# Applied Statistics for Data Science

**Zusammenfassung**

Stephan Stofer

3. Dezember 2020

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# 1 Einleitung

Applied statistics wenden wir Statistik auf konkrete Alltagsprobleme an. Dazu wenden wir die vierstufige Problemlösungsstrategie an.

## 1.1 Vierstufige Problemlösungsstrategie

1. Erste Schritte Es ist nicht immer klar was die effizienteste Antwort auf das Problem ist. Informationen organisieren/sammeln. Problem gegebenenfalls mit eigenen Worten formulieren. Sind alle Infos da, die wir für die Lösung des Problems brauchen?
2. Plan ausarbeiten Herausfinden, welche Schritte nötig sind, um das Problem zu lösen.
3. Plan ausführen Die Schritte in den im Punkt 2 definierten Abfolge ausführen.
4. Resultat interpretieren Überprüfung ob das Resultat möglich und sinnvoll ist. Interpretation des Resultats in den Worten der Problemstellung.

# 2 Deskriptive Statistik - Eindimensionale Daten

## 2.1 Datensätze

Datensätze sind Zusammenstellungen von Daten die in vielen Formen vorkommen können.

### 2.1.1 Liste

Eine Liste von Daten ist die einfachste Variante eines Datensatzes. Sie enthält zum Beispiel die Körpergrösse in Meter von fünf Personen.

$$1.75, 1.80, 1.72, 1.65, 1.54$$

Solche Listen heissen auch *eindimensionale Datensätze* oder *Messreihen*

### 2.1.2 Tabellen

Die häufigste Form von Datensätzen sind Tabellen oder *zweidimensionale Datensätze*. Bei Einträgen mit Zahlen spricht man von *quantitativen* Daten, sprich Messwerte und können theoretisch jede beliebigen Zahlenwert annehmen. Andere (zB. ein Spalte Geschlecht oder Nationalität) sind sogenannte *qualitative* Daten und können nur eine bestimmte Anzahl Werte annehmen. Sie können aber auch Zahlen sein.

## 2.2 Deskriptive Statistik

Die deskriptive Statistik befasst sich mit der *Darstellung* von Datensätzen (lat. describere, beschreiben). Dabei werden Datensätze durch gewisse Zahlen charakterisiert (z.B. Mittelwert) und/oder graphisch dargestellt. Quantitative Daten werden organisiert und zusammengefasst. Die Interpretation und darauffolgende statistische Analyse soll vereinfacht werden. Wir erledigen dies mit Hilfe von:

- graphischen Darstellungen wie Histogramme und Boxplots
- *Kennzahlen*, die Daten numerisch zusammenfassen, die Durchschnitt und Standardabweichung

  Daten sollten immer mit Hilfe von graphischen *und* Kennzahlen dargestellt werden. Nur auf diese Weise können (teils unerwartete) Strukturen und Besonderheiten entdeckt werden.

Man muss sich ausserdem bewusst sein: Werden *Daten zusammengefasst*, gehen *Informationen verloren*!

### 2.2.1 Bezeichnung von Daten

Im Folgenden werden Daten mit $x1, x2, ..., xn$ bezeichnet, wobei $n$ der *Umfang* der Messreihe genannt wird.

## 2.3 Kennzahlen

Meistens ist es sinnvoll, Datensätze durch eine Zahl, also numerisch zusammenzufassen und damit zu beschreiben. Sie werden dabei auf eine odere mehrere Zahlen reduziert. Die zwei wichtigsten sind:

- Lageparameter oder Lagemasse; beschreiben *wo* sich Daten befinden. Beschreibt als Bsp. die *mittlere Lage* der Messwerte (muss nicht Durchschnitt gemeint sein)
- Streuungsparameter oder Streuungsmasse; beschreiben *wie* sich die Daten um die mittlere Lage verteilen. Die *Variabilität* oder *Streuung* der Messwerte gibt die durchschnittliche Abweichung von der mittleren Lage an

### 2.3.1 Arithmetisches Mittel

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

Bekannteste Grösse für eine mittlere Lage ist der *Durchschnitt* oder das **Arithmetische Mittel** $\bar{x}$. In der Notation $\bar{x}_n$ beschreibt $n$ wieder den Umfang der Messreihe.


Abbildung 2.1: Graphische Darstellung des arithmetischen Mittelwertes

#### 2.3.1.1 Arithmetisches Mittel mit R

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

### 2.3.2 Empirische Varianz und Standardabweichung

Das arithmetische Mittel beschreibt einen Datensatz nur unvollständig. In der Abbildung 2.2 ersichtlich, dass zwei Datensreihen dasselbe arithmetische Mittel haben. Allerdings liegen die Punkte der ersten Datenreihe weiter vom Mittelpunkt entfernt, als die Punkte der ersten. Wir sprechen von unterschiedicher *Streuung* der Daten um den Durchschnitt.

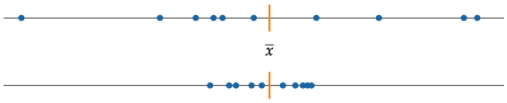


Abbildung 2.2: “Grosse” und “kleine” Streuung von zwei Messreihen

Die gebräuchlichsten Masse für die Streuung oder Variablität von Messwerten sind die *empirische Varianz* und *empirische Standardabweichung*.

#### 2.3.2.1 Mathematische Definition Empirische Varianz

Bei der Varianz werden die Abweichungen quadriert, dadurch können sich diese nicht gegenseitig aufheben. In einigen Bücher steht bei der Definition für die Varianz im Nenner $n$, anstatt $n-1$. Für kleine Datensätze spielt dies eine Rolle, bei grossen jedoch vernachlässigbar. r verwendet mit dem Befehl `r var(x)` auch $n-1$.

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

#### 2.3.2.2 Mathematische Definition Empirische Standardabweichung

Durch das Quadrieren erhalten die Werte eine neue Einheit (z.B. $cm^2$). Durch das Wurzelziehen führen wir diese wieder ihrer ursprünglichen Einheit zu und erhalten damit die Standardabweichung.

$$s_x = \sqrt{Var(x)}$$

> Nur die Standardabweichung $s_x$ lässt sich korrekt interpretieren. Der Wert der empirischen Varianz hat keine physikalische Bedeutung. Wir wissen nur, je grösser der Wert, umso grösser die Streuung.

#### 2.3.2.3 Empirische Varianz bzw. Standardabweichung mit R

```r
var(waageA)
# [1] 0.000574359
sd(waageA) # sd = standard deviation
## [1] 0.02396579
```

### 2.3.3 Median

Der Median ist ein Lagemass für denjenigen Wert, bei dem rund die Hälfte der Messwerte kleiner oder gleich und die andere Hälfte grösser oder gleich diesem Wert sind. Um den *Median* zu bestimmen, müssen alle Daten erst geordnet werden. Ist die Anzahl der Daten *ungerade*, gibt es eine *mittlere* Beobachtung. Bei einer *geraden* Anzahl gibt es zwei *gleichwertige mittlere* Beobachtungen. Als Median benützen wir in diesem Fall den Durchschnitt der beiden mittleren Beobachtungen $\frac{m_1+m_2}{2}$.

- Der Median muss *kein* Wert aus dem Datensatz sein
- Er wird auch *Zentralwert* oder *mittlerer Wert* genannt
- der Median ist sehr *robust*, dies bedeutet dass er weniger stark durch extreme Beobachtungen (Ausreisser) beeinflusst wird als das arithmetische Mittel.

In der Abbildung 2.3 erkennt man, wie durch den blauen Punkt (zweiter Zahlenstrahl ganz rechts) der Durchschnitt von $x_n$ zu $x_n^*$ verändert wird.


Abbildung 2.3: Die Robustheit des Median

#### 2.3.3.1 Median mit R

```
median(waageA)
## [1] 80.03
```

#### 2.3.3.2 Bermerkungen zum Median

Der Median ist gerechter. Weshalb er auch für das berechnen des mittleren Einkommens verwendet wird. Die beiden Lagemasse für die mittlere Lage sollten immer gemeinsam betrachtet werden. Eine grosse Abweichung zwischen den Werten deutet auf besondere Verteilung der Daten hin.

### 2.3.4 Quartile

Die Quartile sind analog dem Median definiert, aber nicht für 50% der Daten die grösser oder kleiner sind, sonder für 25% bzw. 75% der Daten. Das *untere* Quartil ist derjenige Wert, bei welchem 25% aller Beobachtungen kleiner oder gleich und 75% grösser oder gleich diesem Wert sind. Entsprechend ist das *obere* Quartil derjenige Wert, bei dem 75% aller Beobachtungen kleiner oder gleich und 25% grösser oder gleich diesem Wert sind.

Hat eine Messreihe 13 Messpunkte sind 25% davon 3.25. Wir runden jeweils auf –> der vierte Wert wird dann zum unteren Quartil.

#### 2.3.4.1 Quartil in R

```
# Syntax fuer das untere Quartil: p = 0.25, type definiert den verwendeten Algorithmus
# https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile
quantile(waageA, p = 0.25, type = 2)
## 25%
## 80.02

# Syntax fuer das obere Quartil: p = 0.75
quantile(waageA, p = 0.75, type = 2)
## 75%
## 80.04
```

### 2.3.5 Quartilsdifferenz

Die Quartilsdifferenz ist definiert als die Differenz der beiden Quartile: $oberesQuartil - unteresQuartil$ Sie ist ein Streuungsmass für die Daten. Es misst die Länge des Intervalls, das etwa die Hälfte der mittleren Beobachtungen enthält. Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte beim Median und umso kleiner ist die Streuung. Dieses Streuungsmass ist robust.

