Grundkurs Statistik: Formeln und R-Funktionen

Lukas Stammler

2022-02-17

Contents

Kennzahlen **3**

Umfang . . . 3

Arithmetisches Mittel, Mittelwert . . . 3

Median . . . 4

Standardabweichung . . . 4

Grafiken **5**

Histogramm . . . 5

Boxplot . . . 6

Kreuztabelle, absolute Häufigkeiten . . . 6

Kreuztabelle, relative Häufigkeiten . . . 7

Balkendiagramm . . . 7

Wahrscheinlichkeiten **8**

Wahrscheinlichkeit . . . 8

Ereignis und Gegenereignis . . . 8

Bedingte Wahrscheinlichkeiten . . . 8

Unabhängigkeit . . . 8

Normalverteilung **8**

68-95-99.7-Regel . . . 9

z-Wert . . . 9

Perzentilen in R berechnen . . . 9

QQ-Plot . . . 10

Binomialverteilung **11**

Erwartungswert (Mittelwert) der Binomialverteilung . . . 11

Standardabweichung der Binomialverteilung . . . 11

Bedingungen für Binomialverteilung . . . 11

Wahrscheinlichkeiten der Binomialverteilung . . . 11

Normalapproximation . . . 12

Grundlagen der Inferenzstatstik **12**

Zentraler Grenzwertsatz . . . 12

Konfidenzintervalle . . . 13

Zuverlässigkeit vs. Präzision . . . 14

Hypothesentest für einen Mittelwert . . . 15

Inferenz für quantitative Daten **17**

Hypothesentests für gepaarte Mittelwerte . . . 17

Hypothesentest für unabhängige Mittelwerte . . . 18

T-Verteilung und t-Tests **19**

Inferenz für einen Mittelwert . . . 19

Inferenz für zwei Mittelwerte . . . 21

Nichtparametrische Tests **23**

Wilcoxon-Vorzeichenrangtest . . . 23

Mann-Whitney-U-Test . . . 24

Varianzanalyse, ANOVA **24**

Hypothesen . . . 24

Quadratsummenzerlegung . . . 25

Bedingungen für ANOVA . . . 26

Post-Hoc paarweise Vergleiche . . . 26

Inferenz für qualitative Daten **28**

Zentraler Grenzwertsatz für relative Häufigkeiten (engl. *proportions*) . . . 29

Konfidenzintervall für eine relative Häufigkeit . . . 30

Hypothesentest für eine Stichprobe . . . 30

Vergleich von zwei relativen Häufigkeiten . . . 32

Chi-Quadrat-Test . . . 34

Korrelation **35**

Einfache lineare Regression **36**

Lineares Modell 37

Bedingungen für das lineare Regressionsmodell 38

Bestimmtheitsmass R^2 40

R-Funktionen **41**

Hilfe erhalten 41

Libraries verwenden 41

Arbeitsverzeichnis 41

Vektoren (Variablen) 41

Datentypen 43

Logische Operatoren 44

Mathematische Funktionen 44

Datensätze 46

Kennzahlen

Umfang

n = Stichprobenumfang
N = Umfang der Population

Arithmetisches Mittel, Mittelwert

$\bar{x}$ = Stichprobenmittelwert
μ = Populationsmittelwert

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

```
mean()
```

Beispiel:

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

```
## [1] 4
```

Median

wenn n ungerade

$$\tilde{x} = x_{\frac{n+1}{2}}$$

wenn n gerade

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

```
median()
```

Beispiel:

```
x <- c(2, 3, 4, 4, 5, 6, 10)
median(x)
```

```
## [1] 4
```

Varianz

s^2 = Stichprobenvarianz
σ^2 = Varianz der Population

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

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

```
var()
```

Beispiel:

```
x <- c(2, 3, 4, 4, 5, 6, 10)
var(x)
```

```
## [1] 6.809524
```

Standardabweichung

s = Standardabweichung der Stichprobe
σ = Standardabweichung der Population

$$s = \sqrt{s^2}$$

$$\sigma = \sqrt{\sigma^2}$$

```
sd()
```

Beispiel:

```
x <- c(2, 3, 4, 4, 5, 6, 10)
sd(x)
```

```
## [1] 2.609506
```

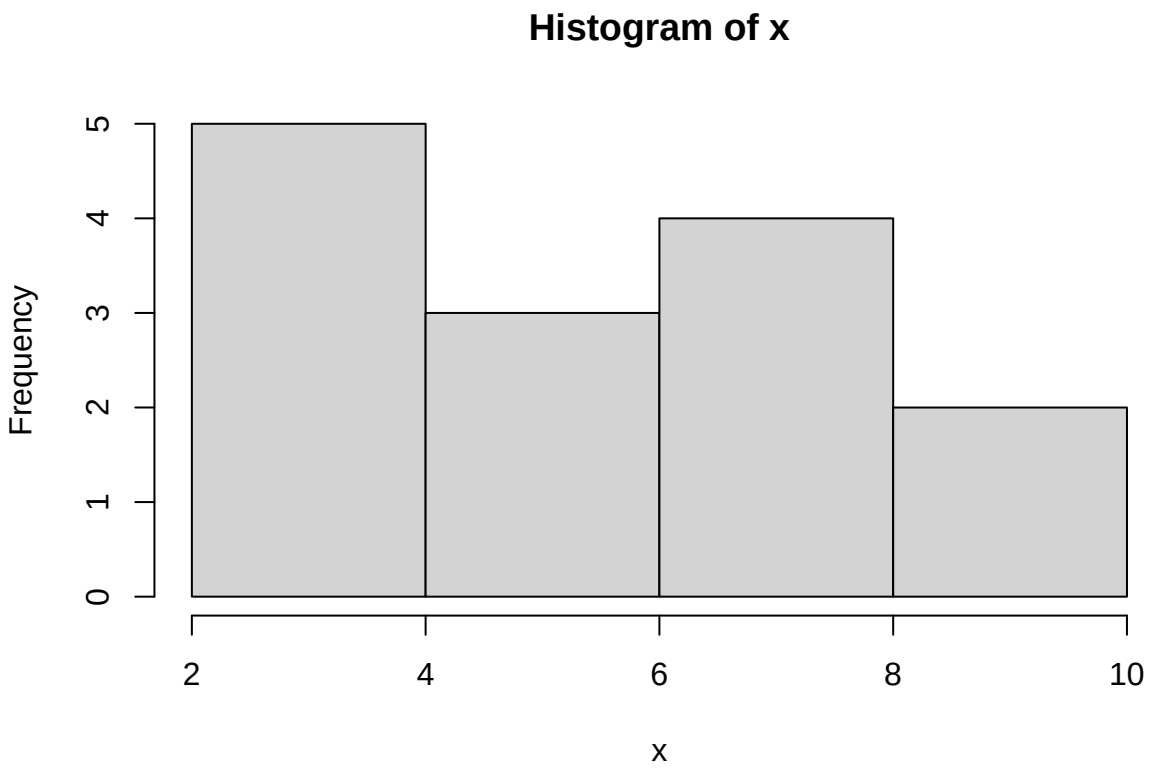
Grafiken

Histogramm

```
hist()
```

Beispiel:

```
x <- c(2, 3, 4, 4, 5, 6, 10, 9, 8, 7, 7, 7, 5, 4)
hist(x)
```

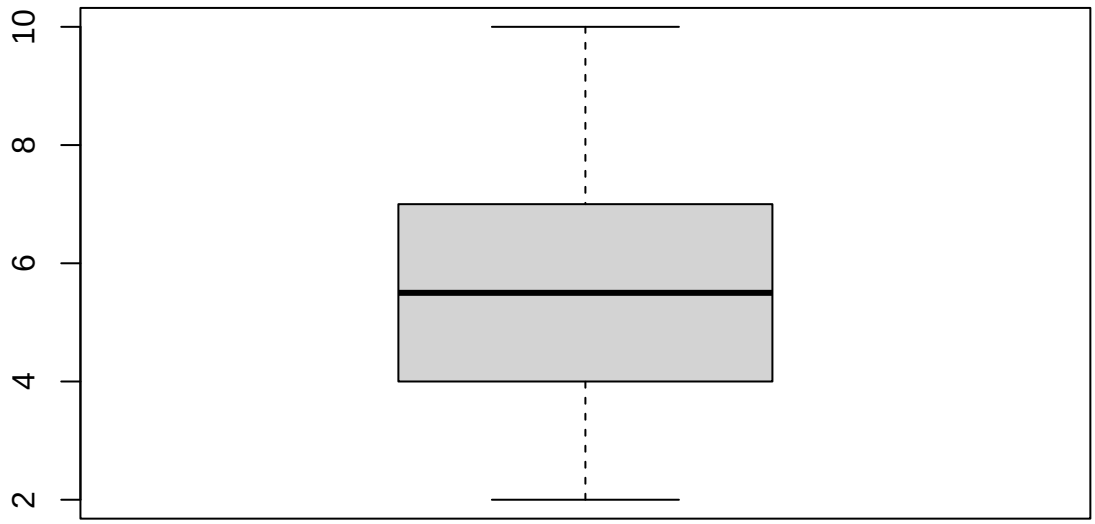


Boxplot

```
boxplot()
```

Beispiel:

```
x <- c(2, 3, 4, 4, 5, 6, 10, 9, 8, 7, 7, 7, 5, 4)
boxplot(x)
```



Kreuztabelle, absolute Häufigkeiten

```
table()
```

Beispiel:

```
x <- c("a", "a", "b", "b", "b", "c")
table(x)
```

```
## x
## a b c
## 2 3 1
```

