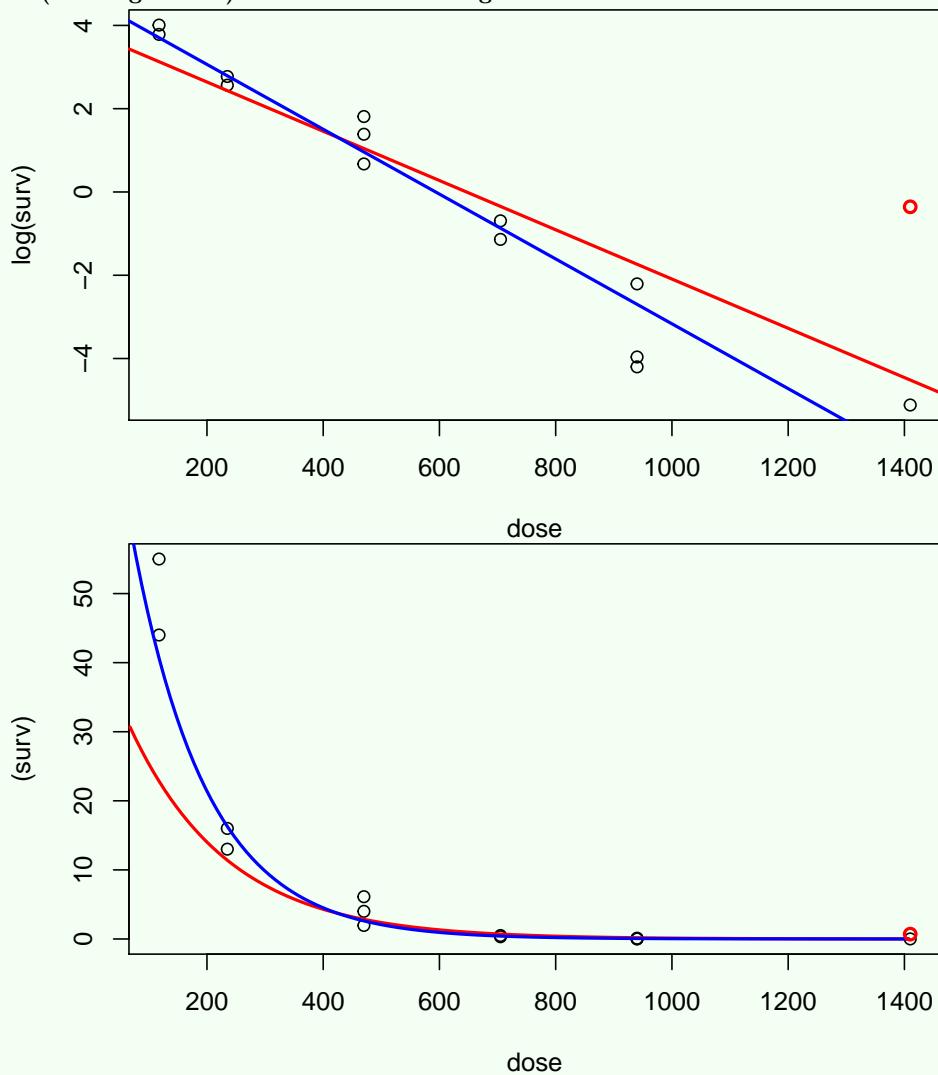
Wenn die Dosis um 1 Einheit steigt, wird Überlebensdauer um 0.6 % fallen

Residual Diagnostic Plots

```
par(mfrow=c(2,2)); plot(fit); par(mfrow=c(1,1))
```



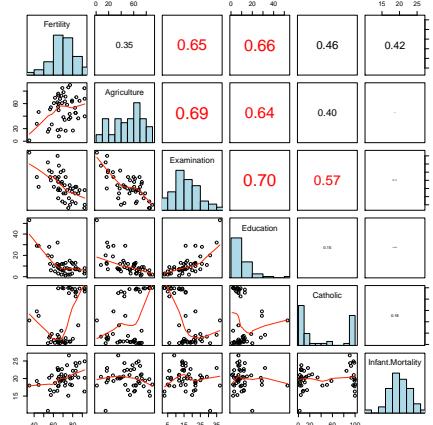
Heelpunkt (Leverage Point) und ihre Auswirkung auf das Modell



Multiples Regressionsmodell - Beispiel Schweiz

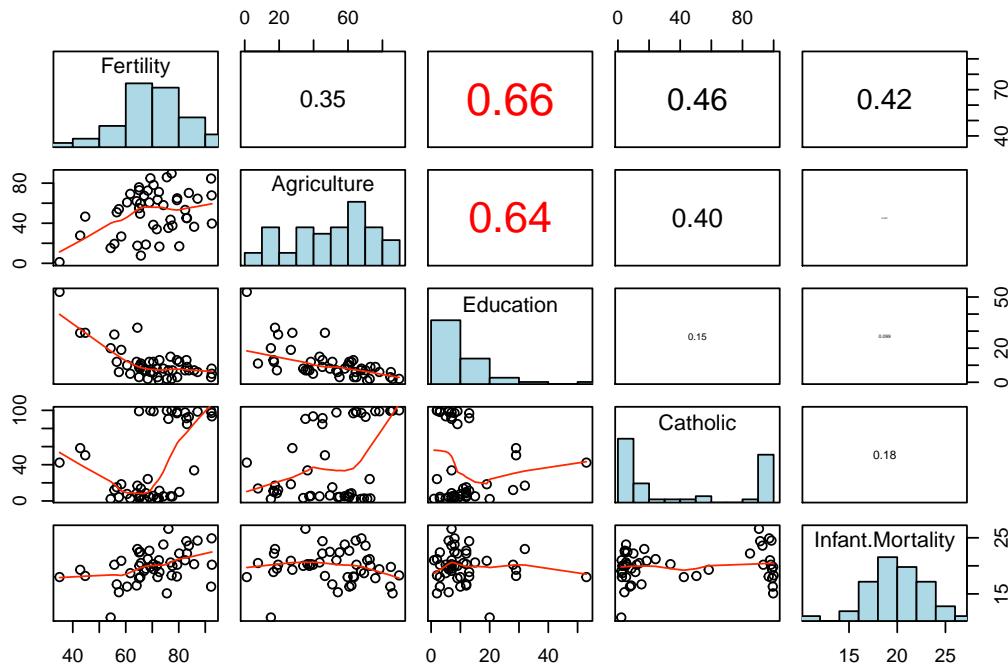
```
summary(swiss)
```

```
##      Fertility      Agriculture      Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70  1st Qu.:35.90  1st Qu.:12.00  1st Qu.: 6.00
##  Median :70.40  Median :54.10  Median :16.00  Median : 8.00
##  Mean   :70.14  Mean   :50.66  Mean   :16.49  Mean   :10.98
##  3rd Qu.:78.45  3rd Qu.:67.65  3rd Qu.:22.00  3rd Qu.:12.00
##  Max.   :92.50  Max.   :89.70  Max.   :37.00  Max.   :53.00
##      Catholic      Infant.Mortality
##  Min.   : 2.150   Min.   :10.80
##  1st Qu.: 5.195   1st Qu.:18.15
##  Median :15.140   Median :20.00
##  Mean   :41.144   Mean   :19.94
##  3rd Qu.:93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```



Filtern von Examination