#### 2.3.5.1 Quartilsdifferenz in R

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

Dies bedeutet, dass die Hälfte aller Messwerte in einem Bereich der Länge 0.02 liegen.

### 2.3.6 Quantile

Mit den *Quantilen* kann das Konzept der Quartile auf jede beliebige Prozentzahl verallgemeinert werden. Das *empirische* $\alpha$*-Quantile* ist derjenige Wert, bei dem $\alpha * 100$ der Datenpunkte kleiner oder gleich und $(1 - \alpha) * 100$ der Punkte görsser oder gleich diesem Wert sind.

#### 2.3.6.1 Quantil in R

```r
quantile(waageA, p = 0.1, type = 2)
## 10%
## 79.98

quantile(waageA, p = 0.7, type = 2)
## 70%
## 80.04
```

Weiteres Beispiel mit versch. Quantilen in einer Zeile

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

Rund 20% der Lernenden haben also eine 3.6 oder waren schlechter und rund 80 % der Lernenden waren gleich oder besser als dieser Wert. Genau 20% der Lernenden ist nicht möglich, da dies 4.8 Lernenden entsprechen würde. Das 60%-Quantil besagt, dass rund 60 Prozent der Lernenden Noten von 5 oder weniger haben. Folglich haben rund 40% eine 5 oder sind besser.

## 2.4 Graphische Methoden

Daten graphisch dazustellen ist ein sehr wichtiger Aspekt der Datenanalyse.

### 2.4.1 Histogramm

Histogramme helfen bei der Frage, in welchem *Wertebereich* besonders viele Datenpunkte liegen. Besonders dann, wenn die Datenmenge gross ist und es keinen Sinn macht, alle Werte einzeln zu betrachten.

#### 2.4.1.1 Histgramm in R

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

- `rnorm(n = 200, mean = 100, sd = 15)` wählt zufällig 200 normalverteilte Daten mit Mittelwert 100 und einer Standardabweichung von 15 aus
- `hist(iq, ...)` zeichnet das Histogramm für `iq`

- xlab ist das x-Label
- ylab ist das y-Label
- col definiert die Farbe
- main steht für Haupttitel

Beim Aufbau eines Histogramm werden die Daten in Klassen eingeteilt. Dabei wird die *Anzahl* der Klassen (Balken) anhand verschiedenen Faustregeln gebildet. Bei weniger als 50 Messungen sind es 5 bis 7, bei mehr als 250 wählt man 10 bis 20 Klassen. Die Wahl der Anzahl ist relevant für die Aussagekraft eines Histogrammes. Es gibt keine allgemeine Grundregel für die Wahl.

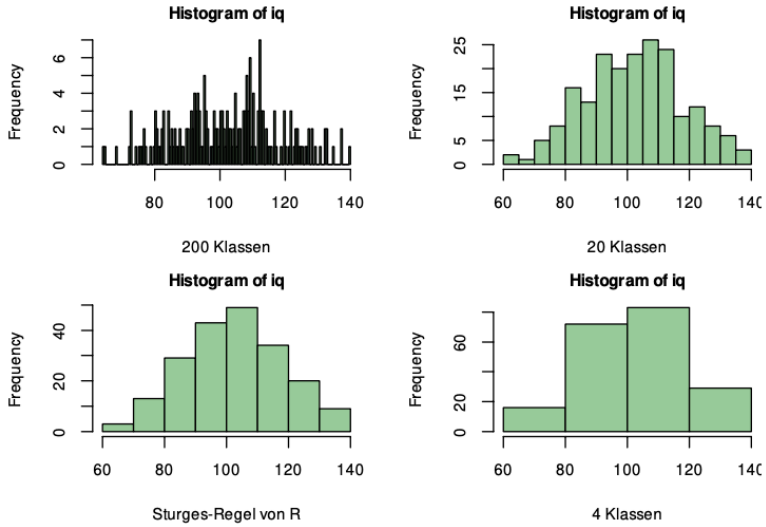


Abbildung 2.4: Vergleich der Histogramme mit verschiedener Klassenwahl

#### 2.4.1.2 Bimodales Verhalten

*Bimodales* Verhalten ist sichtbar, wenn es zwei "Hügel" im Histogramm gibt

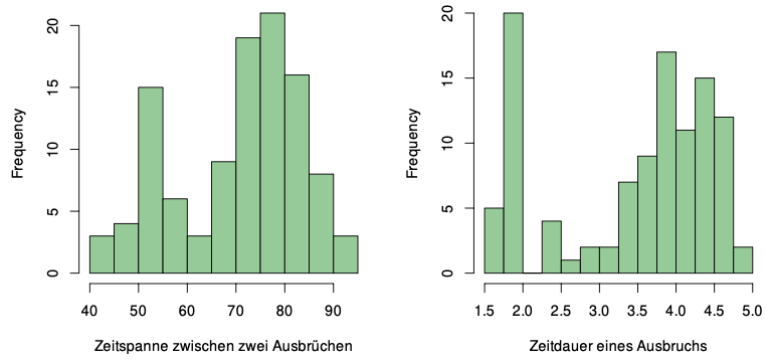


Abbildung 2.5: Bimodales Verhalten in zwei Histogrammen

#### 2.4.1.3 Schiefe von Histogrammen

Wir betrachten die Histogramme in Abbildung 2.6

- Das Histogramm links ist symmetrisch bezüglich 5. Die Daten sind um 5 auf beiden Seiten ähnlich verteilt.
- In einem *rechtsschiefen* Histogramm sind die meisten Daten links im Histogramm

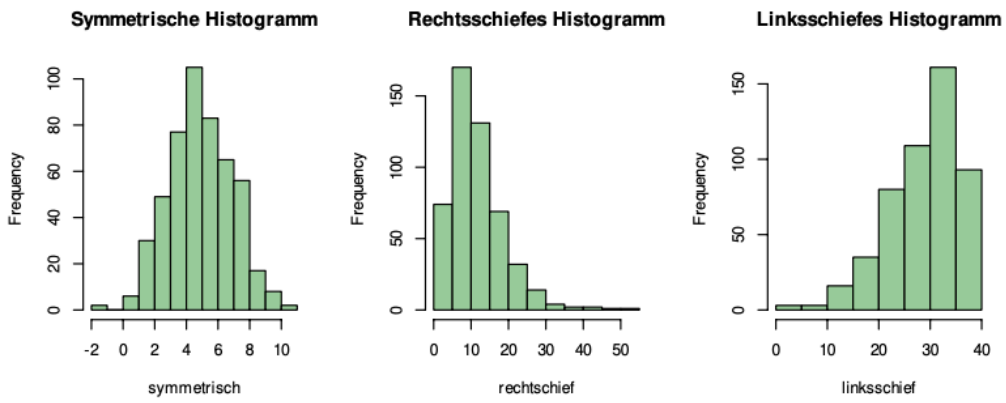


Abbildung 2.6: Symmetrisches, rechts- und linksschiefes Histogramm

- In einem *linksschiefen* Histogramm sind die meisten Daten rechts im Histogramm

Die Bezeichnung "rechts" und "links" bezieht sich immer auf die Richtung von *weniger* Daten sind.

#### 2.4.1.4 Normiertes Histogramm

In den vorherigen Histogrammen ist die Höhe der Balken gerade der Anzahl der Beobachtungen in einer Klasse. In einem normierten Histogramm wird die Balkenhöhe so gewählt, dass die *Balkenfläche* dem Anteil der jeweilgen Beobachtungen an der Gesamtanzahl entspricht. Die Gesamtfläche der Balken muss dann gleich eins sein. Auf der vertikalen Achse ist dann die *Dichte* aufgetragen (entspricht *nicht* Prozentwerten).

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

- mit `freq = F` (frequency false) wird das Histogramm normiert gezeichnet
- Die Option 'ylim = c(0, 25) gibt an, in welcem Bereich die vertikale Achse gezeichnet werden soll
- `rect` zeichnet ein Rechteck in eine vorgegebene Grafik. Die ersten beieden Zahlen sind die Koordinaten des Punktes links unten und die zweiten beiden Zahlen die Koordinaten des Punktes rechts oben.

Mit Hilfe der normierten Histogrammen lassen sich insbesondere solche Datenstämme vergleichen, die sehr unterschiedlich viele Messpunkte enthalten.