Kreuztabelle, relative Häufigkeiten

```
prop.table()
```

Beispiel:

```
x = c("a", "a", "b", "b", "b", "c")
prop.table(table(x))
```

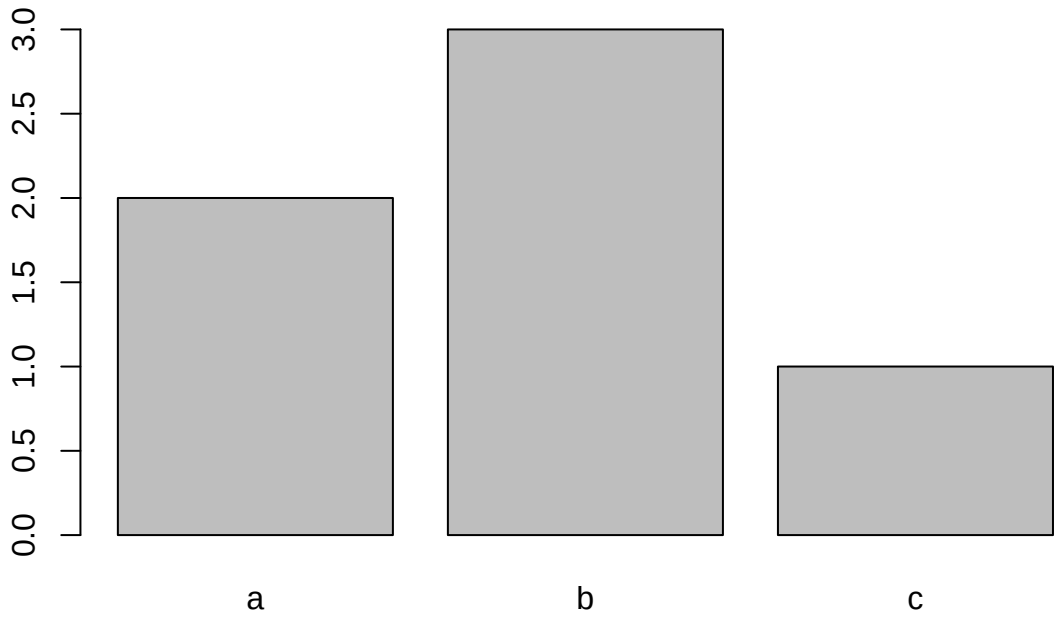
```
## x
##         a         b         c
## 0.3333333 0.5000000 0.1666667
```

Balkendiagramm

```
barplot()
```

Beispiel:

```
x = c("a", "a", "b", "b", "b", "c")
barplot(table(x))
```



Wahrscheinlichkeiten

Wahrscheinlichkeit

- Unter Wahrscheinlichkeit versteht man die Chance, dass bei einem Zufallsexperiment ein bestimmtes Ereignis auftritt.
- Wahrscheinlichkeiten können nur Werte zwischen 0 (unmögliches Ereignis) und 1 (sicheres Ereignis) zugeordnet werden.
- Nach *Laplace* ist die Wahrscheinlichkeit für ein günstiges Ereignis $p(A)$:

$$p(A) = \frac{n_A}{N_{gesamt}} = \frac{Anzahl\ der\ günstigen\ Ereignisse}{Anzahl\ der\ möglichen\ Ereignisse}$$

Ereignis und Gegenereignis

$$p(A) + p(Nicht\ A) = 1$$

Bedingte Wahrscheinlichkeiten

- Die bedingte Wahrscheinlichkeit $p(A|B)$ quantifiziert die Wahrscheinlichkeit des Ereignisses A unter der Bedingung, dass das Ereignis B eingetreten ist.

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

- Das Zeichen $\cap$ ist das mathematische Symbol für UND (Schnittmenge von A und B).
- Das *Theorem von Bayes* gibt an, wie man eine bedingte Wahrscheinlichkeit $p(A|B)$ aus der umgekehrten bedingten Wahrscheinlichkeit $p(B|A)$ berechnen kann.

$$p(A|B) = \frac{p(A) \times p(B|A)}{p(B)}$$

Unabhängigkeit

- Zwei Ereignisse A und B sind unabhängig, wenn das Eintreffen oder Nicht-Eintreffen des Ereignisses B die Wahrscheinlichkeit für ein Ereignis A nicht verändert.

$$p(A) = p(A|B)\ ,\ p(B) = p(B|A)$$

Normalverteilung

$$X \sim N(\mu, \sigma)$$

68-95-99.7-Regel

- 68% in $\mu \pm 1\sigma$
- 95% in $\mu \pm 2\sigma$, genauer $\mu \pm 1.96\sigma$
- 99.7% in $\mu \pm 3\sigma$

z-Wert

$$z = \frac{x_i - \bar{x}}{s}$$

- Der z-Wert einer Beobachtung x_i gibt an, um wieviele Standardabweichungen die Beobachtung über oder unter dem Mittelwert liegt.
- Der z-Wert des Mittelwerts ist 0
- Ungewöhnliche Beobachtungen haben einen z-Wert von $|z| > 2$.

Perzentilen in R berechnen

```
# Fläche links von x
pnorm(x, mean, sd)

# Fläche rechts von x
1 - pnorm(x, mean, sd)
pnorm(x, mean, sd, lower.tail = FALSE)

# Wert auf einer bestimmten Perzentile
qnorm(percentile, mean, sd)
```

Beispiel:

```
x <- c(2, 3, 4, 4, 5, 6, 10, 9, 8, 7, 7, 7, 5, 4)
mittelwert <- mean(x)
stdabw <- sd(x)

# Wahrscheinlichkeit für den Wert kleiner oder gleich 7
pnorm(7, mittelwert, stdabw)
```

```
## [1] 0.6991514
```

```
# Wahrscheinlichkeit für den Wert gleich oder grösser 7
1 - pnorm(7, mittelwert, stdabw)
```

```
## [1] 0.3008486
```

```r
# Wert auf der 40%-Perzentile
qnorm(.4, mittelwert, stdabw)
```

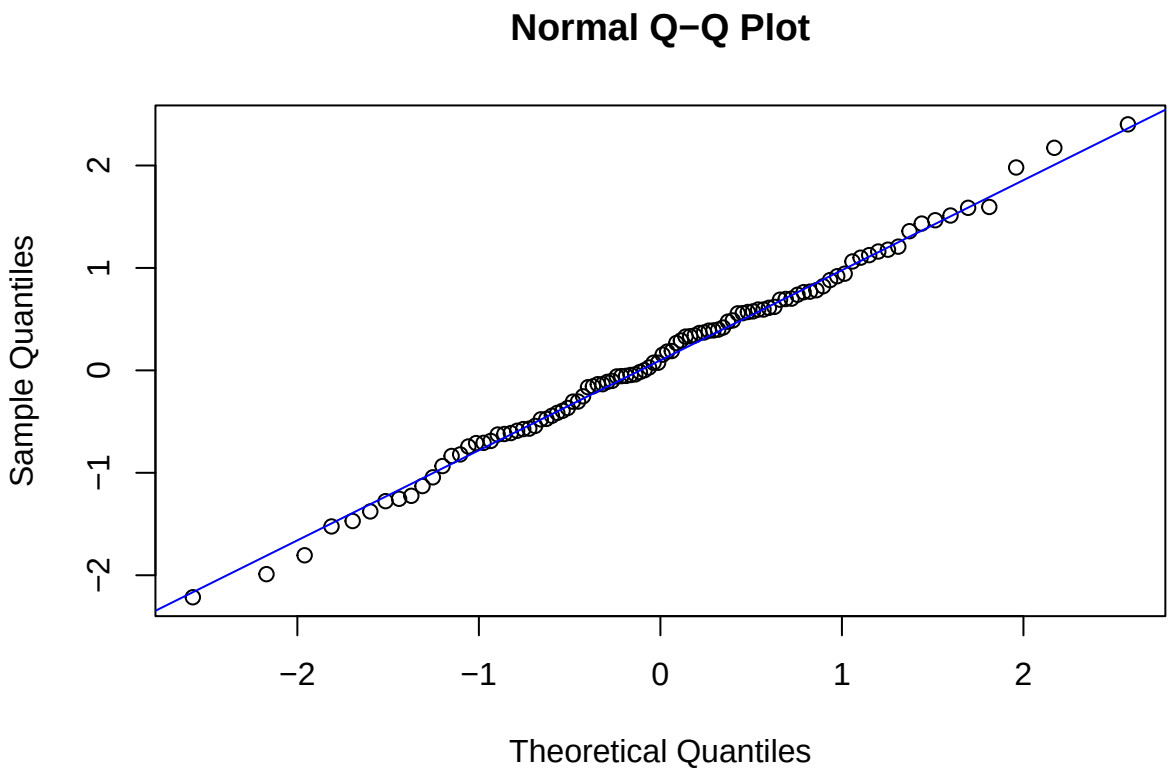
```
## [1] 5.19633
```

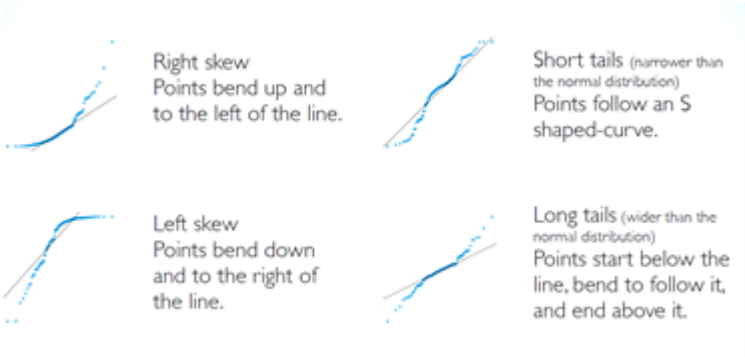
QQ-Plot

```r
# Punkte in Streudiagramm darstellen
qqnorm()

# Linie in QQ-Plot einzeichnen
qqline()
```

Beispiel:

```r
# simulation von 100 normalverteilten Werten, mean = 0, s = 1
set.seed(1)
x <- rnorm(100)

## qq-plot erstellen
qqnorm(x)

## Linie in qq-plot einzeichnen
qqline(x, col = "blue")
```





Binomialverteilung

$$X \sim Bin(n, p)$$

- n = Anzahl Versuche
- p = Eintrittswahrscheinlichkeit

Erwartungswert (Mittelwert) der Binomialverteilung

$$\mu = n \times p$$

Standardabweichung der Binomialverteilung

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

Bedingungen für Binomialverteilung

- Die Versuche müssen unabhängig sein.
- Die Anzahl der Versuche muss bekannt sein.
- Jedes Versuchsergebnis ist entweder ein Erfolg oder ein Misserfolg.
- Die Wahrscheinlichkeit für einen Erfolg muss für jeden Versuch gleich sein.