```
pairs(swiss[,-3], lower.panel = panel.smooth, diag.panel = panel.hist, upper.panel = panel.cor)
```



Modellanpassung

```
##
## Call:
## lm(formula = Education ~ . - Examination, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.4029 -2.7803 -0.7571  2.4934 12.8590 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 49.99303  6.18641  8.081 4.31e-10 ***
## Fertility   -0.52070  0.07869 -6.617 5.14e-08 ***
## Agriculture -0.22880  0.03906 -5.857 6.37e-07 ***
## Catholic     0.08333  0.02179  3.825 0.000428 ***
## Infant.Mortality 0.28437  0.30040  0.947 0.349243  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.224 on 42 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7048 
## F-statistic: 28.46 on 4 and 42 DF,  p-value: 1.804e-11
```

Modellselektion

```
##
## Call:
## lm(formula = Education ~ . - Examination - Infant.Mortality,
##      data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.0852 -2.9521 -0.6678  3.2519 12.9706 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 53.85051  4.64907 11.583 8.30e-15 ***
## Fertility   -0.48883  0.07104 -6.881 1.91e-08 ***
## Agriculture -0.23799  0.03779 -6.298 1.35e-07 ***
## Catholic     0.08440  0.02173  3.884 0.00035 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

## 
## Residual standard error: 5.218 on 43 degrees of freedom
## Multiple R-squared:  0.7247, Adjusted R-squared:  0.7055
## F-statistic: 37.73 on 3 and 43 DF,  p-value: 4.123e-12

```

Modellgleichungen im Vergleich

Variante 1

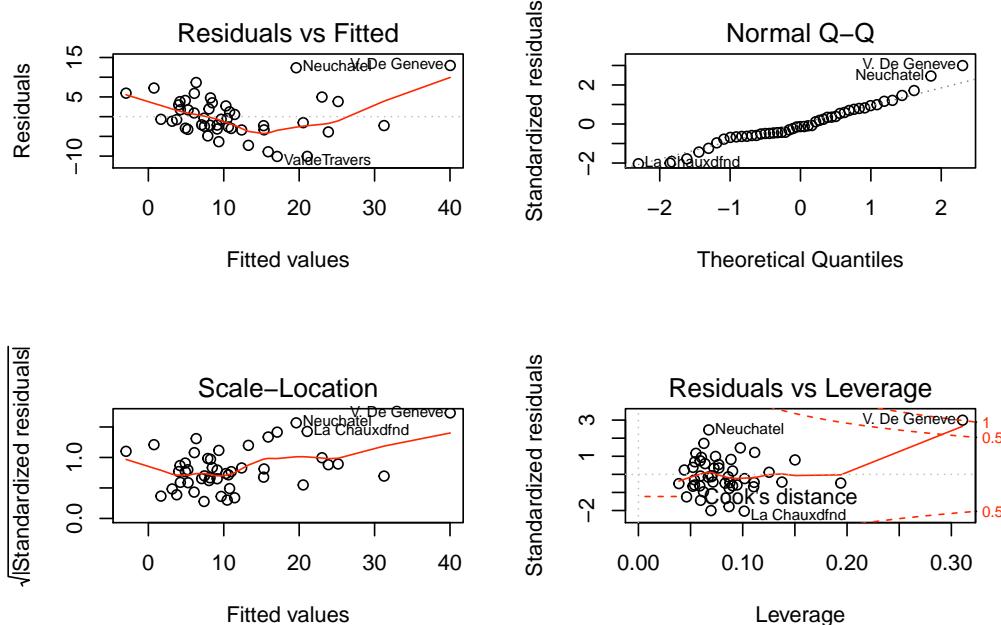
$$\text{Education} = 49.99 + (-0.52) \cdot \text{Fertility} + (-0.23) \cdot \text{Agriculture} + 0.08 \cdot \text{Catholic} + 0.28 \cdot \text{Infant.Mortality}$$

Variante 2 ohne Infant Mortality

$$\text{Education} = 53.85 + (-0.49) \cdot \text{Fertility} + (-0.24) \cdot \text{Agriculture} + 0.08 \cdot \text{Catholic}$$

gute Modellanpassung bei R^2 72.47%

Residual Quality Plots



Hebelpunkt erkennen und entfernen