### 2.4.2 Boxplot

Ein Boxplot ist in Abbildung 2.7 schematisch dargestellt. Er besteht aus:

- einem Rechteck dessen Höhe vom empirischen 25%- und 75%-Quantil begrenzt wird (grüne Fläche)
- horizontalem Strich in der Box für den Median (schwarz)
- *whiskers*, blaue Linien, die zum kleinsten und grössten "normalen" Beobachtung führen (normal heisst *höchstens* 1.5 mal die Quartilsdifferenz von oberen und unteren Quartil)
- kleine roten Kreisen, die Ausreisser, welche ausserhalb der normalen Beobachtungen liegen

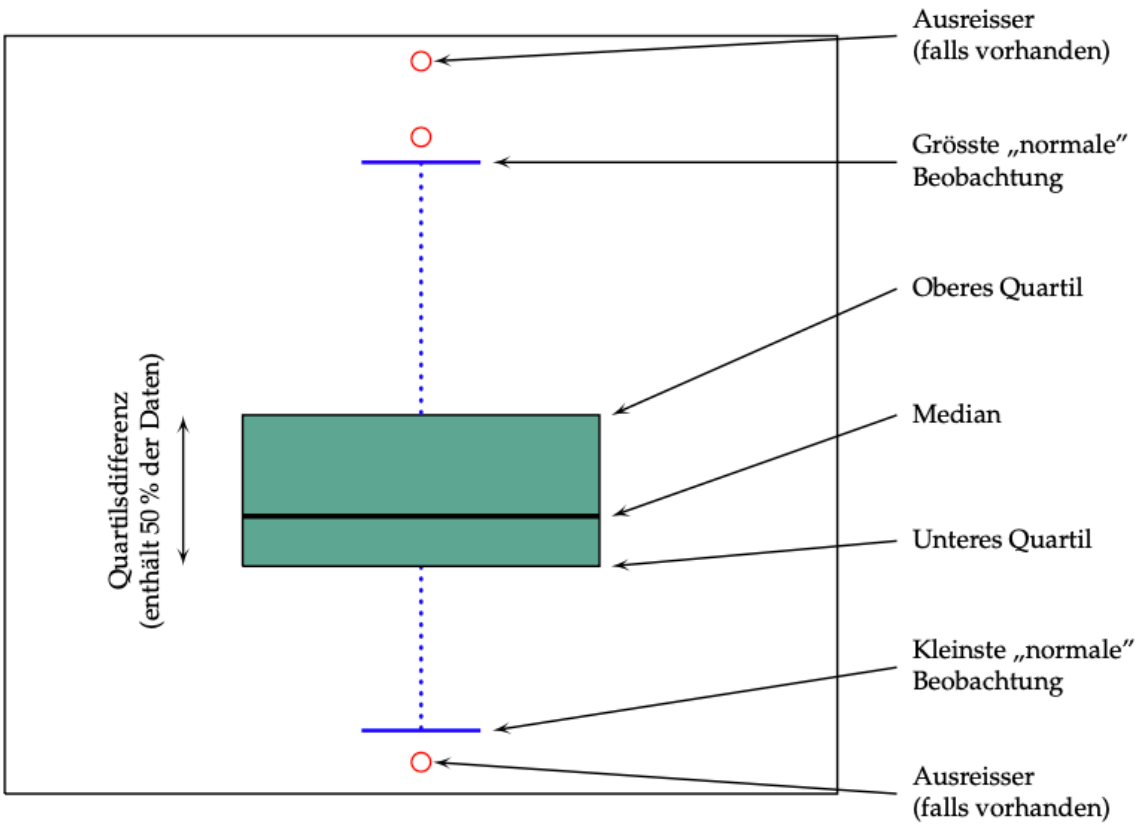


Abbildung 2.7: Schematischer Aufbau eines Boxplotes

#### 2.4.2.1 Boxplot in R

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

Boxplotte sind vorallem dann geeignet, wenn die Verteilung der Daten in verschiedenen Gruppen (versch. Versuchsbedingungen) verglichen werden sollen.

# 3 Deskriptive Statistik - Zweidimensionale Daten

Bei zweidimensionalen Daten werden an *einem* Versuchobjekt jeweils *zwei* verschiedene Grössen ermittelt. Als Beispiel dient uns das *Versuchsobjekt* Mensch mit den Messungen zu der *Körpergrösse* und *Körpergewicht*.

## 3.1 Streudiagramme

Zweidimensionale Daten werden häufig mit *Streudiagrammen* (Scatterplot) dargestellt. Dabei werden die beiden Messungen einer Versuchseinheit als *Koordinaten* in einem Korrdinatensystem interpretiert und dargestellt. Sind die Daten in dieser Form gegeben, interessieren wir uns in erster Linie für die *Zusammenhänge* und *Abhängigkeiten* zwischen den beiden Variablen.

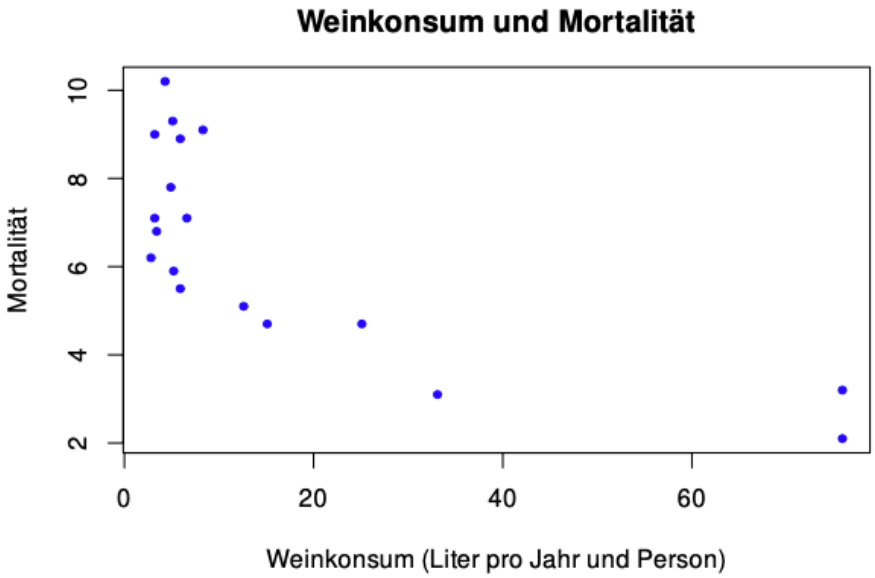


Abbildung 3.1: Streudiagramm für die Mortalität und Weinkonsum in 18 Länder

Dabei ist ein Punkt als $(x_1, y_1)$ mit den Grössen $x = Weinkonsum$ und $y = Mortalitaet$ zu interpretieren.

- Einfluss auf andere Körperorgane wird hier nicht berücksichtigt
- *Kausaler* Zusammenhang muss nicht zwingend zwischen den beiden Grössen vorhanden sein
- Die Zuweisung Grösse und X/Y-Achse könnte auch vertauscht werden. Die Entscheidung hängt von der Problemstellung ab
- Die Punkte im Streudiagramm werden auch als *Punktwolke* bezeichnet

### 3.1.1 Streudiagramm in R

```
1 wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9, 6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9)
2 mort <- c(6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5, 7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1)
3 plot(wein,
```

```
    mort,
    xlab = "Weinkonsum (Liter pro Jahr und Person)",
    ylab = "Mortalität",
    main = "Zusammenhang zwischen Weinkonsum und Mortalität", pch = 20,
    col = "blue"
)
```

## 3.2 Abhängigkeit und Kausalität

Bei Streudiagrammen müssen wir aufpassen, dass *Abhängigkeit* und *Kausalität* nicht miteinander verwechselt werden. Nur weil eine Gesetzmässigkeit vorhanden ist, heisst das nicht, dass diese Gesetzmässigkeit auch kausal erkärt werden kann. Man muss sich *bewusst* sein, auf *welchen Daten* das Streudiagramm basiert. Man soll sich *nie* blindlings auf Grafiken verlassen. Die Daten müssen auf anderen Weg auf einen kausalen Zusammenhang untersucht werden.

## 3.3 Einfache lineare Regression

Weil wir wissen möchten, *wie* sich Daten verhalten, versuchen wir einem Muster ein Form zu geben. Dies kann eine Gerade sein. Die Beschreibung dieser Form geschieht in der Sprache der Mathematik. Dabei wird auch von einer *Modellierung* der Daten gesprochen.

Liegt ein Modell vor, können wir *Vorhersagen* machen. Dieses Modell erlaubt uns, Daten die nicht exakt *gemessen* vorliegen abzuschätzen. Als Beispiel die Wettervorhersage oder den Preis einer Ware bei x-Stücken.

### 3.3.1 Methode der kleinsten Quadrate

Wie finden wir nun die Gerade die *möglichst gut* zu allen Punkten passt? Dazu verwenden wir den Begriff **Residuum**. Ein *Residuum* $r_i$ ist die vertikale Differenz zwischen einem Datenpunkt $(x_i, y_i)$ und dem Punkt $(x_i, a+bx_i)$ auf der gesuchten Geraden:

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

Bei der Methode der kleinsten Quadrate werden die Summen der *Quadrate der Abweichungen* aufsummiert $r_1^2 + r_2^2 + ... + r_n^2 = \sum_i r_i^2$. Die Parameter $a$ und $b$ werden so gewählt, dass die Summe minimal wird. Eine Gerade pass also dann am besten zu den Punkten im Streudiagramm, wenn die Summe der Quadrate der vertikalen Abweichungen minimal ist (Optimierungsproblem).