Wahrscheinlichkeiten der Binomialverteilung

- Wenn p die Wahrscheinlichkeit für einen Erfolg ist, ist $1-p$ die Wahrscheinlichkeit für einen Misserfolg. n gibt die Anzahl der Versuche an und k die Anzahl der Erfolge.

$$p(k, n) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Wahrscheinlichkeit für k Erfolge in n Versuchen mit der Erfolgswahrscheinlichkeit p in `R` berechnen:

```
dbinom(k, n, p)
```

- Anzahl Kombinationen von k Erfolgen in n Versuchen berechnen (Binomialkoeffizient)

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

```
choose(n, k)
```

Normalapproximation

- Eine Binomialverteilung mit mindestens 10 erwarteten Erfolgen und mindestens 10 erwarteten Misserfolgen folgt annähernd einer Normalverteilung.

$$np \geq 10$$
$$n(1-p) \geq 10$$

- Falls diese Bedingung erfüllt ist, gilt:

$$Bin(n,p) \sim N(\mu, \sigma)$$

- wobei

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

Grundlagen der Inferenzstatstik

Zentraler Grenzwertsatz

Die Verteilung von Stichprobenkennzahlen (z.B. Mittelwert) folgt annähernd einer Normalverteilung. Ihr Mittelwert liegt in der Nähe des Populationsmittelwertes μ mit einer Standardabweichung geteilt durch die Quadratwurzel des Stichprobenumfangs.

$$\bar{x} \sim N(Mittelwert = \mu, SE = \frac{\sigma}{\sqrt{n}})$$

Wenn σ unbekannt ist (was eigentlich immer der Fall ist), wird die Standardabweichung s der Stichprobe als Schätzer für σ eingesetzt.

$$SE = \frac{s}{\sqrt{n}}$$

Bedingungen für die Gültigkeit des zentralen Grenzwertsatzes:

- Die Beobachtungseinheiten in der Stichprobe sind unabhängig voneinander (zufällige Auswahl, zufällige Zuordnung zu Gruppen).

- Faustregel: Stichprobenumfang $n > 30$

Beispiel für die Berechnung des Standardfehlers SE in `R`

```
# simulation von 100 normalverteilten Werten, mean = 0, s = 1
set.seed(1234)
x <- rnorm(100)

# Stichprobenumfang von x ermitteln
n <- length(x)

# Standardabweichung von x berechnen
s <- sd(x)

# Berechnung von SE
SE <- s/sqrt(n)

# Output SE
SE
```

```
## [1] 0.1004405
```

Konfidenzintervalle

Konfidenzintervalle (Vertrauensintervalle, CI) können auf jedem Konfidenzniveau berechnet werden. Um die Sache nicht allzu kompliziert zu machen, wird hier v.a. exemplarisch die Berechnung von 95%-Konfidenzinteravallen vorgestellt.

- Signifikanzniveau $= \alpha$
- Konfidenzniveau $= 1 - \alpha$

$$CI^* = \bar{x} \pm z^* \times SE$$

$$z^* = |\frac{(1 - CI^*)}{2}|$$

$$z^* \times SE = z^* \times \frac{s}{\sqrt{n}}$$

$z^* \times SE$ wird auch als Fehlerbereich (engl. *margin of error*, ME) bezeichnet.

Der Wert von z^* ist abhängig vom Konfidenzniveau.

```
# z für ein 95% CI
CI <- .95
z95 <- abs(qnorm((1 - CI)/2))
z95
```

```
## [1] 1.959964
```

```
# z für ein 90% CI
CI <- .9
z90 <- abs(qnorm((1 - CI)/2))
z90
```

```
## [1] 1.644854
```

```
# z für ein 99% CI
CI <- .99
z99 <- abs(qnorm((1 - CI)/2))
z99
```

```
## [1] 2.575829
```

Beispiel für die Berechnung eines 95% Konfidenzintervalls

```
m <- 95.6       # Stichprobenmittelwert
s <- 15.8       # Standardabweichung der Stichprobe
n <- 100        # Stichprobenumfang

# gesucht ist das 95% Konfidenzintervall für den Populationsmittelwert
CI <- .95       # Konfidenzniveau 95%
z <- abs(qnorm((1-CI)/2))
ME <- z * CI    # Fehlerbereich berechnen

# Obere und untere Grenze für 95%-Konfidenzintervall berechnen
CI95 <- m + c(-1, 1) * ME
CI95
```

```
## [1] 93.73803 97.46197
```

Zuverlässigkeit vs. Präzision

Wenn wir das Konfidenzniveau erhöhen (Konfidenzintervall wird breiter, z.B. von 95% auf 99%) nimmt die Zuverlässigkeit, dass wir den wahren Populationsparameter im Intervall haben zu, allerdings auf Kosten der Präzision.
Wie können wir Zuverlässigkeit und Präzision gleichzeitig verbessern? Antwort: Stichprobenumfang erhöhen.

Stichprobenumfang für einen bestimmten Fehlerbereich berechnen:

$$ME = z^* \times \frac{s}{\sqrt{n}} \rightarrow n = (\frac{z^* \times s}{ME})^2$$

Beispiel: Im Beispiel oben betrug unser ME = 1.862. Wir möchten den ME halbieren und bestimmen den benötigten Stichprobenumfang. (Kennzahlen wie oben)

```
ME.alt <- 1.862
ME.neu <- ME.alt/2

# neues 95%-Konfidenzintervall berechnen
CI95.neu <- m + c(-1, 1) * ME.neu
CI95.neu
```

```
## [1] 94.669 96.531
```

```r
# Stichprobenumfang für das neue 95%-CI berechnen
n.neu <- ((z * s)/ME.neu)^2
n.neu
```

```
## [1] 1106.397
```

Hypothesentest für einen Mittelwert

Hypothesentests werden immer für einen Popultionsparameter, z.B. μ durchgeführt und nicht für eine Stichprobe.

1. Formuliere die wissenschaftliche Hypothese

- $H_0 : \mu = Nullwert$
- $H_A : \mu < oder > oder \neq Nullwert$
- Es wird empfohlen H_A : immer zweiseitig formulieren ausser in begründeten Ausnahmefällen.

2. Berechne den Punktschätzer $\bar{x}$ für μ

3. Überprüfe die Testvoraussetzungen

- Beobachtungseinheiten in der Stichprobe sind unabhängig.
- Stichprobe stammt aus eine annähernd normalverteilten Population.
- Der Stichprobenumfang $n \geq 30$ oder grösser bei stark schiefer Verteilung.

4. Skizziere die Stichprobenverteilung, zeichne deinen Verwerfungsbereich ein und berechne die Teststatistik.

$$z = \frac{\bar{x} - \mu}{SE}, \quad SE = \frac{s}{\sqrt{n}}$$

5. Liegt z im Verwerfungsbereich wird H_0 zu Gunsten von H_A zurückgewiesen.

6. Interpretiere dein Resultat im Zusammenhang mit der Fragestellung.

p-Werte berechnen

Definition:

$$p - Wert = P(beobachtete\ oder\ extremere\ Teststatistik \mid H_0\ wahr)$$

Der p-Wert quantifiziert die Evidenz gegen H_0. Ein kleiner p-Wert (üblicherweise $p \leq 0.05$) bedeutet, dass du ausreichend Evidenz dafür hast, H_0 zu Gunsten von H_A zu verwerfen.

Einseitiger Hypothesentest anhand von p-Werten

1. Fall

$H_A : \mu > Nullwert$

$$z = \frac{\bar{x} - Nullwert}{SE_{\bar{x}}}$$

p-Wert in `R` berechnen:

```
p <- 1 - pnorm(z)
```

2. Fall

$H_A : \mu < Nullwert$

$$z = \frac{\bar{x} - Nullwert}{SE_{\bar{x}}}$$

p-Wert in `R` berechnen:

```
p <- pnorm(z)
```

Zweiseitiger Hypothesentest anhand von p-Werten

Zweiseitige Hypothesen sind der Normalfall. Einseitige Hypothesen sollten nur in begründeten Ausnahmefällen formuliert werden.

$H_A : \mu \neq Nullvalue$

$$z = \frac{\bar{x} - Nullwert}{SE_{\bar{x}}}$$

p-Wert in `R` berechnen:

```
p <- 2 * pnorm(abs(z), lower.tail = FALSE)

# Alternative
p <- 2 * pnorm(-abs(z))
```

Entscheidungsfehler

- Fehler 1. Art: H_0 wird verworfen wenn H_0 wahr ist.
- Fehler 2. Art: H_0 wird nicht verworfen wenn H_A wahr ist.

Bei einem Signifikanzniveau $\alpha = 0.05$ nehmen wir ein Risiko von 5% in Kauf, einen Fehler 1. Art zu begehen.

- α : Wahrscheinlichkeit, einen Fehler 1. Art zu begehen.
- β : Wahrscheinlichkeit, einen Fehler 2. Art zu begehen.
- $1 - \beta$: Power (Trennschärfe) eines Tests; Wahrscheinlichkeit, für H_A zu entscheiden, wenn H_A wahr ist.