```

## 
## Call:
## lm(formula = Education ~ . - Examination - Infant.Mortality,
##      data = swiss[-45, ])
## 
## Residuals:
##    Min      1Q   Median      3Q     Max 
## -9.0144 -2.4407 -0.7688  2.6409 14.0202 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 45.26863   4.91660   9.207 1.24e-11 ***
## Fertility   -0.38468   0.07120  -5.403 2.86e-06 ***
## Agriculture -0.20240   0.03566  -5.676 1.16e-06 ***
## Catholic     0.06188   0.02070   2.989  0.00466 ** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.697 on 42 degrees of freedom
## Multiple R-squared:  0.6216, Adjusted R-squared:  0.5945 
## F-statistic: 22.99 on 3 and 42 DF,  p-value: 5.776e-09

```

Variante 1

$$\text{Education} = 49.99 + (-0.52) \cdot \text{Fertility} + (-0.23) \cdot \text{Agriculture} + 0.08 \cdot \text{Catholic} + 0.28 * \text{Infant.Mortality}$$

Variante 2 ohne Infant Mortality

$$\text{Education} = 53.85 + (-0.49) \cdot \text{Fertility} + (-0.24) \cdot \text{Agriculture} + 0.08 \cdot \text{Catholic}$$

Variante 3 ohne Infant Mortality und ohne Val de Genenve

$$\text{Education} = 45.27 + (-0.38) \cdot \text{Fertility} + (-0.20) \cdot \text{Agriculture} + 0.06 \cdot \text{Catholic}$$

Einfluss des Hebelepunkts!!!

moderate Modellanpassung bei R^2 62.16%

Genf ein "guter Hebelepunkt" -> Modellanpassung wird besser

Zusammenhänge zwischen einer kategorialen und metrischen Variablen - logistische Regression

Für kontinuierliche Variablen X und Y kann man sich grundsätzlich über viele Arten des Zusammenhangs Gedanken machen, wir haben uns auf lineare Abhängigkeit konzentriert, aber auch erkannt, dass man nichtlineare Abhängigkeiten auch durch Transformation linear machen kann. Doch nun wollen wir ein y modellieren, welches nur 2 Kategorien hat: "Erfolg" - "Misserfolg", wobei wir diese beiden so allgemein verstehen, dass auch "positiver Krankheitstest" oder "Diabetesdiagnose" solche Ausprägungen von Interesse sein können.

Die dabei zur Anwendung kommende Methode der logistischen Regression wird in der Medizin und Pharmacologie ebenso verwendet, wie in der IT oder Ökonomie. Anstatt einer Regressionsgerade werden wir hier Schwellwerte benötigen anhand derer in die eine oder andere Kategorie klassifiziert wird.

Inferenz durch logistische Regression

Bei linearer Regression hatten wir simultane Messungen von zwei metrischen Merkmalen (x_i, y_i) um eine Gerade mit der *Modellgleichung*

$$y = \alpha + \beta \cdot x.$$

Hierbei wird aus dem Kontext x als die **unabhängige Variable** bezeichnet, was damit zusammenhängt, dass diese bei Experimenten bewusst eingestellt werden kann, während y als die **abhängige Variable** bezeichnet wird.

Wir möchten eine ähnliche Methode wie für numerische y-Werte verwenden, um binäre Outcomes anzupassen, also Klassifikation in 2 Kategorien zu betreiben.

Binäre Daten werden mithilfe der Binomialverteilung modelliert, es ist also

$$f(y | p) = p^y \cdot (1-p)^{1-y}$$

wobei p die Erfolgswahrscheinlichkeit ist und y die Werte 0 (Misserfolg) oder 1 (Erfolg) annehmen kann.

Der Erwartungswert für den Ausgang y ist $\mathbb{E}(y) = p$. Also in $p \cdot 100\%$ der Fälle erwartet man sich für y den Wert 1, in $(1-p) \cdot 100\%$ der Fälle erwartet man sich für y den Wert 0.

Um nun ein Modell für den Erwartungswert $E(y_i) = \mu_i = p_i$ einer binären Variable zu formulieren, soll wieder ein *linearer Prädiktor* $x_i^\top \beta$ verwendet werden.

Problem: Während der lineare Prädiktor $x_i^\top \beta$ prinzipiell alle reellen Werte annehmen kann, liegt der Erwartungswert (= Erfolgswahrscheinlichkeit) $\mu_i = p_i$ immer im Intervall $[0, 1]$.

Lösung: Verwende eine *Link-Funktion* g , die das Intervall $[0, 1]$ auf die reellen Zahlen abbildet, so daß

$$g(\mu_i) = x_i^\top \beta$$

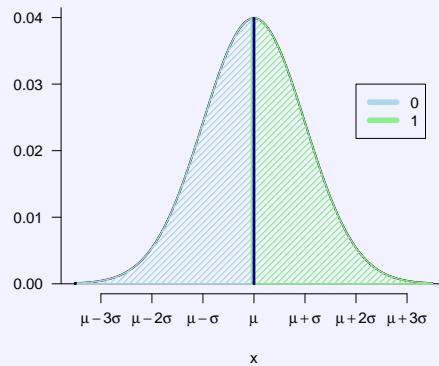
Link-Funktionen der verallgemeinerten linearen Regression

Frage: Was ist eine *geeignete* Link-Funktion?

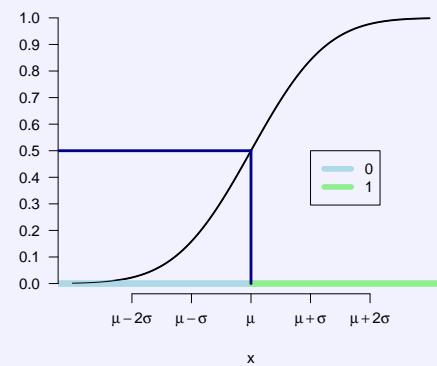
Es gibt unterschiedliche Herangehensweisen, wie man eine Link-Funktion erstellt, also eine Funktion, welche von beliebigen reellen Werten auf "0" und "1" umrechnet und dadurch einen Zusammenhang (Link) zwischen reellen Werten, die die Regressionsgleichung berechnet, und Kategorien, die die eigentlichen y-Werte sind, herstellt.

- **Probit-Link-Funktion der Probit-Regression:** Die probit-Link-Funktion benutzt die Wahrscheinlichkeitsverteilungsfunktion der Normalverteilung als Methode zur Umrechnung.

Dichtefunktion der Normalverteilung

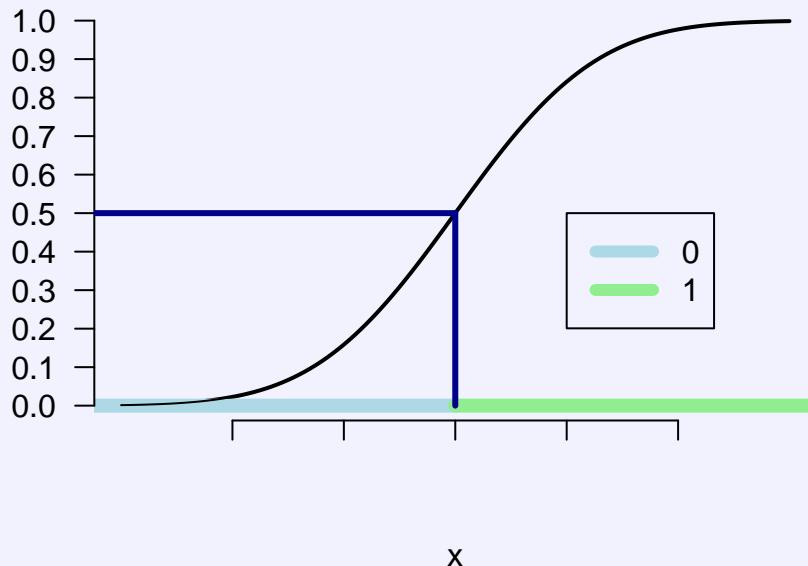


Verteilungsfunktion der Normalverteilung



- **Logit-Link-Funktion der logistischen Regression:** Eine ähnliche Struktur wie die Normalverteilungsverteilungsfunktion besitzt die Sigmoidkurve der logistischen Funktion.

Sigmoidkurve der logistischen Funktion



Als Link-Funktion in der logistischen Regression wird also die sogenannte Logit-Funktion

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

verwendet. Diese berechnet die logarithmierten Chancenverhältnisse (log Odds Ratios).

Dadurch wird die verallgemeinerte lineare Regression

$$E(y_i) = \mu_i = p_i$$

mit Linkfunktion

$$g(\mu_i) = x_i^\top \beta$$

als Modell für die binäre Variable y angepasst.

Logistische Regression

Bei der logistischen Regression passt man das lineare Modell für die Fraktion der Wahrscheinlichkeiten an

$$\text{logit}(y) = \log \left(\frac{\mathbb{P}[y_i = 1|\mathbb{X}]}{\mathbb{P}[y_i = 0|\mathbb{X}]} \right) = \alpha + \beta \cdot x_i + \varepsilon_i$$

Das Ergebnis der logistischen Regressionsfunktion

$$\text{logit}(y) = \log \left(\frac{\mathbb{P}[y_i = 1|\mathbb{X}]}{\mathbb{P}[y_i = 0|\mathbb{X}]} \right) = \alpha + \beta \cdot x_i + \varepsilon_i$$

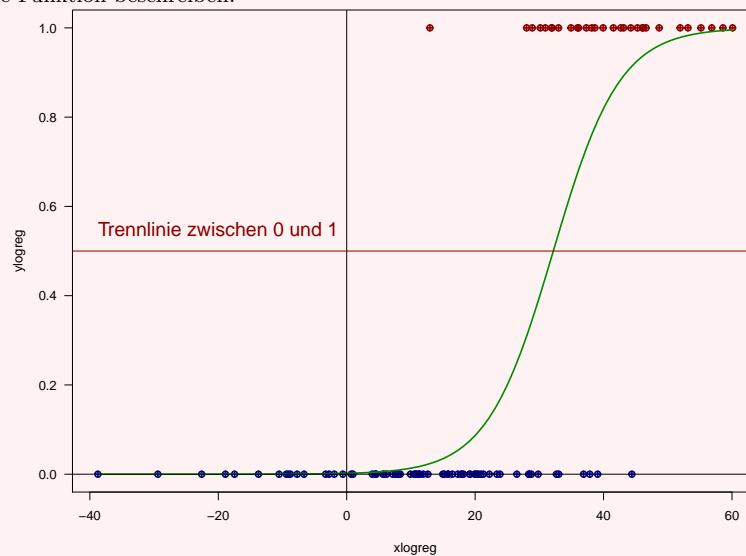
ist die Ermittlung der Wahrscheinlichkeiten in Gruppe 0 oder 1 zu fallen:

$$\mathbb{P}[y = 1|\mathbb{X}] = \frac{\exp(\alpha + \beta \cdot x)}{1 + \exp(\alpha + \beta \cdot x)}$$

und

$$\mathbb{P}[y = 0|\mathbb{X}] = \frac{1}{1 + \exp(\alpha + \beta \cdot x)}$$

welche die logistische Funktion beschreiben.