#### 3.3.1.1 Gerade mit R

```
seite <- c(seq(50, 500, 50))
preis <- c(6.4, 9.5, 15.6, 15.1, 17.8,23.4, 23.4, 22.5, 26.1, 29.1)
plot(seite,
    preis,
    xlab = "Seitenzahl",
    ylab = "Buchpreis",
    pch = 16,
    col = "blue"
)
```

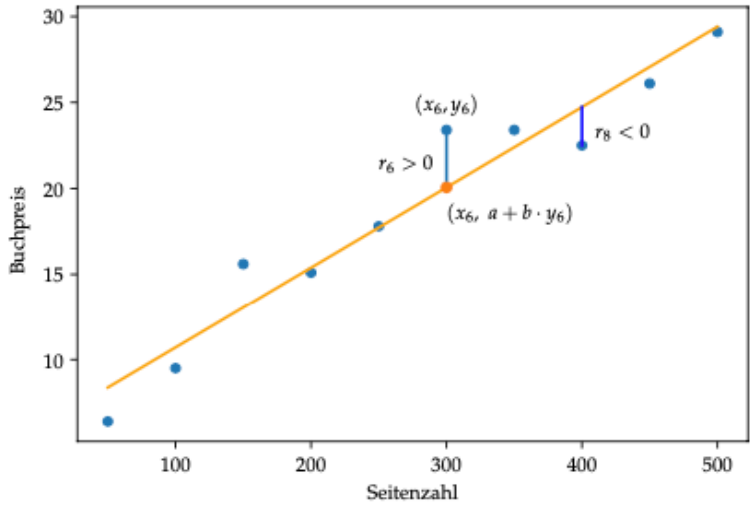


Abbildung 3.2: Residuen für Buchpreis in Abhängigkeit der Seitenanzahl

```r
abline(lm(preis ~ seite), col = "orange")

lm(preis ~ seite)
# Call:
# lm(formula = preis ~ seite)
#
# Coefficients:
# (Intercept)        seite
#     6.04000      0.04673
# y = 6.04 + 0.047x
```

- lm() steht für *linear model*
- Mit lm(y ~ x) passt R ein Modell der Form $y = a + bx$ an die Daten an
- Die Variabeln sind verglichen mit dem plot(x,y) vertauscht
- Die Gerade wird *Regressionsgerade* genannt

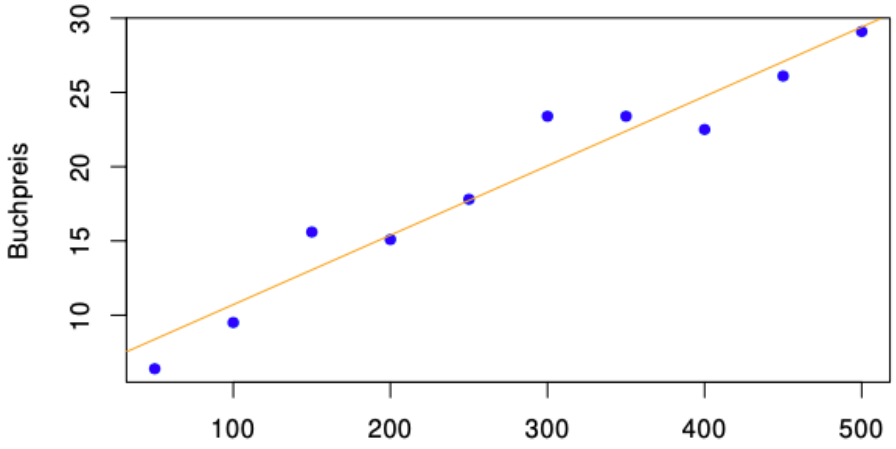


Abbildung 3.3: Streudiagramm mit Regressionsgeraden aus obigem R Code

#### 3.3.1.2 Extrapolationen

Extrapolationen sind Vorhersagen der $y$-Werte des Modelles, die *ausserhalb* des Bereiches ($x$) liegen, womit das Modell erstellt wurde. Für Extrapolationen, die weit ausserhalb des gültigen Bereichs liegen, können die Vorhersagen problematisch wenn nicht sogar sinnlos werden.

#### 3.3.1.3 Interpolationen

Interpolationen sind Vorhersagen der $y$-Werte des Modelles, die *innerhalb* des Bereiches liegt, womit das Modell erstellt wurde. Die Interpolationen sind unproblematischer, aber auch nur *begrenzt* gültig.

### 3.3.2 Empirische Korrelation

Die *empirische Korrelation* ($r$ als Kennzahl oder auch $\hat{p}$ ist die quantitative Zusammenfassung der *linearen* Abhängigkeit von zwei Grössen. Es ist eine dimensionslose Zahle zwischen $-1$ und 1 und misst die Stärke und die Richtung der *linearen Abhängigkeit* zwischen den Daten $x$ und $y$. *Wichtig:* Der Korrelationskoeffizient misst (erkennt) nur den linearen Zusammenhang!

- Ist $r = +1$, dann liegen Punkte auf einer steigenden Geraden ($y = a + bx$ mit $b > 0$)
- Ist $r = -1$, dann liegen die Punkte auf einer fallenden Geraden ($y = a + bx$ mit $b < 0$)
- Sind $x$ und $y$ unabhängig (es besteht kein Zusammenhang), so ist $r = 0$. Die Umkehrung gilt Allgemein aber nicht!

#### 3.3.2.1 Empirische Korreltation in R

```
1 cor(seite, preis)
2 # [1] 0.968112
```

Dieser Wert liegt sehr nahe bei 1 und somit besteht ein *enger* linearen Zusammenhang. Dazu ist der Wert positiv, was einem "je mehr, desto mehr" entspricht.

# 4 Wahrscheinlichkeit

## 4.1 Wahrscheinlichkeitsmodelle

Wir verwenden oft Modelle um die Wahrscheinlichkeit zu bestimmen. Dazu treffen wir Annahmen (z.B. dass ein Würfel fair ist) die dann im Modell umgesetzt/berechnet werden. Mit Hilfe des Modells können wir dann auch untersuchen, ob dieser Würfel fair ist. Wenn wir einen wiederholt werfen und oft die Zahl 2 vorkommt, können wir annehmen, dass der Würfel nicht fair ist.

### 4.1.1 Definition Wahrscheinlichkeitsmodelle

Wir betrachten *Zufallsexperimente*, bei denen der Ausgang *nicht exakt* vorhersagbar ist (z.B. # Anrufe in Callcenter). Das Wahrscheinlichkeitsmodell beschreibt grob welche Ergebnisse möglich sind und beinhaltet die Wahrscheinlichkeiten, wie die Ergebnisse eintreten können.

Das Wahrscheinlichkeitsmodell besteht aus folgenden Komponenten:

- *Grundraum* $\Omega$, der aus der Menge der *Elementarereignissen* $\omega$ besteht
- *Elementarerignisse* sind mögliche Ergebnisse oder Ausgänge des Experiments, die alle zusammen den Grundraum bilden.
- *Ergebnisse A, B, C,…* als Teilmengen von $\Omega$
- *Wahrscheinlichkeiten P*, die zu den Ereignissen $A, B, C, ...$ gehören

#### 4.1.1.1 Grundraum, Elementarereignisse

Bei einem Experiment wird aus dem Grundraum *ein* Elementarereignis *zufällig* gewählt. Als Beispiel das Würfeln. Grundraum gegeben durch

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

Element $\omega$ ist ein Elementarereignis und bedeutet, dass bei Würfeln die Zahl 2 geworfen wird.

#### 4.1.1.2 Ereignis

Unter einem Ereignis versteht man eine Teilmenge von $\Omega$. Das Ereignis $A$ bedeutet, dass das Ergebnis $\omega$ des Experiments zu $A$ gehört.

Beispiel für ein Ereignis $A$: “eine ungerade Augenzahl würfeln”, dann ist $A = \{1, 3, 5\}$ und tritt ein, wenn eine der drei Zahlen gewürfelt wird. Eine Leere Menge ist auch ein Ereignis $\{\} \subset \Omega$.

## 4.2 Disjunkte Ereignisse

Zwei Ereignisse $A$ und $B$ heissen *disjunkt*, wenn sich $A$ und $B$ gegenseitig ausschliessen und nicht gemeinsam eintreten können. Dann gilt

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

und ist somit unmöglich.

## 4.3 Axiome und Rechenregeln der Wahscheinlichkeitsrechnung

Die Wahrscheinlichkeitsrechnung baut auf die folgenden drei Grundregeln (Kolmogorov Axiome):

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

### 4.3.1 Rechenregeln

Für dieses Modul relevante Rechenregeln:

1. $P(\overline{A} = 1 - P(A)$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Die zweite Regel für diskjunkte Ereignisse. Die Schnittemenge wird doppelt gezählt, weshalb wir diese einmal abziehen müssen.

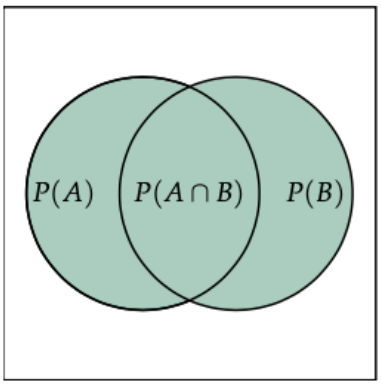


Abbildung 4.1: Wahrscheinlichkeit für nicht disjunkte Ereignisse

## 4.4 Diskrete Wahrscheinlichkeit

Mit diskret sind endliche und unendliche Mengen gemeint, welche ganzzahlige Elemente in $\Omega$ enthalten.