Hypothesentests mit Konfidenzintervallen

- Ein zweiseitiger Hypothesentest mit einem Signifikanzniveau α entspricht einem Konfidenzintervall mit dem Konfidenzniveau $1 - \alpha$.
- Ein einseitiger Hypothesentest mit einem Signifikanzniveau α entspricht einem Vertrauensintervall mit einem Konfidenzniveau von $1 - (2 \times \alpha)$.
- Enthält ein 95% Vertrauensintervall den Nullwert nicht, wird H_0 verworfen.
- Enthält ein 95% Vertrauensintervall den Nullwert, wird H_0 nicht verworfen.

Inferenz für quantitative Daten

Hypothesentests für gepaarte Mittelwerte

- Gepaarte (auch verbundene) Daten:
  - Gleiche Beobachtungseinheiten: Vorher-Nachher-Messungen, Messwiederholungen
  - Unterschiedlieche Beobachtungseinheiten (jedoch abhängig): Zwillingsstudien, Partner
- Parameter: μ_Δ = Mittelwert der paarweisen Differenzen in der Population
- Punktschätzer: $\bar{x}_\Delta$ = Mittelwert der paarweisen Differenzen in der Stichprobe
- Teststatistik: z-Wert
- Hypothesen:
  - $H_0 : \mu_\Delta = Nullwert$
  - $H_A : \mu_\Delta \neq Nullwert$ (zweiseitige H_A)

Vorgehen

1. Wissenschaftliche Hypothesen formulieren
2. Punktschätzer berechnen
3. Annahmen prüfen
   - Unabhängigkeit der Beobachtungseinheiten
   - Paarweise Differenzen sind annähernd normalverteilt.
   - Stichprobenumfang $n \geq 12$ oder grösser bei stark schiefen Verteilungen
4. Stichprobenverteilung skizzieren, Verwerfungsbereich einzeichnen und Teststatistik berechnen

$$z = \frac{\bar{x}_\Delta - \mu_\Delta}{SE_{\bar{x}_\Delta}}$$

5. Liegt z im Verwerfungsbereich wird H_0 zu Gunsten von H_A zurückgewiesen.
6. Resultat im Zusammenhang mit der Fragestellung interpretieren.

Konfidenzintervall für gepaarte Daten

$$CI^* = \bar{x}_\Delta \pm z * \times SE_\Delta$$

$$CI^* = \bar{x}_\Delta \pm z^* \times \frac{s_\Delta}{n}$$

Hypothesentest für unabhängige Mittelwerte

- unabhängige Daten:
  - Unterschiedliche Beobachtungseinheiten, z.B. Vergleich von zwei Gruppen
- Parameter: $\mu_1 - \mu_2$, z.B. Differenz der Mittelwerte von zwei Populationen
- Punktschätzer: $\bar{x}_1 - \bar{x}_2$ z.B. Differenz der Mittelwerte von zwei Stichproben
- Teststatistik: z-Wert
- Hypothesen:
  - $H_0 : \mu_1 = \mu_2$ bzw. $H_0 : \mu_1 - \mu_2 = 0$
  - $H_A : \mu_1 \neq \mu_2$ bzw. $H_A : \mu_1 - \mu_2 \neq 0$ (zweiseitige H_A)

Vorgehen

1. Wissenschaftliche Hypothese formulieren
2. Punktschätzer berechnen
3. Annahmen prüfen
   - Unabhängigkeit der Beobachtungseinheiten innerhalb und zwischen den Gruppen
   - Stichprobe stammt aus eine annähernd normalverteilten Population.
   - Stichprobenumfang $n_1 \geq 30$ und $n_2 \geq 30$ oder grösser bei stark schiefen Verteilungen
4. Stichprobenverteilung skizzieren, Verwerfungsbereich einzeichnen und Teststatistik berechnen

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}}$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

5. Liegt z im Verwerfungsbereich wird H_0 zu Gunsten von H_A zurückgewiesen.

6. Resultat im Zusammenhang mit der Fragestellung interpretieren.

Konfidenzintervall für unabhängige Daten

$$CI^* = (\bar{x}_1 - \bar{x}_2) \pm z^* \times SE_{\bar{x}_1 - \bar{x}_2}$$

T-Verteilung und t-Tests

Die T-Verteilung

- kann als Variante der Normalverteilung aufgefasst werden.
- hat immer den Mittelwert 0.
- hat eine Standardabweichung, die vom Stichprobenumfang n abhängig ist.
- Wird nur durch einen einzigen Parameter, die Anzahl Freiheitsgrade *df* (engl. degrees of freedom), definiert.
- Die T-Verteilung wird mit wachsendem n schmaler und geht für $n \to \infty$ in die Normalverteilung über.

$$df = n - 1$$
$$t \sim T(df)$$

Die T-verteilung wird verwendet, wenn

- der Stichprobenumfang klein ist ($n \leq 30$)
- die Standardabweichung σ der Population unbekannt ist und mit Hilfe der Stichprobenstandardabweichung s geschätzt werden muss.
- also eigentlich immer; die Software rechnet standardmässig mit der T-Verteilung.
- Die Teststatistik von T-Tests sind t-Werte. t-Werte werden gleich interpretiert wie z-Werte.

Inferenz für einen Mittelwert

Ziel: Vergleich eines Mittelwerts mit einem Vergleichswert (= Nullwert)

Hypothesen:

- $H_0 : \mu = Nullwert$
- $H_A : \mu \neq Nullwert$ (zweiseitig)

Konfidenzintervall

$$CI^* = \bar{x} \pm t^*_{df} \times SE_{\bar{x}}$$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}, \quad df = n - 1$$

Quantilen für den kritischen t-Wert (Grenzen des Verwerfungsbereichs) und für die Konstruktion von Konfidenzintervallen mit Rberechnen:

```
# für Signifikanzniveau 0.05, 95% CI
t <- qt(.025, df = n - 1)

# für Signifikanzniveau 0.1, 90% CI
t <- qt(0.05, df = n - 1)

# für Signifikanzniveau 0.01, 99% CI
t <- qt(0.005, df = n - 1)
```

Die R-Funktion `qt()` gibt die untere Grenze an. Da die T-Verteilung symmetrisch ist, entspricht die obere Grenze dem Absolutwert der unteren Grenze.

Beispiel zur Berechnung des Konfidenzintervalls in R

```
# Kennzahlen einer Stichprobe
n <- 25
m <- 15
s <- 3


ci_level <- .95                 # Konfidenzniveau
SE <- s/sqrt(n)                 # Standardfehler
t_df <- qt(.025, df = 25 - 1)   # kritischer t-Wert

CI <- m + c(-1, 1) * abs(t_df) * SE
CI
```

```
## [1] 13.76166 16.23834
```

Einstichproben-t-Test

$$t = \frac{\mu - Nullwert}{SE}$$

Beispiel: Vergleich des Mittelwerts einer Stichprobe mit dem Nullwert in einer Population. Der Nullwert sei $Nullwert = 13$

Hypothesen:

- $H_0 : \mu = 13$
- $H_A : \mu \neq 13$ (zweiseitig)

```r
# Kennzahlen unserer Stichprobe und Nullwert
n <- 25
m <- 15
s <- 3
nullwert <- 13

SE <- s/sqrt(n)                # Standardfehler
t <- (m - nullwert)/SE         # Teststatistik

# p-Wert für zweisseitige Hypothese berechnen
p <- 2 * pt(abs(t), df = n - 1, lower.tail = FALSE)

# output von t und p
paste("t =", t, ", p =", p)
```

```
## [1] "t = 3.33333333333333 , p = 0.00277631418305654"
```

Einfacher geht es mit der Funktion `t.test()`

```r
# Daten simulieren (muss man nicht verstehen)
rnorm2 <- function(n, mean, sd) { mean + sd * scale(rnorm(n)) }
x <- rnorm2(n = 25, mean =15, sd = 3)

# t-Test in R
t.test(x, # Stichprobendaten mit m = 15, s = 3, n = 25
       mu = 13, # Nullwert
       alternative = "two.sided") # zweiseitiger Test
```

```
##
##  One Sample t-test
##
## data:  x
## t = 3.3333, df = 24, p-value = 0.002776
## alternative hypothesis: true mean is not equal to 13
## 95 percent confidence interval:
##  13.76166 16.23834
## sample estimates:
## mean of x
##        15
```

Der Einstichproben-t-Test eignet sich auch als Test für gepaarte Daten mit der Prüfgrösse mu_{Δ}.

Inferenz für zwei Mittelwerte

Ziel: Vergleich von zwei Mittelwerten aus zwei Stichproben

Hypothesen:

- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu_1 \neq \mu_2$ (zweiseitig)

Konfidenzintervall

$$CI^* = (\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times SE_{\bar{x}_1 - \bar{x}_2}$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = n_1 + n_2 - 2$$

Die Formeln für die Berechnung von SE und df sind etwas vereinfacht; genaue Formeln findet man in Statistiklehrbüchern.

Zweistichproben-t-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}}$$

```r
# Kennzahlen unserer Stichprobe und Nullwert
n1 <- 25
m1 <- 15
s1 <- 3
n2 <- 20
m2 <- 18
s2 <- 4

# Standardfehler
SE <- sqrt((s1^2 / n1) + (s2^2/n2))

# Teststatistik
t <- (m1 - m2)/SE

# Freiheitsgrade n2 - 1 ist kleiner als n1 - 1
df <- n1 + n2 - 2

# p-Wert für zweisseitige Hypothese berechnen
p <- 2 * pt(abs(t), df, lower.tail = FALSE)

# output von t und p
paste("t =", t, ", p =", p)
```

```
## [1] "t = -2.78543007265578 , p = 0.00791948797764193"
```

Einfacher geht es mit der Funktion `t.test()`

```r
# Daten simulieren (muss man nicht verstehen)
rnorm2 <- function(n, mean, sd) { mean + sd * scale(rnorm(n)) }
x1 <- rnorm2(n = 25, mean =15, sd = 3)
x2 <- rnorm2(n = 20, mean =18, sd = 4)

# t-Test in R
t.test(x = x1, # Gruppe 1
```

```
       y = x2, # Gruppe 2
       paired = FALSE,
       alternative = "two.sided")# zweiseitiger Test
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -2.7854, df = 34.428, p-value = 0.00863
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.187792 -0.812208
## sample estimates:
## mean of x mean of y
##        15        18
```

Nichtparametrische Tests

Nichtparametrische Tests kommen zur Anwendung, wenn die Annahme der Normalverteilung fraglich ist.

