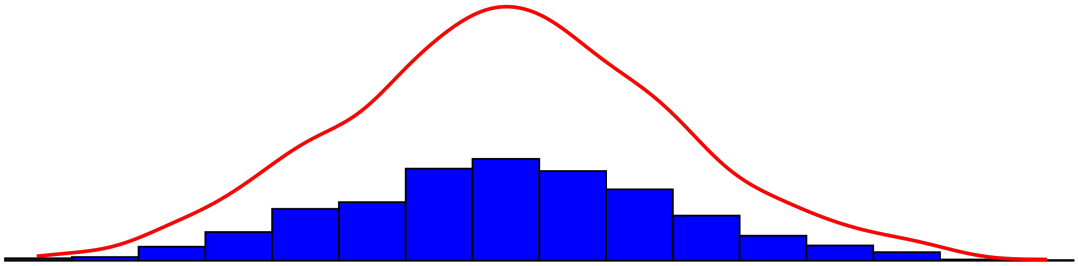


ASTAT

Applied Statistics for Data Science

Author:
Eldar Omerovic
https://www.github.com/omeldar

Project Repository:
https://www.github.com/omeldar/hslu

Credits:
I completed these exercises during my ASTAT class
at the Lucerne University of Applied Sciences and Arts.

June 2024

Inhaltsverzeichnis

1 Einführung in R

1.1 Vektoren

Vektoren in R sind eindimensionale Datenstrukturen, die Elemente des gleichen Typs enthalten. Sie dienen als grundlegende Bausteine für Datenmanipulation und -analyse. Vektoren können numerische Werte, Zeichenketten oder logische Werte speichern und ermöglichen effiziente Berechnungen und Operationen.

1.1.1 Erstellung von Vektoren

In R haben wir die Möglichkeit, Vektoren zu erstellen. Folgend sehen wir, wie wir einen Vektor mithilfe der `combine`-Funktion (also `c()`) erstellen und einer Variablen `x` zuweisen können:

```
1 x <- c(4, 2, 1, 3, 3, 5, 7)
```

1.1.2 Zugriff auf Vektorelemente

Wir können einzelne Werte aus einem Vektor auslesen. Hier ein Beispiel, wie man den dritten Wert in einem Vektor ausliest und in eine Variable speichert:

```
1 dritterWert <- x[3]
```

In diesem Beispiel hat die Variable `dritterWert` den Wert `1`. Beachten Sie, dass die Indexierung bei Vektoren in R mit `1` beginnt, nicht wie in vielen anderen Programmiersprachen mit `0`.

Wir können auch mehrere Werte gleichzeitig aus einem Vektor auslesen. Zum Beispiel können wir den ersten und den dritten Wert wie folgt auslesen:

```
1 werte <- x[c(1, 3)]
```

Hier übergeben wir einen Vektor mit den Indizes der gewünschten Stellen. Der Wert von `werte` nach Ausführung dieses Codes ist `[4 1]`.

1.1.3 Arithmetische Operationen

Mit Vektoren kann man auch Berechnungen durchführen. Hier ein Beispiel, wie man jeden Wert im Vektor um `2` erhöht:

```
1 x + 2
```

Diese Operation ergibt für den Vektor `x` das Resultat `[6 4 3 5 5 7 9]`. Dies funktioniert auch für Subtraktion, Multiplikation und Division.

1.1.4 Filtern von Vektorelementen

Man kann Bedingungen auf einen Vektor anwenden, um bestimmte Werte zu filtern. Hier prüfen wir, welche Werte im Vektor `x` kleiner oder gleich `3` sind:

```
1 x <= 3
```

Dies generiert den Output `[FALSE TRUE TRUE TRUE TRUE FALSE FALSE]`. Um die Werte zu erhalten, die diese Bedingung erfüllen, verwenden wir den folgenden Code:

```
1 x[x <= 3]
```

Dies gibt die Liste `[2 1 3 3]` zurück.

1.1.5 Anwenden von Funktionen

Die Länge eines Vektors bestimmen wir mithilfe der `length()`-Funktion:

```
1 length(x)
```

Dies ergibt den Wert `7`, was der tatsächlichen Anzahl der Elemente im Vektor entspricht.