## 4.5 Laplace Wahrscheinlichkeit

Beim *Modell von Laplace* wird für jedes Elementarereignis die gleiche Wahrscheinlichkeit angenommen. Um diese Wahrscheinlichkeit zu berechnen zählen wir die Anzahl der *günstigen* Elemtentarereignisse durch die Anzahl der *möglichen* Elementarereignisse. Wenn alle Ereignisse $E$ gleich wahrscheinlich sind, ist das Eintreten des Ereignisses $E$ nach dem Laplace-Modell:

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

## 4.6 Der Begriff der Unabhängigkeit

Hat der Ausgang von Ereignis $A$ keinen Einfluss auf den Ausgang des Ereignisses $B$, sind die Ereignisse Ereignisse $A$ und $B$ stochastisch unabhängig. Dann gilt

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

Als Beispiel: $A$ sei mit einem fairen Würfel eine eis oder zwei zu würfeln und Ereignis $B$ sei Kopf beim Werfen einer fairen Münze. Weil die beiden Ereignisse keinen Einfluss aufeinander haben gilt obige Formel.

Sind Ereignisse *nicht* stochastisch unabhängig, gibt es keine allgemeine Formel für die Brechnung der Wahrscheinlichkeit von zwei Ereignissen.

# 5 Zufallsvariable

Bei einem Zufallsexperiment mit dem Grundraum $\Omega$ ordnen wir mit der *Funktion* $X$ jedem Elementarereignis $\omega$ von $\Omega$ einen *Zahlwert* zu. Die Funktion $X$ wird als *Zufallsvariable* bezeichnet und ordnet damit jedem Element *eine Zahl* zu.

Eine Wertemenge bezeichnet die Werte, welche die Zufallsvariable annehmen kann.

**Bermerkungen**

- die *Zufallsvariable wird mit einem Grossbuchstaben bezeichnet ($X$)
- der entsprechende *Kleinbuchstabe* $x$ stellt einen konkreten Wert dar, den die Zufallsvariable annehmen kann (die Zahl)
- für das Ereignis, welches $X$ annimmt, schreiben wir $X = x$
- bei der Zufallsvariable ist nicht die Funktion $X$ zufällig, sondern das Argument $\omega$. Je nach Ausgang erhalten wir einen anderen Wert $X(\omega) = x$
- $x$ wird auch als *Realisierung* der Zufallsvariable $X$ bezeichnet

## 5.1 Wahrscheinlichkeitsverteilung einer Zufallsvariable

Berechnen wir für *jede* Realisierung einer Zufallsvariable die zugehörige Eintretenswahrscheinlichkeit, so bilden alle diese Wahrscheinlichkeiten zusammen die *Wahrscheinlichkeitsverteilung* dieser Zufallsvariablen. Dabei gilt

$$P(X = x_1) + P(X = x_2) + ... + P(X = x_n) = 1$$

oder

$$\sum_{i=1}^{n} P(X = x_n) = 1$$

## 5.2 Kennzahlen einer Verteilung

Eine beliebige *diskrete* Verteilung kann durch zwei Kennzahlen, den *Erwartungswert* $E(X)$ und die *Standardabweichung* $\sigma(X)$. Der Erwartungswert beschreibt die mittlere Lage der Verteilung:

$$E(X) = x_1 * P(X = x_1) + x_2 * P(X = x_2) + ... + x_n * P(X = x_n) = \sum_{i=1}^{n} x_i * P(X = x_i)$$

Die Standardabweichung oder *Varianz* beschreibt die Streuung der Verteilung.

$$Var(X) \sum_{i=1}^{n} (x_n - (E(X))^2 * P(X = x_n)$$

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

### 5.2.1 Standardabweichung mit R

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

Beispiel eines nicht-fairen Würfels auch mit R berechnet:

```
x <- 1 : 6
p <- c(4, 2, 1, 3, 1, 1) / 12
E_X <- sum(x * p)
E_X
## [1] 2.833333
var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)
sd_X
## [1] 1.674979
```

Der Erwartungswert ist 2.8333 und die Standardabweichung ist 1.675 (die Werte weichen im "Durchschnitt" so viel ab).

## 5.3 Unterschied empirischer und theoretischer Kennzahlen

### 5.3.1 Unterschied Mittelwert und Erwartungswert

Der arthmetische Mittelwert $\overline{x}$ wird aus *konkreten* Daten gewonnen. Wir haben also Messwerte $x_1, ..., x_n$ und können $\overline{x}$ berechnen. Der Erwartungswert $\sigma_X$ ist ein *theoretischer* Wert, der sich aus dem Modell der Wahrscheinlichkeitsverteilung ergibt.

Die Hoffnung ist, dass sich das arithmetische Mittel für immer mehr Versuche an den theoretischen Wert annähert, sofern sich die konkreten Daten der Wahrscheinlichkeitsverteilung von $X$ folgen.

Derselbe Unterschied wie für Mittelwert und Erwartungswert gilt auch für die *empirische Standardabweichung* $s_X$ und die *Standardabweichung* $\sigma_X$

# 6 Bedingte Wahrscheinlichkeit

Die *bedingte Wahrscheinlichkeit* betrachtet nicht die gesamte Grundmenge, sondern nur einen Teil davon. Die neue Grundmenge in der Formel ist die Variable nach dem Längsstrich. Die bedingte Wahrscheinlichkeit bedeutet die Wahrscheinlichkeit von A unter der Voraussetung von B.

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

Das wir wissen, dass B schon eingetreten ist, haben wir einen neuen Grundraum $\Omega' = B$. Damit müssen wir nur noch den Teil anschauen, der auch in $B$ auftritt (daher $A \cap B$). Dies muss dann noch in Relation zur Wahrscheinlichkeit von $B$ (neue Grundmenge) gesetzt werden.

## 6.1 Rechenregeln der bedingten Wahrscheinlichkeit

![Rechenregeln der bedingten Wahrscheinlichkeit}(rechenregeln_bedWahrscheinlichkeit.png){width=60%}

## 6.2 Bayes Theorem und totale Wahrscheinlichkeit

### 6.2.1 Bayes' Theorem

Das Theorem ist oft sehr nützlich. Es erlaubt uns die Wahrscheinlichkeit $P(A|B)$ zu berechnen, falls $P(B|A)$.

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

### 6.2.2 Totale Wahrscheinlichkeit

Bei der totalen Wahrscheinlichkeit wird erst eine Menge $A$ in Menhen $A_1, ..., A_k$ unterteilt, die untereinander keine Schnittmenge haben und zusammen (Vereinigung) die ganze Menge $A$ bilden. Dies wird *Partitionierung* genannt.

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

# 7 Normalverteilung

## 7.1 Stetige Zufallsvariable und Wahrscheinlichkeitsverteilungen

Der Unterschied einer *stetigen* und *diskreten* Zufallsvariable besteht darin, dass die stetige *jeden* Wert eines bestimmten Bereiches annehmen. Diskrete hingegeben aus einer definierten und endlicher Menge (z.B. von 0 bis 100, jeweils ganze Zahlen). Wichtig auch, es kann kein Wert zwischen zwei Werten aus einer Wertemenge ausgewählt werden. Dabei gilt:

- die Variable $X$ ist eine Funktion
- die Variable $x$ ist ein konkreter Wert (*Realisierung*) von $X$

### 7.1.1 Stetige Verteilungen

Stetig sind die Verteilungen wenn keine Lücken in einem Bereich vorhanden sind. Die jeweilige Wahrscheinlichkeit ist $P(X = x) = 0$. Diese werden *Punktwahrscheinlichkeiten* genannt. Diese Wahrscheinlichkeit bringt uns aber nicht weiter. Weshalb wird die Wahrscheinlichkeit zwischen zwei Punkten berechnen. Dazu wird das Konzept der *Wahrscheinlichkeitsdichte* angewendet.

### 7.1.2 Wahrscheinlichkeitsdichte

Wahrscheinlichkeitsdichten können unter der Berücksichtigung folgender Einschränkungen fast beliebig aussehen. Relevant für uns sind aber lediglich die Normalverteilung und die dazu verwandte $t$-Verteilung.

#### 7.1.2.1 Eigenschaften Wahrscheinlichkeitsdichte

Für eine Wahrscheinlichkeitsdichte $f(x)$ gelten folgende Eigenschaften:

- es gilt $f(x) \geq 0$ -> Kurve auf oder oberhalb der x-Achse
- die Wahrscheinlichkeit $P(a < X \leq b)$ entspricht der Fläche zwischen $a$ und $b$ unter $f(x)$
- die gesamte Fläche unter der Kurve ist 1 -> W'keit dass *irgendein* Wert gemessen wird.

[Dichte einer Zufallsvariable und der Wahrscheinlichkeit in ein Intervall $(a,b],$ zu fallen (grüne Fläche).](dichte_zv.png){width=60%}

Bei stetigen Wahrscheinlichkeitsverteilungen entsprechen Wahrscheinlichkeiten den Flächen unter der Dichtefunktion.

### 7.1.3 Quantile

Bei stetigen Verteilungen ist das $\alpha$-Quantil $q_\alpha$ derjenige Wert, wo die Fläche unter der Dichtefunktion von $-\infty$ bis $q_\alpha$ gerade $\alpha$ entspricht. Das 50%-Quantil heisst der *Median*.

Dies bedeuetet, dass 75% einer Menge maximal diese Messgrösse erreichen.