Wilcoxon-Vorzeichenrangtest

Vergleicht einen Median mit einem vorgegebenen Referenzmedian. Annahmen:

- quantitative oder ordinal skalierte Daten
- unabhängige Beobachtungseinheiten
- Daten sind annähernd symmetrisch um den Median verteilt

```
wilcox.test(x, mu = Referenzwert, alternative = "two.sided")
```

Beispiel:

```
# Daten generieren
X <- c(3, 3, 4, 5, 6, 7, 7, 8, 1, 2)
nullwert <- 6.5

# Wilcoxon-Vorzeichenrangtest
wilcox.test(x = X, mu = nullwert, alternative = "two.sided")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  X
## V = 8.5, p-value = 0.05835
## alternative hypothesis: true location is not equal to 6.5
```

Mann-Whitney-U-Test

Wird auch *Wilcoxon Rangsummen-Test* genannt.

Testet nicht ganz dasselbe wie der t-Test

- $H_0 : P(X > Y) = P(Y > X)$, m.a.W: Es besteht eine 50%-Wahrscheinlichkeit dafür, dass ein zufällig gezogener Wert aus X grösser ist als ein zufällig gezogener Mittelwert aus Y (und umgekehrt)
- $H_0 : P(X > Y) \neq P(Y > X)$, m.a.W: Die Wahrscheinlichkeit ist nicht 50%, dass ein zufällig gezogener Wert aus X grösser ist als ein zufällig gezogener Mittelwert aus Y (und umgekehrt)

Annahmen

- quantiative oder ordinal skalierte Daten
- unabhängige Beobachtungen

```
wilcox.test(x, y, alternative = "two.sided", paired = FALSE)
```

Beispiel:

```
# Daten generieren
X <- c(3, 3, 4, 5, 6, 7, 7, 8, 1, 2)
Y <- c(2, 3, 2, 5, 6, 2, 3, 8, 1, 2)

# Mann-Whitney-U-Test
wilcox.test(x = X, y = Y, alternative = "two.sided", paired = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  X and Y
## W = 66, p-value = 0.2351
## alternative hypothesis: true location shift is not equal to 0
```

Varianzanalyse, ANOVA

- ANOVA steht für Varianzanalyse (engl. Analysis of Variance) und wird verwendet um die Mittelwerte von mehr als 2 Gruppen zu vergleichen.

Hypothesen

- $H_0 : \mu_1 = \mu_2 ... = \mu_n$
- H_A : *Die Mittelwerte sind nicht alle gleich*

Quadratsummenzerlegung

- Bei der Varianzanalyse wird die "Gesamtvarianz" der abhängigen Variablen y in die Varianz zwischen den Gruppenmittelwerten und die Varianz zwischen den Messwerten innerhalb der Gruppen zerlegt.

$$SS_{total} = SS_{between} + SS_{within}$$

- Die **Gesamtquadratsumme** SS_{total} misst die totale Variabilität der abhängigen Variablen.

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- y_i : Wert der abhängigen Variablen für jede Beobachtung
- $\bar{y}$: Mittelwert der abhängigen Variablen (sog. *grand mean*)
- Die **Quadratsumme zwischen den Gruppen** misst die Variabilität zwischen den Gruppen; entspricht der Variabilität, die durch die Gruppierungsvariable erklärt wird (erklärte Variabilität).

$$SS_{between} = \sum_{j=1}^{k} n_j (\bar{y}_i - \bar{y})^2$$

- n_j : Anzahl Beobachtungen in Gruppe j
- $\bar{y}_j$: Mittelwert der abhängigen Variablen in Gruppe j
- $\bar{y}$: Mittelwert der abhängigen Variablen (*grand mean*)
- Die **Quadratsumme innerhalb der Gruppen** misst die Variabilität innerhalb der einzelnen Gruppen; entspricht der Variabilität, die nicht durch die Gruppierungsvariable beschrieben wird, also andere Gründe hat (unerklärte Variabilität)

$$SS_{within} = SS_{total} - SS_{between}$$

- Freiheitsgrade für ANOVA
  - Total: $df_{tota} = n - 1$
  - für SS_{between}: $df_{between} = k - 1$
  - für SS_{within}: $df_{within} = df_{total} - dfbetween$
- Die **mittleren Quadratsummen** beschreiben die durchschnittliche Variabilität zwischen und innerhalb der Gruppen.

$$MSS_{between} = SS_{between}/df_{between}$$

$$MSS_{within} = SS_{within}/df_{within}$$

- Teststatistik F

$$F = \frac{MSS_{between}}{MSS_{within}}$$

- p-Wert
  - gibt die Wahrscheinlichkeit eine so grosse oder noch grössere Teststatistik F unter der Annahme, dass die Mittelwerte aller Gruppen gleich gross sind.
  - ist die Fläche unter der Kurve der F-Verteilung mit den Freiheitsgraden $df_{between}$ und df_{within} oberhalb des F-Werts.
  - in `R`

```
pf(F, df_between, df_within, lower.tail = FALSE)
```

Bedingungen für ANOVA

- Unabhängigkeit der Messungen
  - zwischen den Gruppen: Die Gruppen müssen voneinander unabhängig sein, andernfalls ist eine ANOVA für Messwiederholungen (repeated measures anova) durchzuführen.
  - innerhalb der Gruppen: Die Beobachtungseinheiten müssen unabhängig voneinander sein.
- Normalverteilung: Die Daten innerhalb jeder Gruppe sollten annähernd normalverteilt sein.
- Die Gruppen sollten annähernd gleiche Varianzen haben.

Post-Hoc paarweise Vergleiche

- Das Signifikanzniveau muss für die Anzahl der Vergleiche angepasst werden. Es existieren verschiedene Verfahren. Am einfachsten ist die **Bonferroni-Korrektur**. Vergleiche den p-Wert für jeden Test mit dem Signifikanzniveau α^*.

$$\alpha^* = \frac{\alpha}{K}$$

$$K = Anzahl\ Vergleiche = \frac{k(k-1)}{2}$$

- Standardfehler für mehrere paarweise Vergleiche

$$SE = \sqrt{\frac{MSS_{within}}{n_1} + \frac{MSS_{within}}{n_2}}$$

- Teststatistik t

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{MSS_{within}}{n_1} + \frac{MSS_{within}}{n_2}}}$$

```
# p-Wert für zweiseitigen t-Test
2 * pt(t-Wert, df, lower.tail = FALSE)
```

Beispiel

```
library(dplyr)

# create sample data ------------------------------------------------------------
data <- tibble(
  gruppe = c(rep("G1", 7), rep("G2", 7), rep("G3", 7)),
  score = c(92, 93, 100, 104, 89, 91, 99, 71, 62, 85, 94, 78, 66, 71, 64, 73, 87, 91, 56, 78, 87)
)

# create anova summary table ----------------------------------------------------
anova <- aov(score ~ gruppe, data = data)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## gruppe       2   1780   890.1    8.18 0.00297 **
## Residuals   18   1959   108.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

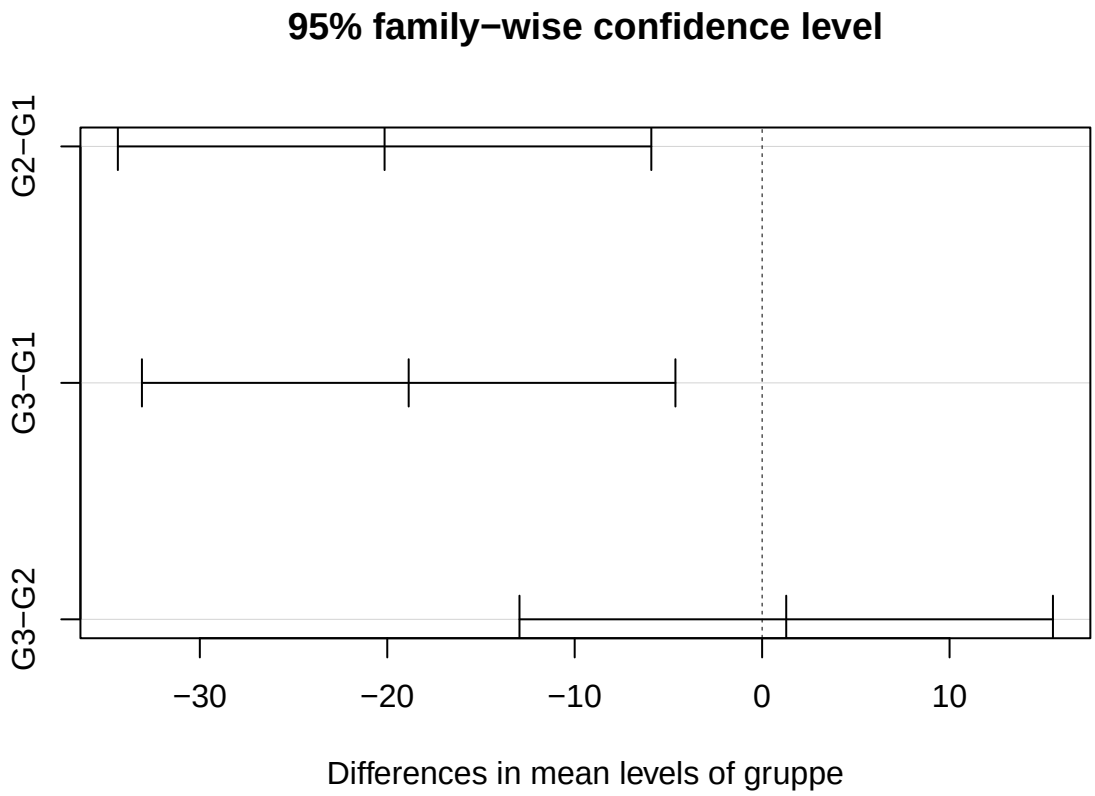
```
# post hoc analysis -------------------------------------------------------------
pairwise.t.test(data$score, data$gruppe,
                p.adjust.method = "bonferroni",
                paired = FALSE,
                alternative = "two.sided")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$score and data$gruppe
##
##    G1    G2
## G2 0.006 -
## G3 0.010 1.000
##
## P value adjustment method: bonferroni
```