Ein weiteres Beispiel ist die Funktion `sum`, die die Summe aller Werte im Vektor berechnet:

```
1 sum(x)
```

Diese Funktion gibt für den Vektor `x` die Summe aller Elemente zurück. Um beispielsweise die Summe der Quadrate zu berechnen, kombinieren wir Funktionen:

```
1 sum(x * x)
```

Für den Vektor `x` ergibt dies den Wert `113`, da jeder Wert quadriert und die Ergebnisse anschliessend summiert werden.

1.1.6 Sortieren von Vektoren

In R können wir Vektoren einfach sortieren. Die Funktion `sort()` sortiert die Elemente eines Vektors in aufsteigender Reihenfolge. Hier ein Beispiel:

```
1 sortierterVektor <- sort(x)
```

Dieser Befehl sortiert den Vektor `x` und weist das Ergebnis der Variablen `sortierterVektor` zu. Angenommen, `x` ist `[4, 2, 1, 3, 3, 5, 7]`, dann ist `sortierterVektor [1, 2, 3, 3, 4, 5, 7]`.

1.1.7 Indexsortierung von Vektoren

Mit der Funktion `order()` können wir die Indizes der sortierten Elemente eines Vektors erhalten. Dies ist nützlich, wenn wir wissen möchten, in welcher Reihenfolge die Elemente eines Vektors sortiert würden, ohne den Vektor selbst zu sortieren. Hier ein Beispiel:

```
1 sortierIndizes <- order(x)
```

Dieser Befehl gibt einen Vektor zurück, der die Indizes der sortierten Elemente enthält. Für den Vektor `x [4, 2, 1, 3, 3, 5, 7]` würde `sortierIndizes [3, 2, 4, 5, 1, 6, 7]` sein.

1.1.8 Ersetzen von Werten in Vektoren

In R können wir Werte in einem Vektor leicht ersetzen. Um beispielsweise den vierten Wert im Vektor `x` durch `10` zu ersetzen, verwenden wir den folgenden Befehl:

```
1 x[4] <- 10
```

Nach der Ausführung dieses Codes wäre der Vektor `x [4, 2, 1, 10, 3, 5, 7]`. Dies zeigt, wie man gezielt auf Elemente in einem Vektor zugreifen und deren Werte ändern kann.

1.1.9 NA (not available)

Der Wert NA steht für not available. Diese stehen für fehlende Daten und kommen in Statistiken recht häufig vor. Wenn wir von einem Vektor der NA Werte enthält den Mittelwert berechnen möchten, müssen wir die NA Werte ignorieren. Dies können wir wie folgt machen:

```
1 x <- c(4, 10, 3, NA, NA, 1, 8)
2 mean(x, na.rm = TRUE)
```

Die Option `na.rm = TRUE` entfernt die NA Werte vor der Berechnung des Mittelwerts.

Fürs Sortieren von Vektoren mit NA Werten gibt es eine ähnliche Option:

```
1 x <- c(4, 10, 3, NA, NA, 1, 8)
2 sorted_x <- sort(x, na.last = TRUE)
3 order_x <- order(x, na.last = TRUE)
```

Hier können wir uns dazu entscheiden, ob die NA Werte alle an den Schluss gesetzt werde oder nicht. Möchte man einen Vektor rückwärts sortieren, können wir `decreasing` verwenden. Dies ignoriert NA Werte komplett:

```
1 x <- c(4, 10, 3, NA, NA, 1, 8)
2 mean_narm <- mean(x, na.rm = TRUE)
3 sorted_na_last <- sort(x, na.last = TRUE)
4 sorted_decreasing <- sort(x, decreasing =
      TRUE)
```

```
mean_narm 5.2
sorted_decreasing num [1:5] 10 8 4 3 1
sorted_na_last num [1:7] 1 3 4 8 10 NA NA
x num [1:7] 4 10 3 NA NA 1 8
```

1.2 Aufgaben

1.2.1 1.3 Temperaturrechner

Gegeben sind folgende Temperaturen in Grad Fahrenheit (°F)

$$51.9, \quad 51.8, \quad 51.9, \quad 53$$

a) Bilden Sie einen Vektor `fahrenheit` mit diesen Werten.

b) Berechnen Sie diese Temperaturen in Grad Celsius (°C) um. Die Umrechnungsformel lautet

$$C = \frac{5}{9}(F - 32)$$

Bilden Sie dazu einen Vektor `celsius`.

c) Gegeben sind weitere Temperaturen:

$$48, \quad 48.2, \quad 48, \quad 48.7$$

Bestimmen Sie die Differenz zu den ursprünglichen Temperaturen. Benützen Sie wieder Vektoren.

```
fahrenheit <- c(51.9, 51.8, 51.9, 53)
celsius <- 5/9 * (fahrenheit - 32)
wetetereTemp <- c(48, 48.2, 48, 48.7)
differenzen <- fahrenheit - wetetereTemp
```

Dies führt zu folgenden Werten der Variablen.