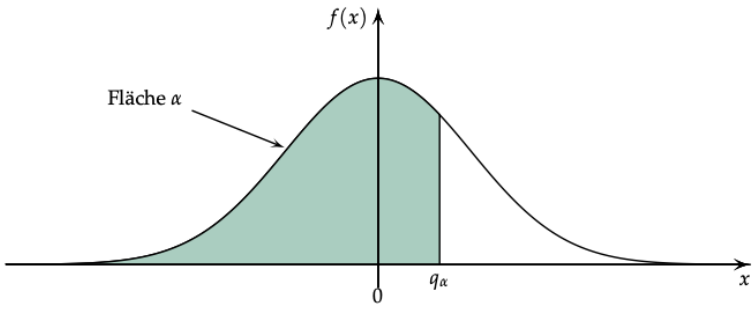


Abbildung 7.1: Quantil $q_\alpha$ anhand der Dichte $f(x)$ für $\alpha = 0.75$

### 7.1.4 Kennzahlen von stetigen Verteilungen

Der Erwartungswert $E(X)$ und die Standardabweichung $\sigma_X$ werden gleich wie im diskreten Fall interpertiert:

- $E(X)$ beschreibt die mittlere Lage der Verteilung
- $\sigma_X$ beschreibt die Streuung der Verteilung um den Erwartungswert

Im stetigen Fall, sind die beiden Funktionen aber mit Integralen statt Summen definiert.

## 7.2 Normalverteilung (Gaussverteilung)

Die *Normalverteilung* ist die häufigste Verteilung für Messwerte und spielt vor allem für Durchschnitte von Messwerten eine wichtige Rolle. Wichtigte Wahrscheinlichkeitsverteilung in der Statistik.

**Normalverteilung**

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

Es ist folgende Schreibweise für die Verteilung einer normalverteilten Zufallsvariable $X$ mit Parameter $\mu$ und $\sigma$ zu verwenden:

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

- $E(X) = \mu$
- $Var(X) = \sigma^2$
- $\sigma_X = \sigma$

Die bedeutet, dass die Parameter $\mu$ und $\sigma^2$ eine natürliche Interpretation als Erwartungswert und Varianz einer Verteilung haben.

### 7.2.1 Graphische Darstellung der Normalverteilung

Alle Normalverteilungskurven haben folgende Eigenschaften:

- Die Kurven sehen aus wie "Glocken", weswegen man auch von *Glockenkurven* spricht
- Die Wahrscheinlichkeitsdichtefunktion der Normalverteilung ist symmetrisch um den Erwartungswert $\mu$
- Der Parameter $\mu$ verschiebt den Graphen um $\mu$-Einheiten nach links ($\mu < 0$) oder nach rechts ($\mu > 0$)
- Je grösser $\sigma$, desto flacher bzw. breiter wird die Dichtekurve. Je näher bei 0, wird die Kurve umso spitzer.

Der Grund dafür, weil $\sigma$ die Streuung um den Erwartungswert $\mu$ angibt. Je grösser $\sigma$ um so mehr sind die Werte um den Erwartungswert $\mu$ verteilt, die Kurve wird also breiter. Ist $\sigma$ nahe bei 0, so weichen die Werte wenig von $\mu$ ab, die Kurve wird schmaler und höher.

#### 7.2.1.1 Wahrscheinlichkeiten mit R

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

Dieser Befehl ermittelt die Fläche (Wahrscheinlichkeit von $-\infty$ bis $q = 130$ unter der Normalverteilungskurve mit $\mu = 100$ und $\sigma = 15$

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

qnorm() bestimmt die Quantile für die Normalverteilung. Bei diesem Beispiel haben wir den Wertebereich von 95% aller Beobachtungen eruiert. Diese liegen zwischen 70 bis 130.

> Ist eine Zufallsvariable normalverteilt, so liegen etwa zwei Drittel aller Messerte etwa eine Standardabweichung um den Erwartungswert.

Für *alle* Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$ gilt, dass die Wahrscheinlichkeit, dass eine Beobachtung höchstens *eine* Standardabweichung vom Erwartungswert abweicht, ist etwa $\frac{2}{3}$:

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

Die Wahrscheinlichkeit, dass eine Beobachtung höchstens *zwei* Standardabweichungen vom Erwartungswert abweicht ist:

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

Aus diesen beiden Wahrscheinlichkeiten lassen sich auch als Flächen interpretieren. Die Fläche der Normalverteilung über dem Intervall $[\mu - \sigma, \mu + \sigma]$ ist ca. $\frac{2}{3}$. Die Flöche über dem Intervall $[\mu - 2\sigma, \mu + 2\sigma]$ ist ca. 0.95.

### 7.2.2 Standardnormalverteilung

Die Normalverteilung $\mathcal{N}(0, 1)$ mit Mittelwert 0 und Standardweichung 1 heisst *Standardnormalverteilung*. Falls $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist die standardisierte Zufallsvariable wieder normalverteilt, hat nun aber den Erwartungswert 0 und die Varianz 1. Man erhält also die Standardnormalverteilung:

$$Z = \frac{Z - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

# 8 Durchschnitte und Summen von Zufallsvariablen

Bisher haben wir lediglich untersucht, wie *eine* Zufallsvariable verteilt ist. In Vielen Anwendungen haben wir es mit *mehreren* Zufallsvariablen zu tun. Überlicherweise messen wir die *gleiche* Grösse mhermals und untersuchen, wie die Summe und der Durchschnitt von *mehreren* Zufallsvariablen verteilt sind.

Die Messungen bezeichnen wir mit $x_1, x_2, ..., x_n$ und fassen diese als Realisierungen der Zufallsvariablen $X_1, X_2, ..., X_n$ auf. Dies ist die Zufallsvariable für das *arithmetische Mittel*. Das arithmetische Mittel $\bar{x}_n$ ist also eine Realisierung der Zufallsvariablen $\bar{X}_n$.

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{100} \sum_{i=1}^{n} X_i$$

Analoges gilt für die *Summe* $S_n$:

$$S_n = X_1 + ... + X_n = \sum_{i=1}^{n} X_i$$

## 8.1 Unabhängigkeit und i.i.d. Annahme

Wir treffen oft die Annahme, dass die Zufallsvariablen *unabhängig* voneinander sind. Es gibt keine gemeinsamen Faktoren, die den Ausgang der verschiedenen Messungen beeinflussen. Damit ist gemeint, dass eine Messung keinen Einfluss auf das Resultat der nachfolgenden Messung hat. Als Beispiel für *nicht" unabhängige Zufallsvariablen die Temperaturmessung an benachbarten Sommertage. Die Tage werden ähnliche Temperaturen haben nicht 28°C und am nächsten -5°C.

Wenn die Zufallsvariablen $X_1, ..., X_n$ unabhängig sind und alle *dieselbe* Verteilung haben, schreiben wir das kurz als

$$X_1, ..., X_n$$

i.i.d.

Die Abkürzung i.i.d. steht für:

**i**ndependent, **i**dentically, **d**istributed

> Sind Zufallsvariablen i.i.d., so wird dasselbe unter den gleichen Bedingungen gemessen.

Die Unabhängigkeit spielt insofern eine Rolle bei den *Regeln für Erwartungswerte und Varianzen* von Summen. Die Beziehung

$$E(X_1 + X_2) = E(X_1) + E(E_2)$$

gilt immer,

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$$

jedoch nur wenn $X_1$ und $X_2$ unabhängig sind.

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

Wir nehmen an dass $X_1, ..., X_n$ i.i.d. Wegen dem zweiten "i" in i.i.d. hat *jedes* $X_i$ dieselbe Verteilung und dieselben Kennzahlen.

### 8.2.1 Varianz und Standardabweichung der Summe

Die Varianz und Standardabweichung nimmt mit zunehmender Anzahl Würfen zu. Je mehr Ereignisse, umso grösser wird der Wertebereich. Die Summen verteilen sich auf mehr Zahlen und damit nimmt auch die Streuung zu. Das Gesetzt für die Varianz und Standardabweichung der Summe lautet:

$$Var(S_n) = nVar(S_n)$$

und für die Standardabweichung gilt

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

### 8.2.2 Erwartungswert des Durchschnittes

Da wir gleichbleibende Ergebnisse erwarten, ob nun 2 oder 50 Experimente gemacht werden, entspricht der Durchschnitt in etwa dem Erwartungswert.

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

Für die Varianz und die Standardabweichung des Durchschnitts gilt folgendes Gesetz:

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

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

Allgemein gilt:

**Kennzahlen von $S_n$**

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

**Kennzahlen von $\overline{X}_n$**

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

Die Standardabweichung von $\overline{\mathrm{X}}_n$ heisst auch der *Standardfehler* des arithmetischen Mittels.

Abbildung 8.1: Regeln für die Kennzahlen von $S_n$ und $\bar{X}_n$

Die Standardabweichung der Summe wächst mit wachsendem $n$, aber langsamer als die Anzahl Beobachtungen $n$. Der Erwartungswert von $\bar{X}_n$ ist gleich demjenigen einer einzelnen Zufallsvariable $X_i$, die *Streeung nimmt jedoch mit wachsendem n ab*.