```
glm(y~x,data=daten,family=binomial(link = "logit"))
```

Exkurs: Vierfeldertafeln, Chancen und Chancenverhältnisse anhand eines Beispiels In der explorativen Datenanalyse wurde bereits der Zusammenhang von zwei kategorialen Merkmalen mit Hilfe ihrer Kontingenztafel und dem zugehörigen Mosaikplot untersucht

Beispiel: Vorsorgeuntersuchung und Geschlecht

Geschlecht vs. Vorsorgeuntersuchung		nein	ja
	Frauen	273	183
	Männer	627	217

Um zu untersuchen, ob sich die Wahrscheinlichkeit für Männer und Frauen unterscheidet, zur Vorsorgeuntersuchung zu gehen, sind die relativen Häufigkeiten bedingt unter dem Geschlecht geeigneter.

Geschlecht vs. Kauf		nein	ja	Summe
	Frauen	0.599	0.401	1
	Männer	0.743	0.257	1

Chancen - Odds

Um diese vier bedingten relativen Häufigkeiten noch weiter zu komprimieren, betrachtet man anstatt der **Wahrscheinlichkeiten** die zugehörigen **Chancen**. Die Chance eines Ereignisses ist die Wahrscheinlichkeit für sein Eintreten geteilt durch die Gegenwahrscheinlichkeit. Im englischen heißen diese **Odds**.

$$\text{Odds(Vorsorgeuntersuchung)} = \frac{P(\text{Vorsorgeuntersuchung})}{1 - P(\text{Vorsorgeuntersuchung})}$$

Bei den Frauen beträgt die Chance auf die Vorsorgeuntersuchung also $0.401/0.599 = \frac{2/5}{3/5} = 0.67$, also in etwa 2:3.

Bei den Männern hingegen beträgt die Chance auf die Vorsorgeuntersuchung $0.257/0.743 = \frac{1/4}{3/4} = 0.346$ also in etwa 1:3.

Chancenverhältnis

Wenn man auch noch diese beiden Zahlen ins Verhältnis setzt, so nennt man das Ergebnis **Chancenverhältnis** oder **Odds Ratio**.

Hier besagt das Chancenverhältnis von $0.346/0.67 = 0.516$, daß die Chancen auf die Vorsorgeuntersuchung bei Männern nur rund halb so groß sind wie bei Frauen.

Das Chancenverhältnis kann auch leicht aus der Originaltabelle berechnet werden.

Geschlecht vs. Vorsorgeuntersuchung		nein	ja
	Frauen	273	183
	Männer	627	217
OddsRatio		$\frac{273 \cdot 217}{627 \cdot 183} = 0.516$	

Umsetzung der logistischen Regression in R

Beispiel: Prüfungsdaten - logistische Regression

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1