```r
# 95%-Konfidenzintervalle für Differenzen -------------------------------------
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = score ~ gruppe, data = data)
##
## $gruppe
##             diff       lwr       upr     p adj
## G2-G1 -20.142857 -34.37399 -5.911722 0.0053757
## G3-G1 -18.857143 -33.08828 -4.626008 0.0088636
## G3-G2   1.285714 -12.94542 15.516849 0.9711637
```

```r
plot(TukeyHSD(anova))
```



Inferenz für qualitative Daten

- p : Populationsparameter
- $\hat{p}$: Punktschätzer für den Populationsparameter (Anzahl Erfolge/Stichprobenumfang)

Zentraler Grenzwertsatz für relative Häufigkeiten (engl. *proportions*)

- Relative Häufigkeiten von Stichproben sind annähernd normalverteilt mit ihrem Zentrum bei der Häufigkeit in der Population und einem Standardfehler, der umgekehrt proportional ist zum Stichprobenumfang.

$$\hat{p} \sim N\left(Mittelwert = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

- Voraussetzungen
  - Unabhängigkeit: Die Beobachtungen müssen voneinander unabhängig sein
  - Stichprobenumfang: Es müssen mindestens 10 Erfolge und 10 Misserfolge vorliegen

$$n \times p \geq 10;\ n \times (1-p) \geq 10$$

Beispiel

- Berechne die Wahrscheinlichkeit $P(\hat{p} > 0.95)$ für ein Ereignis mit der Erfolgswahrscheinlichkeit $p = 0.9$ und einen Stichprobenumfang $n = 200$.

1. Voraussetzungen prüfen

```
p <- 0.9
n <- 200

n * p       # Anzahl Erfolge
```

```
## [1] 180
```

```
n * (1 - p) # Anzahl Misserfolge
```

```
## [1] 20
```

2. Normalverteilungsapproximation

- Mittelwert und SE berechnen

$$\hat{p} \sim N\left(Mittelwert = 0.9, SE = \sqrt{\frac{0.9 \times 0.1}{200}} = 0.0212\right)$$

- z-Wert berechnen

$$z = \frac{0.95 - 0.9}{0.0212} = 2.36$$

- p-Wert berechnen

$$P(Z > 2.36) \approx 0.0091$$

- Unter Verwendung der Binomialverteilung
  - Die erwartete Wahrscheinlichkeit bei 200 Versuchen mit $p = 0.95$ ist $\hat{p} = 200 \times 0.95 = 190$
  - Wie gross ist die Wahrscheinlichkeit für $p \geq 190$ bei 200 Stichproben und einer Wahrscheinlichkeit von $p = 0.9$ in der Population unter der Nullhypothese?

```
# Wahrscheinlichkeit für 190 oder mehr Erfolge

sum(dbinom(190 : 200, size = 200, prob = .9))
```

```
## [1] 0.00807125
```

Konfidenzintervall für eine relative Häufigkeit

Vorgehen

1. Voraussetzungen prüfen: Die Beobachtungen sind unabhängig, $\hat{p}n \geq 10$ und $(1 - \hat{p})n \geq 10$ (Beachte: Wir verwenden hier den Schätzer für p!)
2. Wenn die Voraussetzungen erfüllt sind, ist für die Stichprobenverteilung von $\hat{p}$ näherungsweise das Normalverteilung gültig.
3. Standardfehler SE mit $\hat{p}$ anstelle von p berechnen mit der Formel

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

4. Konfidenzintervall berechnen

$$\hat{p} \pm z \times SE_{\hat{p}}$$

- $z = 1.96$ für ein 95%-Konfidenzintervall

Hypothesentest für eine Stichprobe

Vorgehen

1. Hypothesen formulieren:

- $H_0 : p = Nullwert$
- $H_A : p < oder > oder \neq Nullwert$

2. Punktschätzer $\hat{p}$ berechnen.

3. Voraussetzungen prüfen

- Beobachtungen müssen unabhängig sein.
- $np \geq 10$ und $n(p-1) \geq 10$ (hier p aus der Nullhypothese einsetzen!)

4. Teststatistik z berechnen

$$z = \frac{\hat{p} - p}{SE} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

5. Entscheide und interpretiere im Kontext der Forschungsfrage

   a) Verwerfe H_0, wenn $p \leq \alpha$; die Daten liefern Evidenz gegen H_0.

   b) Verwerfe H_0 nicht, wenn $p > \alpha$; die Daten liefern keine Evidenz gegen H_0.

$\hat{p}$ versus p

- Berechnung eines Konfidenzintervalls: p ist unbekannt und wir setzen den besten Schätzer $\hat{p}$ ein.

$$n\hat{p} \geq 10;\ n(1-\hat{p}) \geq 10$$

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

- Hypothesentest: Wir testen gegen die Nullhypothese und setzen p aus H_0 ein.

$$np \geq 10;\ n(1-p) \geq 10$$

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Beispiel

- In einer Stichprobe von 100 Schüler:innen einer Schule sind 20 Raucher:innen.

$$n = 100,\ \hat{p} = 20/100 = 0.20$$

- Konfidenzintervall berechnen

$$CI = \hat{p} \pm z \times SE$$

$$CI = \hat{p} \pm z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$CI_{95} = 0.20 \pm 1.96 \times \sqrt{\frac{0.20 \times 0.80}{100}}$$

$$CI_{95} = 0.20 \pm 1.96 \times 0.04 \approx [0.12,\ 0.28]$$

- *Interpretation: Wir können zu 95% darauf vertrauen, dass an dieser Schule zwischen 12% und 28% der Schüler:innen rauchen.*

- Hypothesentest
  - Frage: Unterscheidet sich der wahre Anteil an Schüler:innen, die an dieser Schule rauchen von 18%?
  - $H_0 : p = 18$
  - $H_A : p \neq 18$

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

$$z = \frac{0.20 - 0.18}{\sqrt{\frac{0.18(1-0.18)}{100}}} = \frac{0.02}{0.0384} = 0.52$$

```
2 * pnorm(0.52, lower.tail = FALSE)
```

```
## [1] 0.6030636
```

- *Interpretation: $p - Wert > \alpha$, wir haben keine Evidenz gegen die H_0, die besagt, dass der wahre Anteil an Raucher:innen 18% beträgt. Das Ergebnis des Hypothesentests stimmt mit dem berechneten Konfidenzintervall von [0.12, 0.28] überein, das den Wert 0.18 enthält.*

Vergleich von zwei relativen Häufigkeiten

- Anwendung des zentralen Grenzwertsatzes

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(Mittelwert = (p_1 - p_2), SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

- Wir vergleichen die Zahlen der Schule mit einer zweiten Schule: Dort wurde eine Zufallsstichprobe von 120 Schüler:innen erhoben, wovon 30 Raucher:innen waren.

$$n_1 = 100, \hat{p}_1 = 20/100 = 0.20$$

$$n_2 = 120, \hat{p}_2 = 30/120 = 0.25$$

$$CI = (\hat{p}_1 - \hat{p}_2) \pm z \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$CI = (0.20 - 0.25) \pm z \times \sqrt{\frac{0.20 \times 0.80}{100} + \frac{0.25 \times 0.75}{120}}$$

$$CI_{95} = -0.05 \pm 1.96 \times 0.0562 \approx [-0.16,\ 0.06]$$

- *Interpretation: Wir können zu 95% darauf vertrauen, dass der wahre Anteil an Raucher:innen in Schule 1 um -16% tiefer bis 6% höher ist als an Schule 2.*

- Hypothesentest
  - Frage: Unterscheiden sich die Anteile an Raucher:innen zwischen den beiden Schulen?
  - $H_0 : p_1 = p_2$
  - $H_A : p_1 \neq p_2$

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

$$\hat{p}_{pool} = \frac{Anzahl\ Erfolge}{Anzahl\ Fälle} = \frac{20+30}{100+120} \approx 0.23$$

$$z = \frac{(0.20 - 0.25) - 0}{\sqrt{\frac{0.23 \times 0.77}{100} + \frac{0.23 \times 0.77}{120}}} = \frac{-0.05}{0.057} = -0.88$$

```
2 * pnorm(0.88, lower.tail = FALSE)
```

```
## [1] 0.3788593
```

- *Interpretation: $p - Wert > \alpha$; wir haben keine Evidenz gegen die H_0, die besagt, dass es keinen Unterschied zwischen den beiden Schulen bezüglich der Anteile an Raucher:innen gibt. Das 95%-Konfidenzintervall unterstützt dieses Ergebnis, da es Null enthält.*

Chi-Quadrat-Test

auch *Chi-Quadrat-Anpassungstest* oder *Chi-Quadrat-Unabhängigkeitstest*
Untersucht, ob eine Zusammenhang zwischen zwei nominal oder ordinal skalierten Variablen besteht. Hypothesen:

- H_0 : Die Zeilen- und Spaltenvariablen sind voneineinander unabhängig.
- H_A : Die Zeilen- und Spaltenvariablen sind hängen voneinander ab.

Für jede Zelle der Tabelle muss der erwartete Wert E unter der Nullhypothese berechnet werden.

$$E = \frac{Spaltentotal \times Zeilentotal}{Gesamttotal}$$

χ^2-Teststatistik

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

- O : beobachtete absolute Häufigkeiten
- E : erwartete absolute Häufigkeiten
- k : Anzahl Zellen

Die χ^2-Verteilung hat nur einen Paramter: *df*

$$df = (R-1) \times (C-1)$$

- R : Anzahl Zeilen
- C : Anzahl Spalten

Merke: Der χ^2-Test darf nur durchgeführt werden, wenn die erwartete Häufigkeit in jeder Zelle mindestens 5 beträgt. Andernfalls *Fisher's exakten Test* durchführen.

Der χ^2-Test kann in `R` einfach mit der Funktion `chisq.test()` durchgeführt werden.

Der kritische Wert für χ^2 kann in einer Verteilungstabelle abgelesen werden. Bei einer Vierfeldertafel (2 Zeilen und 2 Spalten) ist der Zusammenhang zwischen der Zeilen- und der Kolonnenvariable statistisch signifikant auf dem Niveau von 5% wenn χ^2 grösser als 3.84 ($= 1.96^2$) ist.

Funktion in `R`

```
chisq.test()
```

Beispiel: Untersucht wurde bei 100 Schüler:innen, ob sie Tictoc verwenden.