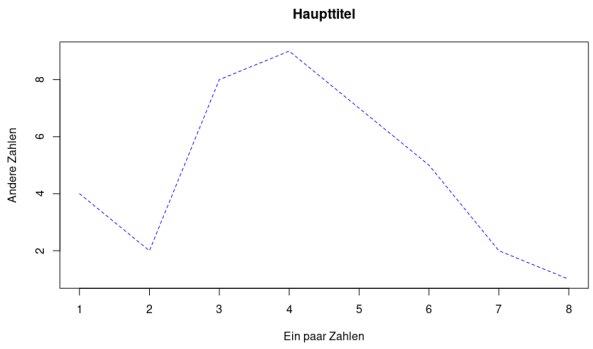
```
celsius num [1:4] 11.1 11 11.1 11.7
differenzen num [1:4] 3.9 3.6 3.9 4.3
fahrenheit num [1:4] 51.9 51.8 51.9 53
weitereTemp num [1:4] 48 48.2 48 48.7
```

1.2.2 1.5 Plotting

Das plotten ist in der Statistik wichtig. Hier ein Beispiel wie man simple Werte plotten kann.

```
z <- c(4, 2, 8, 9, 7, 5, 2, 1)
plot(z,
     type = "l",
     col = "blue",
     lty = 2,
     main = "Haupttitel",
     xlab = "Ein paar Zahlen",
     ylab = "Andere Zahlen"
)
```

Dies sieht dann wie folgt aus.



Nun ist die Aufgabe folgende:

a) Zeichnen Sie eine senkrechte Gerade $x = 3$, durchgezogen, grün.

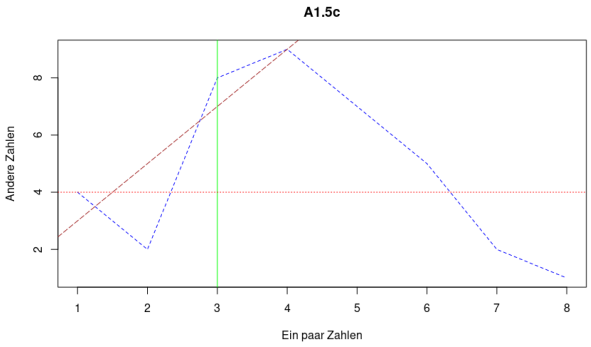
b) Zeichnen Sie eine waagrechte Gerade $y = 4$, gepunktet, rot.

c) Zeichnen Sie eine Gerade $y = 2x + 1$, gestrichelt mit langen Strichen, braun.

Dies kann man folgendermassen erreichen:

```
abline(v = 3, col = "green", lty = 1)
abline(h = 4, col = "red", lty = 3)
abline(a = 1, b = 2, col = "brown", lty = 5)
```

Dies sieht dann wie folgt aus:



Wir sehen hier die hinzugefügten Linien. Wir sehen auch, dass in R wir nicht (x/y) verwenden sondern (h/v). Horizontal und Vertikal.

1.2.3 1.6 Reading Data

a) Laden Sie den Datensatz und speichern Sie diesen unter der Variablen `data` ab.

b) Wählen Sie den Wert der zweiten Zeile und dritten Spalte aus.

c) Wählen Sie die 4. Zeile aus.

d) Wählen Sie die 1. und die 4. Spalte aus. Verwenden Sie dazu die Spaltennamen.

e) Speichern Sie obige Daten unter dem Namen `data1` ab und speichern Sie diese unter dem Namen `weather2.csv`.

f) Wie können Sie herausfinden (mit R natürlich), welches der Name der 3. Spalte ist?

g) Wir möchten den Spalten `Basel` durch `Genf` ersetzen. Wie würden Sie vorgehen?

h) Wir betrachten den Befehl `data3 <- data[order(data[, SZurich"]), ]`.

Um Daten aus einer CSV-Datei in eine Variable in R zu laden, verwenden wir die Funktion `read.csv`. Hier ein Beispiel:

```
# Load data from a CSV file into a variable
data <- read.csv("data.csv")
```

Dieser Code lädt die Daten aus der Datei `data.csv` und speichert sie in der Variablen `data`.