```
glm(Pass~Hours, family=binomial(link = "logit"))
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.08	1.76	-2.32	0.02
Hours	1.50	0.63	2.39	0.02

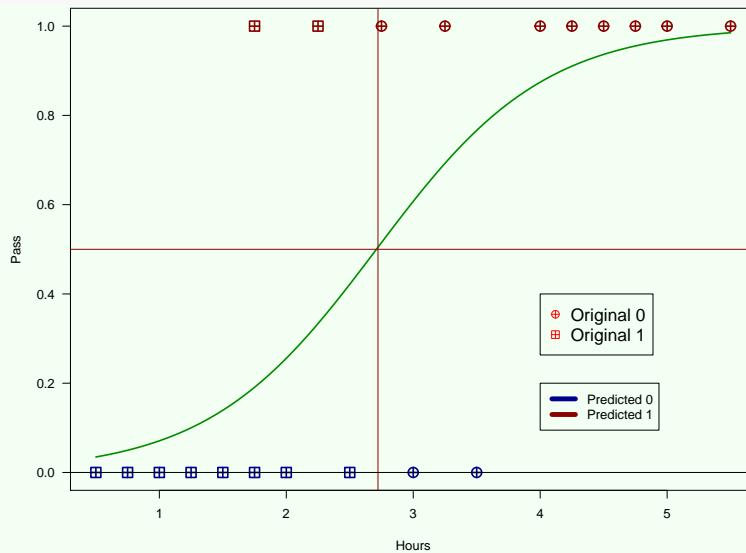
0.017 = Odds-Ratio, die Prüfung zu schaffen, wenn 0 Stunden gelernt wird

4.502 = Faktor, um den der Odds-Ratio, die Prüfung zu schaffen, steigt, für jede Stunden, die gelernt wird

$\exp(-4.078 + 1.505 \cdot Hours)$ = Odds-Ratio, die Prüfung zu schaffen, wenn Hours Stunden gelernt wird

$\frac{1}{1 + \exp(-(-4.078 + 1.505 \cdot Hours))}$ = Wahrscheinlichkeit, die Prüfung mit Hours Stunden Lernen zu bestehen (= probability of passing the exam)

```
newdat <- data.frame(Hours=seq(min(Hours), max(Hours), len=100))
newdat$Pass = predict(logitPass, newdata=newdat, type="response")
schwelle<-newdat[which(abs(newdat$Pass-0.5)<0.01),"Hours"]
schwellePass<-newdat[which(abs(newdat$Pass-0.5)<0.01),"Pass"]
plot(Hours,Pass,type="p",pch=c(10,12)[(predict(logitPass,type="response")<schwellePass)+1],lwd=2,las=1,col=c("darkblue","darkred"))
lines(Pass ~ Hours, newdat, col="green4", lwd=2)
abline(v=0); abline(h=0, col="red4"); abline(v=schwelle, col="red4")
text(x=55,y=0.55,"Trennlinie zwischen 0 und 1",col="red4",cex=1.5)
legend(4,0.4,legend=c("Original 0","Original 1"),col=c(1,1),cex=1.3,pch=c(10,12))
legend(4,0.2,legend=c("Predicted 0","Predicted 1"),col=c("darkblue","darkred"),lwd=6)
```



Erweiterte Aspekte der Regression

Lineare Regression mit Nebenbedingungen: LASSO, Ridge-Regression und Elastic nets

Anstatt das lineare Modell

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} + \varepsilon_i$$

mithilfe der kleinsten Quadrate Methode anzupassen, wird eine Optimierung mit Nebenbedingungen durchgeführt.