**Gesetz der grossen Zahlen**

Für $n \to \infty$ geht die Streuung gegen null. Es gilt das **Gesetz der grossen Zahlen:** Falls $X_1, \ldots, X_n$ i.i.d., dann

$$\overline{X}_n \longrightarrow \mu \quad \text{für} \quad n \to \infty$$

Abbildung 8.2: Gesetzt der grossen Zahlen

**Standardfehler**

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

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

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

Abbildung 8.3: Standardfehler des arithmetischen Mittels

## 8.3 Verteilungen von $S_n$ und $\bar{X}_n$

Die Verteilung der *Mittelwerte* $\bar{X}_n$ (oder auch der Summen) nähert sich mit wachsendem $n$ einer Normalverteilung an. Sind die $X_i$'s i.i.d., dann gilt der *Zentrale Grenzwertsatz*.

**Zentraler Grenzwertsatz**

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

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

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

Abbildung 8.4: Zentraler Grenzwertsatz

wobei die Approximation im allgemeinen besser wird mit grösserem $n$. Überdies ist die Approximation besser, je näher die Verteilung von $X_i$ bei der Normalverteilung $\mathcal{N}(\mu, \sigma_x^2)$ ist.

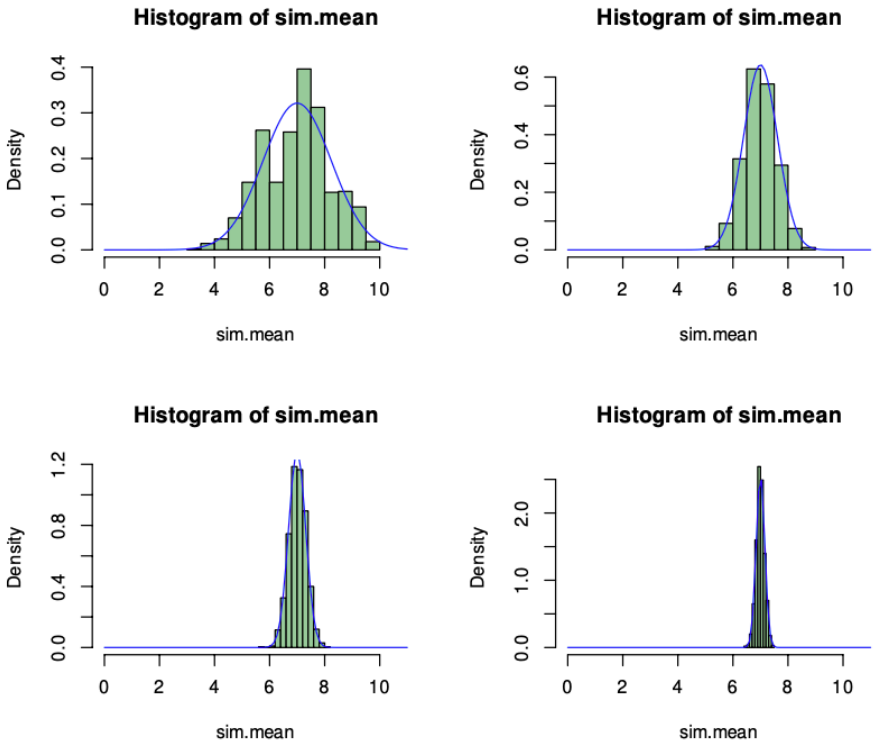


Abbildung 8.5: 4 Histogramme vom Durchschnitt von 16, 64, 256, 1024 Versuchen mit 1000 Ziehungen mit Dichtekurven

# 9 Hypothesentest für Messdaten

Mittels eines Hypothesentests wird eine Angabe (z.B Inhaltsmenge einer Pet-Flasche) überprüft ob diese wahr ist. Solche Tests sind stadardisiert und ist eine reproduzierbares Verfahren. Diese geben auch ein klares Kriterium an, wann ein Druchschnitt zu weit von einer Angabe entfernt ist. Der Hypothesentest ist *nie* ein Beweis das eine Angabe wahr ist oder falsch, sonder lediglich ob die Angabe mit einer gewissen Wahrscheinlichkeit stimmt oder nicht.

## 9.1 Statistische Tests und Vertrauensintervall für eine Stichprobe bei normalverteilten Daten

Messungen können wir als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_i$ betrachten und die Kennzahlen $E(X_i) = \mu$ und $Var(X_i) = \sigma_X^2$ erfassen. Typischerweise sind diese (und andere) Kennzahlen *unbekannt*.Trotzdem möchten wir eine Aussage über das wahre, aber eben unbekannte $\mu$ und $\sigma^2$ machen. Das Ziel ist es den Daten $\mu$ und $\sigma^2$ anzunähern und sprechen dabei von einer *Schätzung* der Parameter $\mu$ und $\sigma^2$. Geschätzte Werte werden mit einem ˆbezeichnet $\hat{\mu}$. Für die (Punkt-)Schätzungen für den Erwartungswert und Varianz gilt:

$$\hat{\mu} = \bar{X_n} = \frac{X_1 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}_X^2 = \frac{(X_1 - \bar{X_n})^2 + ... + (X_1 - \bar{X_n})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2$$

Beachte, dass $\hat{\mu}$ und $\hat{\sigma}_X^2$ selbst Zufallsvariablen sind und für jede neue Messung sich neue $\hat{\mu}$ und $\hat{\sigma}_X^2$ ergeben. Obwohl im Allgmeinen $\hat{\mu} \neq \mu$ und $\hat{\sigma}_X^2 \approx \sigma_X^2$ ist die Hoffnung dass $\hat{\mu} \neq \mu$ und $\hat{\sigma}_X^2 \approx \sigma_X^2$ und damit annähert.

### 9.1.1 Ziel des Hypothesentests

Das Ziel des Hypothesentests ist das feststellen ob der waare Mittelwert wahr ist. Entstehen bei den Schätzungen grössere Standardabweichungen muss die Angabe hinterfragt werden.

## 9.2 Hypothesentest

Hypothesentests sind ein wichtiges statistisches Mittel um zu entscheiden, ob eine Messreihe zu einer gewissen Grösse passt. Wir gehen davon aus, dass wir den wahren Mittelwert *nicht* kennen, gehen aber von einem *Idealwert* oder einem vermuteten Wert aus.

Unter der Annahme, dass die Daten normalverteilt sind, wird überprüft ob eine Messreihe, unter der Annahme von $\mu = a$ (der Mittelwert) warhscheinlich ist oder nicht.

### 9.2.1 Modell

Eine Anuahl Messwertel sind die Realisierung der Zufallsvariablen $X_1, X_2, ..., X_n$, wobei $X_i$ eine kontinuierliche Messgrösse ist. Dabei soll gelten:

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

### 9.2.2 Nullhypothese

$$H_0 : \mu = \mu_0 = a$$

Die Nullhypothese ist eine Annahme über den wahren Mittelwert. Diese Annahme wird mit dem Mittelwert $\hat{\mu}$ überprüft, ob er sich dem wahren Mittelwert annähert.

### 9.2.3 Alternativhypothese

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

oder

$$H_A < oder >$$

Wenn die Annahme nicht gleich $\mu$ ist

### 9.2.4 Teststatistik

Es wird getestet, ob diese Verteilung mit der Annahme $\mu = a$ gerechtfertigt ist. Oder Mathematisch: Die Verteilung der Teststatistik T unter Nullhypothese $H_0$

$$T = \bar{H_n}/mathcalN(\mu, \frac{\sigma^2}{n})$$

Ist die Wahrscheinlichkeit kleiner als 2.5% ist sie zu klein und der Mittelwert zu unwahrscheinlich, als dieser zur Aussgangsgrösse a passen könnte.

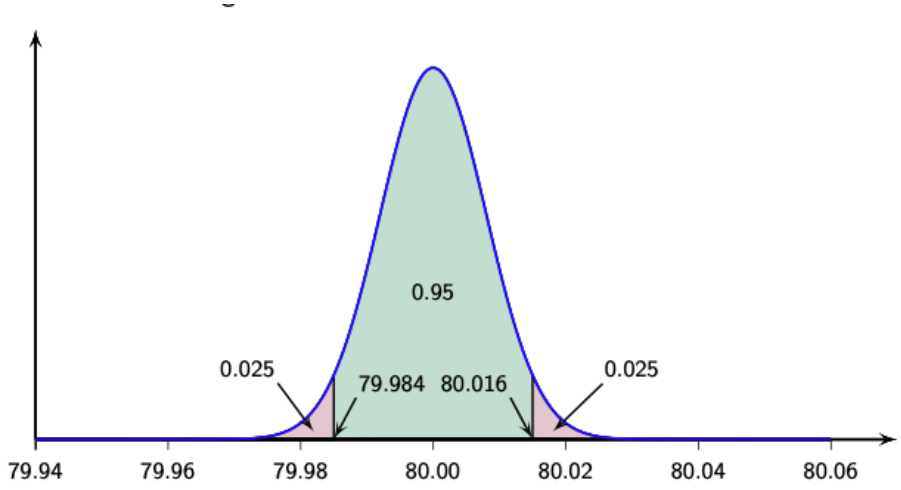


Abbildung 9.1: Normalverteilungskurve eines Hypothesentests

Die *Abmachung*, dass die Grenze 2.5% gilt, kommt daher, dass der symmetrische Teil um den Mittelwert 95% betragen soll.

### 9.2.5 Signifikanzniveau $\alpha$

Das Signifikanzniveau $\alpha$ gibt an, wie hoch das Risiko ist, eine falsche Entscheidung zu treffen. Normalerweise 0.05 bzw. 0.01.