```r
# Beispielaten generieren
tictoc_m <- c(rep("ja", 23), rep("nein", 29))
tictoc_w <- c(rep("ja", 38), rep("nein", 10))
geschlecht <- c(rep("m", length(tictoc_m)), rep("w", length(tictoc_w)))
tictoc <- data.frame(Geschlecht = geschlecht,
                     tictoc = c(tictoc_m, tictoc_w))

# Chi-Quadrat-Test, Ergebnis in chisq speichern
chisq <- chisq.test(table(tictoc))

# Testergebnis anzeigen
chisq
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(tictoc)
## X-squared = 11.379, df = 1, p-value = 0.0007428
```

```r
# Beobachtete Werte anzeigen
chisq$observed
```

```
##           tictoc
## Geschlecht ja nein
##          m 23   29
##          w 38   10
```

```r
# erwartete Werte anzeigen
chisq$expected
```

```
##           tictoc
## Geschlecht    ja  nein
##          m 31.72 20.28
##          w 29.28 18.72
```

Korrelation

Beschreibt die Stärke eines linearen Zusammenhangs zwischen zwei Variablen.
Zwei Korrelationskoeffizienten:

- **Korrelationskoeffizient nach Pearson** r
  - ist empfindlich gegenüber Ausreissern

$$r = \frac{s_{xy}}{s_x \times s_y}$$

s_{xy} bezeichnet die *Covarianz* der beiden Variablen X und Y:

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

- **Rangkorrelationskoeffizient nach Spearman** r_s
  - ist robust gegenüber Ausreissern
  - misst den monotonen Zusammenhang zwischen zwei Variablen

Interpretation Korrelationskoeffizienten

- Wertebereich: $[-1,\ 1]$, 0 (kein Zusammenhang) ± 1 (perfekter Zusammenhang)
- Das Vorzeichen gibt die Richtung des Zusammenhangs an: - (Minus) bedeutet negativer Zusammenhang, + (Plus) bedeutet postitiver Zusammenhang.
- Faustregel zur Interpretation:
  - -0.8 bis -1: starker negativer Zusammenhang
  - -0.8 bis -0.5: mittlerer negativer Zusammenhang
  - -0.5 bis 0.5: schwacher positiver Zusammenhang
  - 0.5 bis 0.8: mittlerer Zusammenhang
  - 0.8 bis 1: starker Zusammenhang

```r
# Korrelationskoeffizient nach Pearson
cor(x, y)

# Rangkorrelationskoeffizient nach Spearman
cor(x, y, method = "spearman")
```

Hypothesentest für Korrelationskoeffizienten

- $H_0 : \rho = 0$, es besteht kein linearer Zusammenhang zwischen zwei Variablen.
- $H_A : \rho \neq 0$, es besteht ein linearer Zusammenhang zwischen zwei Variablen.

```r
# Korrelationskoeffizient nach Pearson
cor.test(x, y)

# Rangkorrelationskoeffizient nach Spearman
cor.test
```

Einfache lineare Regression

Quantifiziert den Zusammenhang zwischen zwei Variablen.
Unterscheide: **abhängige Variable** y und **unabhängige Variable** x**, Prädiktor**

Lineares Modell

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

bzw. mit Stichprobendaten

$$\hat{y} = b_0 + b_1 x + e$$

- $\hat{y}$: geschätzte abhängige Variable
- b_0 : Achsenabschnitt (x = 0), *intercept*
- b_1 : Steigung der Regressionsgeraden, *slope*
- x : Prädiktor
- e : Fehler, *Residuen*

$$e_i = y_i - \hat{y}_i$$

Steigung der Regressionsgeraden b_1

- Wenn x quantitativ ist: Wenn x um eine Einheit erhöht wird, erwarten wir, dass y um $|b_1|$ Einheiten zunimmt bzw. abnimmt.
- Wenn x nominal ist: Der Wert von y nimmt um $|b_1|$ Einheiten gegenüber dem Referenzlevel zu bzw. ab.

$$b_1 = \frac{s_y}{s_y} r$$

- s_y : Standardabweichung von y
- s_y : Standardabweichung von x
- r : Korrelationskoeffizient nach Pearson

Achsenabschnitt b_0

- Wenn x quantitativ ist: Wenn $x = 0$ ist y im Durchschnitt gleich b_0
- Wenn x nominal ist: Der durchschnittliche Wert von y für ein bestimmtes Level von x ist gleich b_0.

$$b_0 = \bar{y} - b_1 \bar{x}$$

- $\bar{y}$: Mittelwert von y
- $\bar{x}$: Mittelwert von x

Bedingungen für das lineare Regressionsmodell

1. Linearität

- Es besteht eine lineare Beziehung zwischen y und x.
- Wird anhand von einem Streudiagramm überprüft.

2. Normalverteilung der Residuen

- Die Residuen sind annähernd normalverteilt, mit einem Mittelwert um 0.
- Wird anhand von einem QQ-Plot für die Residuen überprüft.

3. Konstante Variabilität (Homoskedastizität)

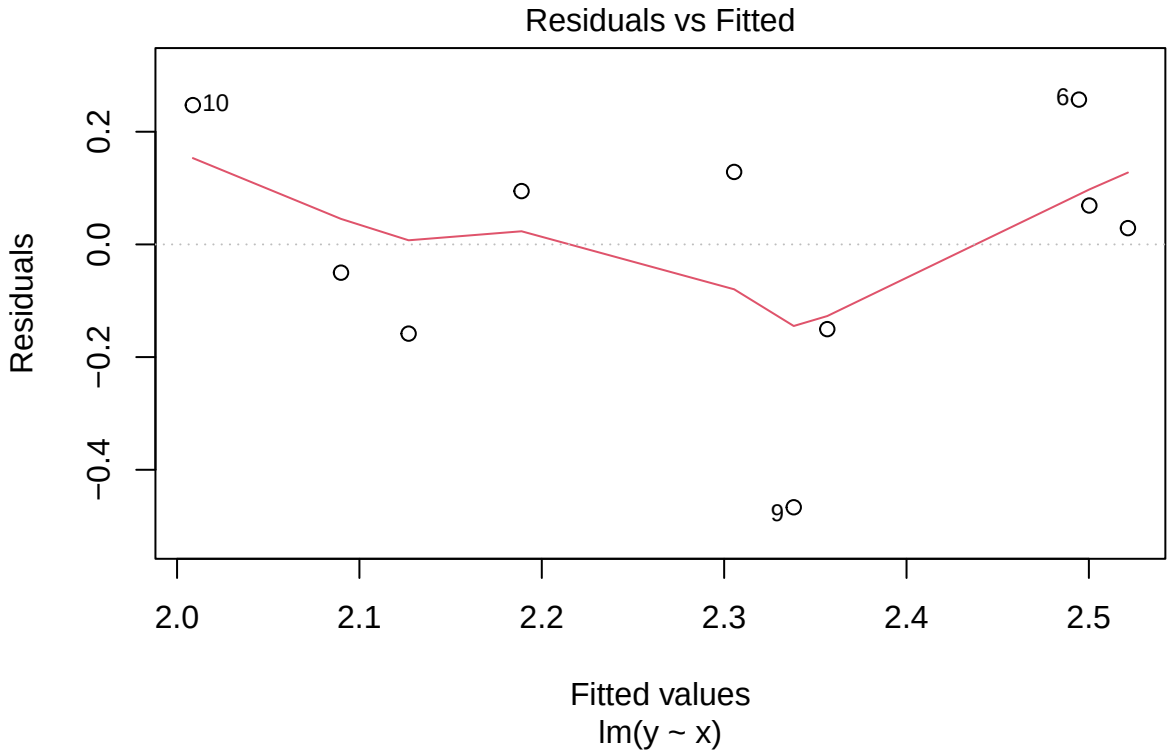
- Die Streuung der Punkte um die Regressionsgerade sollte annähernd konstant sein.
- Das bedeutet, dass die Streuung der Residuen um den Mittelwert 0 annähernd konstant ist.
- Wird an einem Streudiagramm für die Residuen geprüft.

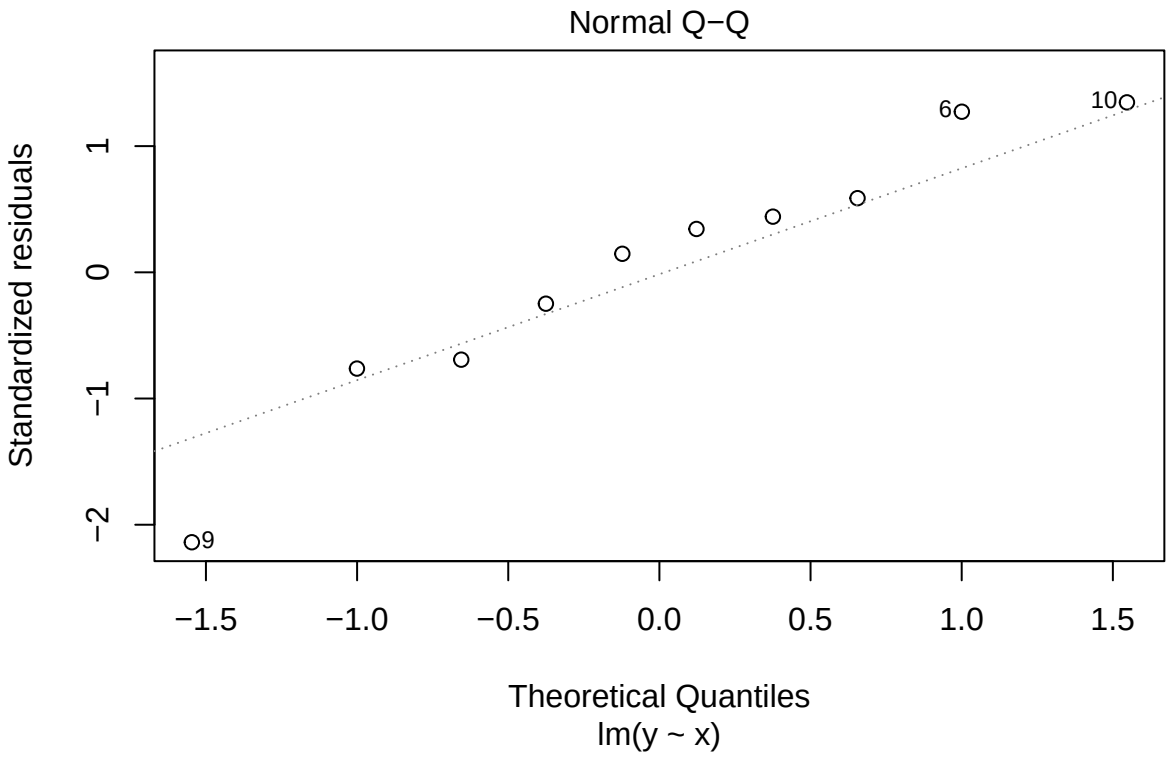
```r
# Beispieldaten generieren
set.seed(1)
b0 <- 2
b1 <- .5
x <- runif(10)
error <- rnorm(10, 0, .2)
y <- b0 + b1 * x + error
daten <- data.frame(x = x, y = y)