- Ridge Regression mit der Nebenbedingung

$$\sum_{i=1}^n \beta_i^2 < c \text{ für ein } c > 0$$

- LASSO least absolute shrinkage and selection operator

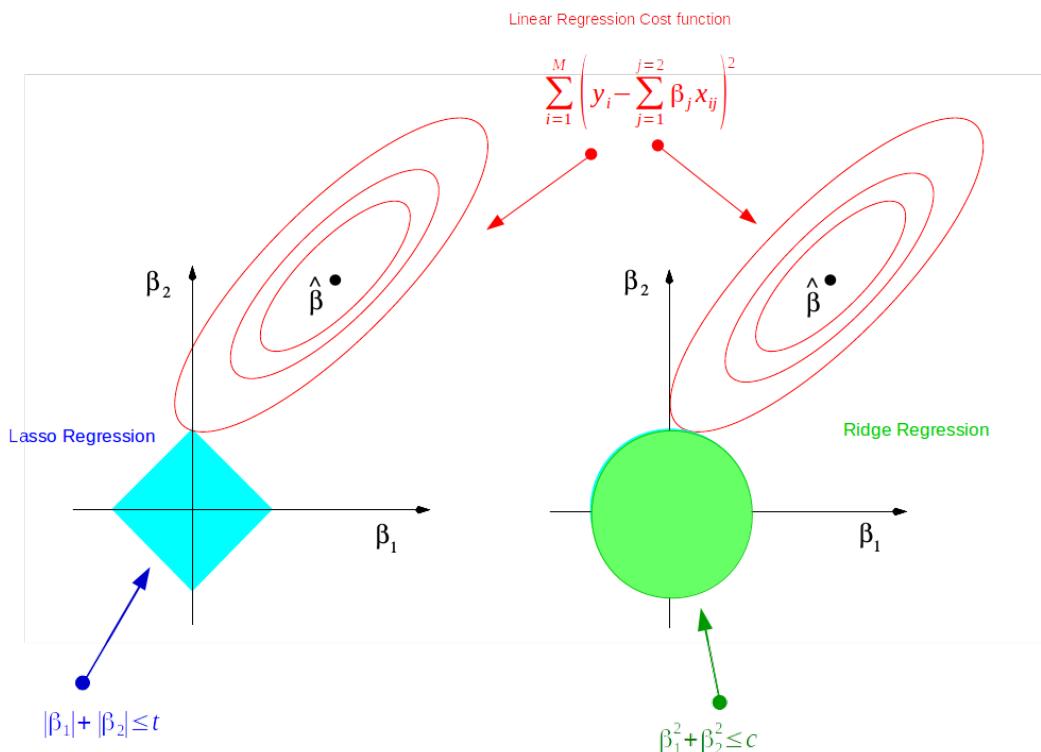
$$\sum_{i=1}^n |\beta_i| < c \text{ für ein } c > 0$$

- sowie ihre Kombination elastic nets in Machine Learning

$$(1 - \alpha) \cdot \sum_{i=1}^n \beta_i^2 + \alpha \cdot \sum_{j=1}^n |\beta_j| < c \text{ für ein } c > 0, p \in [0; 1]$$

Die penalty functions, also die unterschiedlichen Funktionen bei linearer (LASSO) und quadratischer (Ridge) Bestrafung für die Größe der Modellkoeffizienten, wird im folgenden Diagramm dargestellt.

Dimension Reduction of Feature Space with LASSO



Quelle: Scikit

Hier hilft es zu wissen, dass die Kreisgleichung $x^2 + y^2 = r^2$ mit der allgemeinen Ellipsengleichung $(x/a)^2 + (y/b)^2 = r^2$ durch die unterschiedlichen "Streckungs- oder Stauchungsfaktoren" a und b der Ellipsenachse zusammenhängen, also durch Umskalierung der Achsen jede Ellipse zum Kreis werden kann.

Wir haben das Konzept der Normalisierung von Variablen, also das Konzept der z-Transformation $z = \frac{x - \mu}{\sigma}$ kennengelernt, wodurch wir jede Variable standardisieren, also auf Mittelwert 0 und Standardabweichung 1 umtransformieren können. Dadurch passiert eben die Umwandlung einer Ellipse mit allgemeinen Achsen in den Einheitskreis.

Die drei verallgemeinerten Regressionsmethoden

- Ridge Regression
- LASSO Regression
- sowie ihre Kombination elastic nets

sind einfache Techniken, um Modellkomplexität und Überanpassung zu verhindern, die mit einfacher linearer Regression auftreten können.

```
library(glmnet)
glmnet(x,y,family, alpha)
```

Elastic Nets, LASSO und Ridge in R family beschreibt die Art, wie die abhängige Variable y modelliert werden soll.

family="gaussian" ist numerisch wie lineare Regression.

family="binomial" ist binär wie bei logistischer Regression.

alpha ist der "elasticnet mixing parameter", with $0 \leq \alpha \leq 1$.

alpha=1 ist LASSO, alpha=0 Ridge Regression.

```
set.seed(1); n <- 500; p=20
X <- matrix(rnorm(n*p), ncol=p)
X <- scale(X)
eps <- rnorm(n, sd=5)
beta <- 3:5
y <- 2 + X[,1:3] %*% beta + eps
```

Ridge Regression in R The connection between linear regression and Ridge regression is via the penalty term of the squared coefficients.

```
fitLM <- lm(y~X)
round(coef(fitLM),2)
```

```
## (Intercept)      X1      X2      X3      X4      X5
##     2.01      3.04     3.94     4.92    -0.12     0.24
##     X6       X7      X8      X9      X10     X11
##    -0.12     -0.16     -0.15    -0.06    -0.17    -0.08
##     X12      X13      X14      X15      X16     X17
##    -0.14      0.43     -0.08    -0.03     0.02     0.00
##     X18      X19      X20
##    -0.21     -0.18     -0.31
# root sum of squared coefficients
sqrt(sum(coef(fitLM)[-1]^2))
```

```
## [1] 7.044342
```

The package `glmnet` can perform ridge regression in R using the function `glmnet`. To obtain a ridge regression there, the parameter `alpha` needs to be set to zero.

```
lambda.grid <- 10^seq(10, -2, length=100)
fitRR <- glmnet(x=X, y=y, alpha=0, lambda=lambda.grid)
dim(coef(fitRR))
```

```
## [1] 21 100
```

Lasso in R The function `glmnet` can also compute the lasso. For that purpose the parameter `alpha` needs to be set to one.

```
fitL <- glmnet(x=X, y=y, alpha=1, lambda=lambda.grid)
dim(coef(fitL))
```

```
## [1] 21 100
```

```
fitRR$lambda[50]
```