### 9.2.6 Verwerfungsbereich

Liegt der gemessene Mittelwert im roten Bereich (der Abbildung), so zweifelt man an der Nullhypothese und *verwerfen* diese. Diese werden im einem Intervall angegeben:

$$K = (-\infty, a - \alpha] \cup [\, a + \alpha, \infty)$$

### 9.2.7 p-Wert

Der P-Wert ist die Wahrscheinlichkeit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten wie das aktuell beobachtete. Damit wird angedeutet, wie extrem das Ergebnis ist. Je kleiner der p-Wert desto mehr spricht das Ergebnis gegen die Nullhypothese.

- 0: passt gar nicht
- 1: passt sehr gut

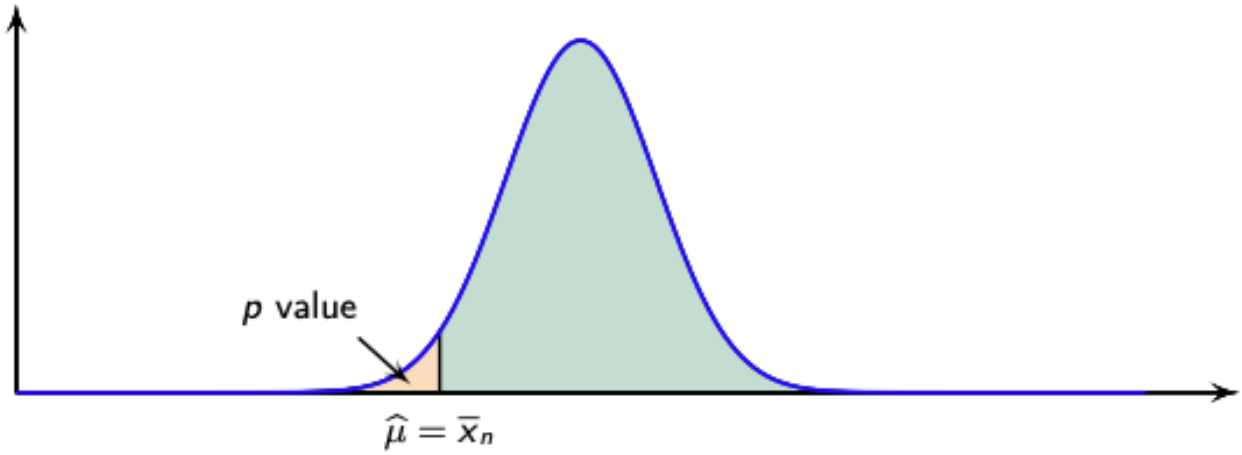


Abbildung 9.2: Wahrscheinlichkeit unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein extremeres zu erhalten

### 9.2.8 p-Wert und Statistischer Test

Bei einem vorgegebenen Signifikanzniveaus $\alpha$ gilt aufgrund der Definition des p-Werts für einen einseitigen Test:

- Verwerfe $H_0$ falls p-Wert $\leq \alpha$
- Belasse $H_0$ falls p-Wert $> \alpha$

Computerprogramme liefern den Testentscheid nur mit p-Wert und *immer* auf Signifikanzniveau an.

#### 9.2.8.1 Signifikanz

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

## 9.3 t-Test

Entgegen dem bisherigen Verfahren (z-Test) wo die Standardabweichung bekannt ist, setzt der t-Test keine Standardabweichung voraus. Dies ist auch praktisch kaum der Falls das SD vorliegt.

### 9.3.1 t-Verteiltung

Die Verteilung der Teststatistik beim t-Test unter der Nullhypothese

$$H_0 : \mu = \mu_0$$

ist gegeben durch

$$T = \bar{X_n} \sim T_{n-1}(\mu, \frac{\hat{\sigma_X}}{\sqrt{n}})$$

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

Die Normalverteilung wird also durch t-Verteilung ersetzt. Gleicht aber Normalverteilung, ist aber flacher, wegen der grösseren Unsicherheit. Dies hängt von der Anzahl Beobachtungen ab.

# 10 Vertrauensintervall, Zweistichprobentest und Wilcoxon-Test

## 10.1 Vertrauensintervall für $\mu$

Das Intervall gibt an, wo der wahre Mittelwert $\mu$ mit einer bestimmten 95%-Wahrscheinlichkeit liegt. Bei der Bestimmung des Verwerfungsbereiches beim $z$-Test gehen wir von einem wahren (aber unbekannten) Wert $\mu$ aus mit einer bekannten Standardabweichung.

```
qnorm(p=c(0.025, 0.975), mean= 5, sd =2)
## 1.080072 8.919928
```

In obiger Ausgabe sehen wir das **Vertrauensintervall**. In diesem Intervall liegt mit bestimmter Wahrscheinlichkeit der **wahre** Mittelwert. Liegt nun $\bar{x}_n$ im Vertrauenintervall, wird $H_0$ nicht verworfen. Liegt der Mittelwert ausserhalb, wird $H_0$ verworfen.

Dies ist eine weitere Möglichkeit für einen Testentscheid. R gibt das Vertrauenintervall auch im $t$-Test als **confidence interval** aus. Dieses besagt, dass bei einem Signifikanzniveau von 5% das wahre $\mu$ zu 95% in diesem Intervall liegt. Je schmaler das Vertrauensintervall ist, umso sicherer sind wir, wo sich der wahre Mittelwert befindet. Bei einem grossen Intervall besteht eine grosse Unsicherheit wo das wahre $\mu$ liegt.

## 10.2 Der Wilcoxon-Test

Der Wilcoxon-Test ist eine Alternative zum $t$-Test, setzt dabei aber weniger voraus. Er wird vorallem bei **nicht-normalverteilten** Daten eingesetzt. Grundsätzlich hat er die grössere *Macht*. Macht ist die Wahrscheinlichkeit, dass die Nullhypothese richtigerweise verworfen wird. Einzige Voraussetzung bei einem Wolcoxon-Test ist dass die Verteilung unter der Nullhypothese *symmertrisch* bezüglich $\mu_0$ ist.

Der Test berechnte den $V$-Wert, der die sogenannte *Rangsumme* repräsentiert. Ist der $V$-Wert zu weit weg vom Median, wird die Nullhypothese verworfen, ansonsten beibehalten.

## 10.3 Statistische Tests bei zwei Stichproben

Wird ein Vergleich zwischen zwei Proben gemacht spricht man von gepaarten und ungepaarten Stichproben.

### 10.3.1 Gepaarte Stichproben

Um eine gepaarte Stichprobe zu sein müssen folgende Voraussetzungen gelten:

- beide Versuchsbedingungen müssen an derselben Versuchseinheit eingesetzt werden
- jeder Versuchseinheit aus der einen Gruppe kann genau eine Versuchseinheit aus der anderen Grupper zugeordnet werden
- die Stichprobengrösse $n$ ist für beide Versuchsbedingungen dieselbe
- $x_i$ und $y_i$ sind abhängig, weil die Werte von der gleichen Versuchseinheit kommen

#### 10.3.1.1 Beispiele gepaarte Stichproben

- Messung vor und nach einem Ereignis
- Zwei Messungen vom selben Punkt/Objekt (Zuordnung ist eindeutig)

#### 10.3.1.2 Statistischer Test für gepaarte Stichproben

Bei der Analyse arbeitet man mit den Differenzen innerhalb der Paare $d_i = x_i - y_i (i = 1, ..., n)$ welche wir als Zufallsvariablen $D_1, ..., D_n$ auffassen. **Kein** Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[D_i] = 0$. Falls die Daten normalverteilt sind eignet sich ein $t$-Test, sonst Wilcoxon-Test. Bei beiden Tests kann für eine gepaarte Stichprobe die Option `paired=TRUE` mitgegeben werden.

**Hinweise**

- man kann auf die Differenzen $d_i$ berechnen und dann den Test durchführen
- das erste Argument im Funktionsaufruf bezieht sich auf das **nachher**, das zweite auf die Messung **vorher**

### 10.3.2 Ungepaarte Stichproben

Oft kann nicht jede Messung einer Messung aus der zweiten Gruppe eindeutig zuordnen. Dann spricht man von einer **ungepaarten** Stichprobe. Im Allgemeinen ist die Anzahl der Messungen beider Gruppen unterschiedlich (muss aber nicht). Entscheidend ist dass $x_i$ und $y_i$ zu verschiedenen Versuchseinheiten gehören und unabhängig angenommen werden.

#### 10.3.2.1 Beispiele für ungepaarte Stichproben

- Messung wird nacheinander von zwei versch. Geräten gemacht (nicht gleichzeitig)
- verschieden lange Messreihen
- Zufällige Zuordnung in zwei Gruppen

## 10.4 Tests mit R bei zwei Stichproben

Folgende Parameter können bei $t$-Test oder Wilcoxon-Test mitgegeben werden:

- `x`: erste Messreihe (nachher)
- `y`: zweite Messreihe (vorher)
- `alternative=`: `"less"`, `"greater"` oder `"two.sided"` als Varianten, je nach Test
- `mu=`: gibt an welcher Unterschied in den Mittelwerten der beiden Gruppen getestet werden soll. Wenn Test prüfen soll ob Gruppenmittelwerte gleich, dann `mu=0`
- `paired=`: `TRUE` oder `FALSE` für gepaarte/ungepaarte Messreihen
- `conf.level=`: definiert Vertrauenintervall - default `0.05`

# 11 Lineare Regression