# lineares modell berechnen
model <- lm(y ~ x, data = daten)

# Zusammenfassung des Modells anzeigen
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = daten)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46653 -0.12534  0.04895  0.12012  0.25701
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9728     0.1530  12.898 1.23e-06 ***
## x             0.5808     0.2437   2.383   0.0443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2308 on 8 degrees of freedom
## Multiple R-squared:  0.4151, Adjusted R-squared:  0.342
## F-statistic: 5.679 on 1 and 8 DF,  p-value: 0.04434
```

```r
# Diagnostische Plots anzeigen
plot(model, which = 1:2)
```





Bestimmtheitsmass R^2

Für die einfache lineare Regression:

$$R^2 = r^2$$

- r : Korrelationskoeffizient nach Pearson
- Ist ein Mass für die Güte eines linearen Modells
- sagt uns, wieviel Prozent der Streuung in y durch x erklärt werden.

$$R^2 = \frac{durch\ x\ erklärte\ Streuung\ von\ y}{Gesamtstreuung\ von\ y}$$

- Die nicht durch R^2 erklärbare Streuung wird durch Faktoren erklärt, die nicht im Modell eingeschlossen sind.
- Wertebereich: [0, 1], 0 = 0%, 1 = 100%
- Interpretation: R^2 der Variabilität von y wird durch x erklärt.

R-Funktionen

Zusammenstellung einiger häufig verwendeter `R`-Funktionen.

Wer etwas mehr Details sucht ist hier gut aufgehoben:

- Einführung in R
- Base R - Cheat Sheet

Hilfe erhalten

Hilfe zu einer bestimmten Funktion (in `RStudio` im Register Help)

```
?mean
```

Struktur eines Objekts anzeigen (in `RStudio` im Register Environment)

```
str(objectname)
```

Libraries verwenden

Eine Library herunterladen und installieren

```
install.packages("libraryname")
```

Eine Library laden

```
library(libraryname)
```

Arbeitsverzeichnis

Arbeitsverzeichnis anzeigen

```
getwd()
```

Arbeitsverzeichnis definieren

```
setwd("C://pfad")
```

Vektoren (Variablen)

Vektoren erzeugen

```
# Werte einem Vektor zuweisen
x <- 2
x
```

```
## [1] 2
```

```r
# Elemente zu einem Vektor verbinden
x <- c(2, 4, 6)
x
```

```
## [1] 2 4 6
```

```r
# Eine ganzzahlige Sequenz erzeugen
x <- 2:6
x
```

```
## [1] 2 3 4 5 6
```

```r
# Eine komplexe Sequenz erzeugen
x <- seq(2, 3, by = .5)
x
```

```
## [1] 2.0 2.5 3.0
```

Vektorfunktionen

```r
# Beispielvariable erzeugen für Demo
x <- c(3, 2, 1, 2, 2, 1, 3, 4)

# Variable sortieren
sort(x)
```

```
## [1] 1 1 2 2 2 3 3 4
```

```r
# Werte zählen und in Tabelle ausgeben
table(x)
```

```
## x
## 1 2 3 4
## 2 3 2 1
```

```r
# Länge einer Variable bestimmen
length(x)
```

```
## [1] 8
```

Vektorelemente auswählen

```r
# das 4. Element
x[4]
```

```
## [1] 2
```

```r
# alle Elemente ausser dem 4. Element
x[-4]
```

```
## [1] 3 2 1 2 1 3 4
```

```r
# Elemente 2 bis 4
x[2:4]
```

```
## [1] 2 1 2
```

```r
# Elemente die gleich 3 sind
x[x == 3]
```

```
## [1] 3 3
```

```r
# Alle Elemente die kleiner als 3 sind
x[x < 3]
```

```
## [1] 2 1 2 2 1
```

Datentypen

R kennt 4 Datentypen

```r
# numeric (quantitativ)
x <- c(1, 0, 1)
str(x)
```

```
##  num [1:3] 1 0 1
```

```r
# chraracter (string)
x <- c("Anna", "Felix", "Lena")
str(x)
```

```
##  chr [1:3] "Anna" "Felix" "Lena"
```

```r
# factor (nominal), Verwendung als Gruppierungsvariable
x <- c("con", "exp", "exp")
geschlecht <- factor(x,  levels = c("con", "exp"))
str(geschlecht)
```

```
##  Factor w/ 2 levels "con","exp": 1 2 2
```

```r
# logical - TRUE, FALSE
x <- c(TRUE, FALSE, FALSE, TRUE)
x
```

```
## [1]  TRUE FALSE FALSE  TRUE
```

```r
# Beispiel für die Verwendung von logischen Variablen
x <- 1:10
x
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```r
x > 5
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```r
x[x > 5]
```

```
## [1]  6  7  8  9 10
```

Logische Operatoren

```r
# a ist gleich b
a == b

# a ist nicht gleich b
a != b

# a ist grösser als b
a > b

# a ist kleiner als b
a < b

# a ist grösser gleich b
a >= b

# a ist kleiner gleich b
a <= B

# fehlende Wert in a
is.na(a)
```

Mathematische Funktionen

```r
# Addieren
1 + 2
```

```
## [1] 3
```

```r
# Subtrahieren
2 - 1
```

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

```r
# Multiplizieren
2 * 3
```

```
## [1] 6
```

```r
# Dividieren
6 / 3
```

```
## [1] 2
```

```r
# Quadrieren
3^2
```

```
## [1] 9
```

```r
# Quadratwurzel ziehen
sqrt(9)
```

```
## [1] 3
```

```r
# Absolutwert
abs(-2)
```

```
## [1] 2
```

```r
# Beispielvariable erzeugen für Demo
x <- c(3, 2, 1, 2, 2, 1, 3, 4)

# Summe berechnen
sum(x)
```

```
## [1] 18
```

```r
# Maximum finden
max(x)
```

```
## [1] 4
```

```r
# Minimum finden
min(x)
```

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

```r
# Wert auf 3 Stellen runden
Wert <- 3.1234567
round(Wert, 3)
```

```
## [1] 3.123
```

```r
# Mittelwert von x
mean(x)
```

```
## [1] 2.25
```

```r
# Median von x
median(x)
```

```
## [1] 2
```

```r
# Varianz von x
var(x)
```

```
## [1] 1.071429
```

```r
# Standardabweichung von x
sd(x)
```

```
## [1] 1.035098
```

```r
# Spannweite, Variationsbreite (gibt min und max)
range(x)
```

```
## [1] 1 4
```

```r
# Interquartilsabstand
IQR(x)
```

```
## [1] 1.25
```

Datensätze

In einem Datensatz haben alle Variablen die gleiche Länge!

```r
# Einen Datensatz erstellen
df <- data.frame(
  variable1 = 1:3,
  variable2 = c("A", "B", "C")
)
df
```

```
##   variable1 variable2
## 1         1         A
## 2         2         B
## 3         3         C
```

```r
# Gesamten Datensatz anzeigen
View(df)

# Erste 6 Zeilen eines Datensatzes anzeigen
head(df)
```

```
##   variable1 variable2
## 1         1         A
## 2         2         B
## 3         3         C
```

```r
# Anzahl Zeilen und Spalten anzeigen
dim(df)
```

```
## [1] 3 2
```

```r
# Spalte des Datensatzes anzeigen
df[, 1]
```

```
## [1] 1 2 3
```

```r
# Zeile des Datensatzes anzeigen
df[2, ]
```

```
##   variable1 variable2
## 2         2         B
```

```r
# Bestimmte Zelle des Datensatzes anzeigen
df[2, 2]
```

```
## [1] "B"
```

```r
# Variable des Datensatzes
df$variable1
```

```
## [1] 1 2 3
```