Vergleich von Ridge Regression und LASSO

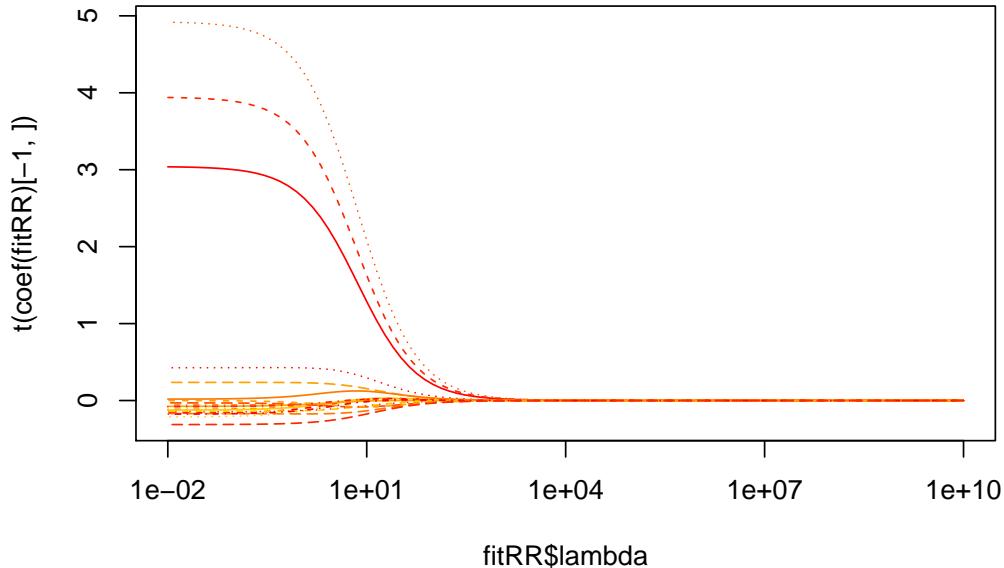
```
## [1] 11497.57
round(coef(fitRR)[,50],2)
```

```

## (Intercept)      V1      V2      V3      V4      V5
## 2.01       0.00     0.00     0.00     0.00     0.00
## V6        V7      V8      V9      V10     V11
## 0.00       0.00     0.00     0.00     0.00     0.00
## V12       V13     V14     V15     V16     V17
## 0.00       0.00     0.00     0.00     0.00     0.00
## V18       V19     V20
## 0.00       0.00     0.00
sqrt(sum(coef(fitRR)[-1,50]^2))

## [1] 0.004567586
matplot(fitRR$lambda, t(coef(fitRR)[-1,]), type="l", log="x")

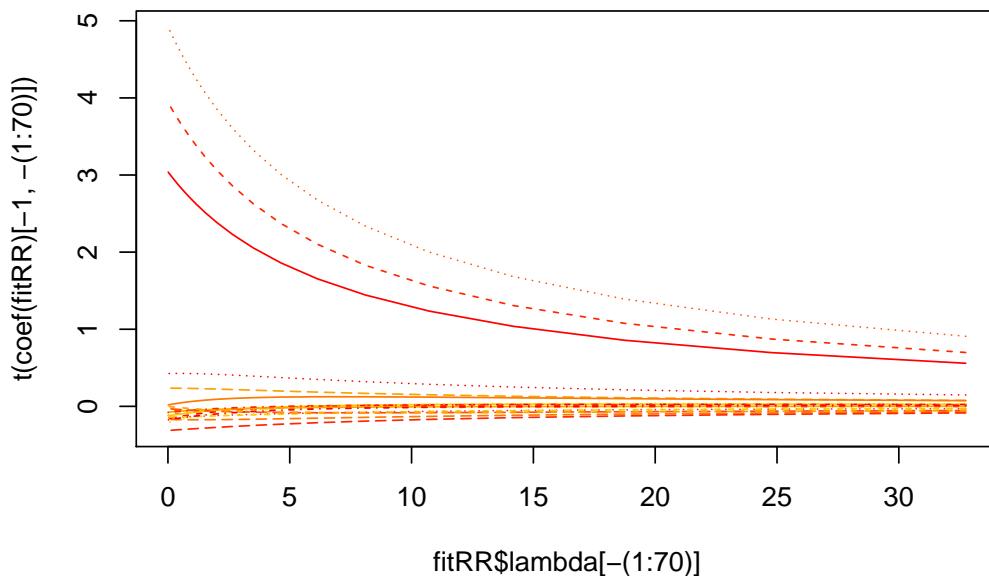
```



```

matplot(fitRR$lambda[-(1:70)], t(coef(fitRR)[-1,-(1:70)]), type="l")