Um einen bestimmten Wert aus den Daten zu lesen, muss man die Reihe und die Spalte angeben. Um also den Wert aus der zweiten Zeile und dritten Spalte zu lesen machen wir folgendes.

```
data[2, 3]
```

Wir können aber auch eine ganze Zeile auslesen, dies machen wir in dem wir keine Spalte angeben.

```
data[4, ]
```

Wir können auch zum Beispiel eine Range von Zeilen wie folgt laden.

```
data[1:5, ]
```

Wir können auch zum Beispiel alle Zeilen aber nur 2 bestimmte Spalten anzeigen. Vergleichbar mit der Projektion mithilfe SELECT im SQL.

```
spalten_d <- data[1:4, c("Luzern", "Zurich")]
print(spalten_d)
```

Dies generiert dann folgenden Output in der Konsole.

	Luzern	**Zurich**
Jan	2	4
Feb	5	0
Mar	10	8
Apr	16	17

Auch können wir bearbeitete Daten in Dateien schreiben. So können wir die Daten persistieren. Dies machen wir wie folgt.

```
write.csv(spalten_d, "source/data/weather2.
    csv", row.names = FALSE)
```

Dies speichert den Output von oben ohne die Monate in eine neue Datei `weather2.csv`.

Wenn man die Spaltennamen nicht kennt oder man etwas dynamisch lösen will, kann diesen auch herauslesen:

```
name_dritte_spalte <- colnames(data)[3]
```

So haben wir nun den Wert `Chur` in die Variable `name_dritte_spalte` gespeichert.

Wir können Spalten auch ersetzen. Dies können wir wie folgt machen:

```
data$Genf <- data$Basel
data$Basel <- NULL
```

Wir kopieren also die Spalte zuerst in einen neuen Wert in `data` und löschen dann die Spalte `Basel` in dem wir sie `NULL` setzen.

Auch Sortierfunktionen können wir auf diese Daten machen. Zum Beispiel können wir die Daten basierend auf der Spalte `Zurich` sortieren. Dies machen wir wie folgt.

```
data3 <- data[order(data[, "Zurich"]), ]
print(data3)
```

Dies generiert folgenden Output.

	Luzern	**Chur**	**Zurich**	**Genf**
Feb	5	1	0	6
Jan	2	-3	4	5
Mar	10	13	8	11
Apr	16	14	17	12
May	21	21	20	23
Jun	25	23	27	21

2 Eindimensionale deskriptive Statistik

2.1 Daten und Statistiken

Daten und Statistiken spielen in vielen Bereichen unseres Alltags eine zentrale Rolle. Sie finden Anwendung in Prognosen bei Wahlen, der Funktionsweise von Suchmaschinen wie Google, Wettervorhersagen und der Analyse von Börsenkursen. Grundsätzlich unterscheidet man zwischen eindimensionalen Datensätzen, wie Listen von Werten (z.B. Körpergrössen), und zweidimensionalen Datensätzen, die als Tabellen organisiert sind und mehrere Merkmale (z.B. Person, Grösse, Gewicht) umfassen.

2.2 Deskriptive Statistik

Die deskriptive Statistik zielt darauf ab, Daten durch numerische Kennwerte und graphische Darstellungen zusammenzufassen. Durch die Berechnung von Kennzahlen und die Erstellung von Diagrammen können grosse Datensätze verständlicher gemacht werden, indem zentrale Tendenzen und Streuungen sichtbar gemacht werden.

2.2.1 Lageparameter

Lageparameter dienen dazu, die „Mitte“ eines Datensatzes zu bestimmen:

Arithmetisches Mittel: Das arithmetische Mittel, oder der Durchschnitt, wird berechnet, indem alle Werte eines Datensatzes summiert und durch die Anzahl der Werte geteilt werden. Es ist sensitiv gegenüber Ausreissern.

Das arithmetische Mittel wird berechnet als:

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

Angenommen, wir haben die folgenden Daten: 4, 7, 10, 5, 8. Das arithmetische Mittel berechnet sich wie folgt:

$$\bar{x} = \frac{4 + 7 + 10 + 5 + 8}{5} = \frac{34}{5} = 6.8$$

Median: Der Median ist der Wert, der die Daten in zwei gleich grosse Hälften teilt. Er ist weniger anfällig für Ausreisser und gibt eine bessere Darstellung der zentralen Tendenz bei schiefen Verteilungen.