```



Unlike Ridge regression LASSO shrinks the irrelevant coefficients to 0.

```
fitL$lambda[50]
```

```

## [1] 11497.57
round(coef(fitL)[,50],2)

## (Intercept)      V1      V2      V3      V4      V5
## 2.01       0.00     0.00     0.00     0.00     0.00
## V6        V7      V8      V9      V10     V11
## 0.00       0.00     0.00     0.00     0.00     0.00

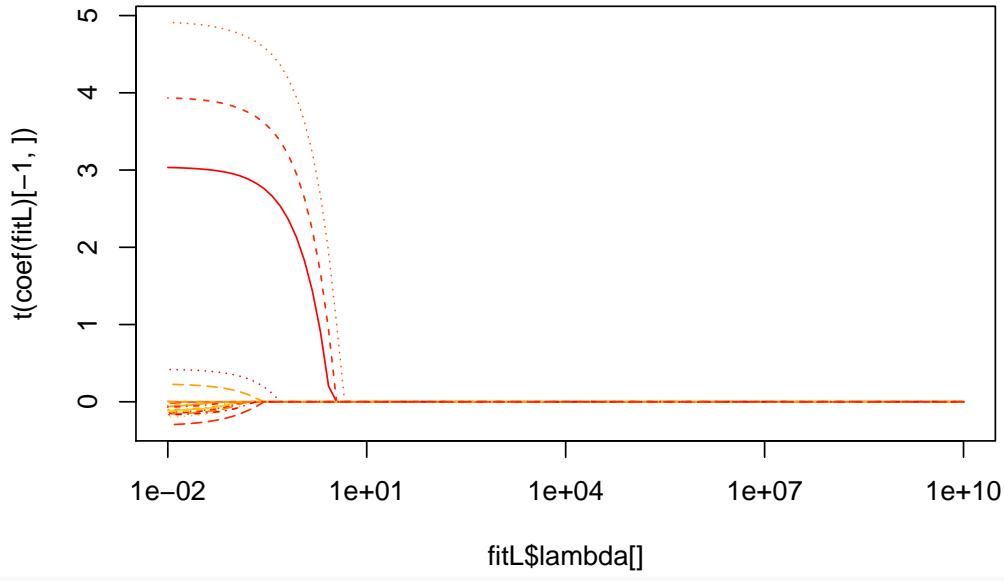
```

```

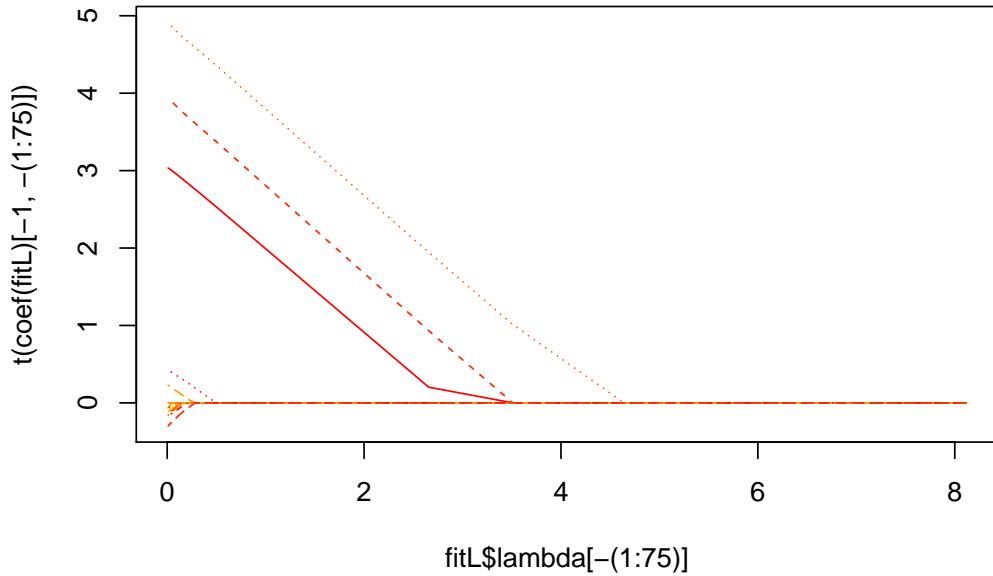
##      0.00      0.00      0.00      0.00      0.00      0.00
##    V12      V13      V14      V15      V16      V17
##    0.00      0.00      0.00      0.00      0.00      0.00
##    V18      V19      V20
##    0.00      0.00      0.00
sqrt(sum(coef(fitL)[-1,50]^2))

## [1] 0
matplot(fitL$lambda[], t(coef(fitL)[-1,]), type="l", log="x")

```



```
matplot(fitL$lambda[-(1:75)], t(coef(fitL)[-1,-(1:75)]), type="l")
```



Regression Trees und Forests

Classification and Regression Tree (CART)

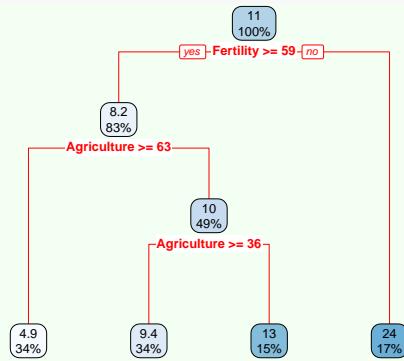
Anstatt ein Modell anzupassen, wird bei Baum-basierten Methoden in Untergruppen partitioniert. Die rekursive Partitionierung basiert auf Schwellwerten aus den zufälligen Untergruppen, sodass die Residuenquadratsumme der Aufteilung minimiert wird. Dementsprechend werden Indikatoren, in welchen Untergruppenbereich Beobachtungen fallen, aus dem Baum vorhergesagt.

```
library(rpart); rpart(y~x1+x2,daten, method)
```

method =“class” for a classification tree OR “anova” for a regression tree

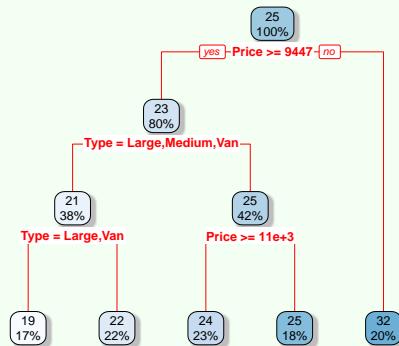
Beispiel für einen linearen Modell-Regression Tree für einen numerischen Output

```
swisstree<-rpart(Education+Fertility+Agriculture+Catholic+Infant.Mortality,data=swiss, method="anova")
rpart.plot(swisstree,type=2)
```

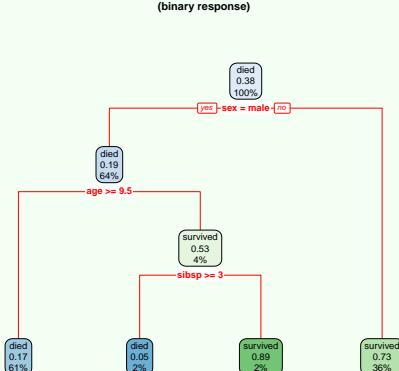


```
anova.model <- rpart(Mileage ~ ., data = cu.summary)
```

```
rpart.plot(anova.model,
           shadow.col = "gray",          # add shadows just for kicks
           main = "miles per gallon\n(continuous response)\n")
```



```
Beispiel für einen logistischen Regression Tree mit kategorialen Zielvariablen
binary.model <- rpart(survived ~ ., data = ptitanic, cp = .02)
rpart.plot(binary.model, main = "Titanic survived\n(binary response)")
```



Vor- und Nachteile von Regression Trees

\

Vorteile:

- einfach und direkt interpretierbar ohne mathematisches Modell
- Prädiktionen sind schnell und einfach
- mit teilweise fehlenden Daten immer noch verwendbar bis zur Verzweigung, wo Daten fehlen
- einfache, schneller Algorithmus

Nachteile:

- hohe Varianz zwischen unterschiedlichen Berechnungen
- sehr hoher Fehler in der Prädiktion

-> -> ->