Für die Daten 3, 5, 7, 9, 11 ist der Median der mittlere Wert:

$$\text{Median} = 7$$

Für die Daten 3, 5, 7, 9, 11, 13 ist der Median der Durchschnitt der beiden mittleren Werte:

$$\text{Median} = \frac{7 + 9}{2} = 8$$

Quantile: Quantile teilen die Daten in gleiche Teile. Die wichtigsten Quantile sind Quartile, die die Daten in vier Teile teilen. Der Median ist dabei das zweite Quartil.

Angenommen, wir haben die folgenden Daten: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20. Die Quartile berechnen sich wie folgt:

- 1. Quartil (Q1): Der Wert bei 25% der Daten.
- 2. Quartil (Q2, Median): Der Wert bei 50% der Daten.
- 3. Quartil (Q3): Der Wert bei 75% der Daten.

In diesem Beispiel sind die Quartile:

$$Q1 = 6, \quad Q2 = 11, \quad Q3 = 16$$

2.2.2 Streuungsparameter

Streuungsparameter beschreiben die Verteilung der Daten um die zentrale Tendenz:

Empirische Varianz: Die empirische Varianz misst die durchschnittliche quadratische Abweichung der Werte vom arithmetischen Mittel. Sie wird berechnet, indem die Differenzen der Werte zum Mittelwert quadriert, summiert und durch die Anzahl der Werte geteilt werden.

Die empirische Varianz wird berechnet als:

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

Für die Daten 4, 7, 10, 5, 8 und das berechnete arithmetische Mittel von 6.8:

$$s^2 = \frac{(4-6.8)^2 + (7-6.8)^2 + ...}{5-1}$$

$$= \frac{7.84 + 0.04 + 10.24 + 3.24 + 1.44}{4} = 5.2$$

Standardabweichung: Die Standardabweichung ist die Quadratwurzel der Varianz und gibt die Streuung der Daten in derselben Einheit wie die ursprünglichen Werte an. Eine hohe Standardabweichung weist auf eine grosse Streuung der Daten hin.

Die Standardabweichung ist die Quadratwurzel der Varianz:

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

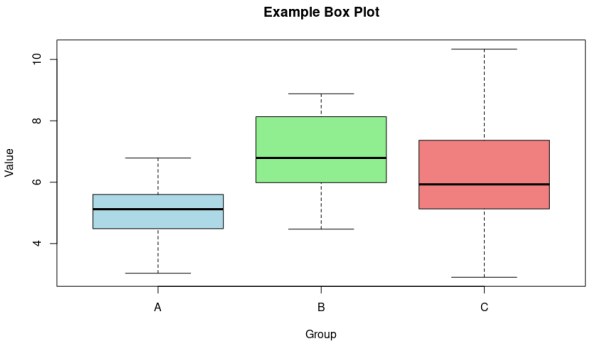
Für die oben berechnete Varianz von 5.2:

$$s = \sqrt{5.2} \approx 2.28$$

Quartilsdifferenz: Die Quartilsdifferenz, auch Interquartilsabstand genannt, ist die Differenz zwischen dem oberen (dritten) und dem unteren (ersten) Quartil. Sie misst die Streuung der mittleren 50% der Daten und ist robust gegenüber Ausreissern.

2.3 Graphische Darstellungen

Ein wichtiges Werkzeug der deskriptiven Statistik sind graphische Darstellungen wie Boxplots. Ein Boxplot visualisiert die Verteilung eines Datensatzes. Der Boxplot zeigt den Median, das erste und dritte Quartil, sowie mögliche Ausreisser.



Box: Der Bereich zwischen erstem (unterem) und drittem (oberem) Quartil wird als Box dargestellt. Der Median wird als Linie innerhalb der Box angezeigt.

Whiskers: Linien, die von der Box zu den kleinsten und grössten „normalen“ Werten reichen, werden als Whiskers bezeichnet.

Ausreisser: Werte, die ausserhalb des normalen Bereichs liegen, werden als einzelne Punkte dargestellt.

2.4 Aufgaben

2.4.1 A2.2 Median und Mittelwert

In einer Klasse wurden in einer Statistik-Prüfung folgende Noten geschrieben:

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

a) Ändern Sie drei Noten im Datensatz so ab, dass der Median gleich bleibt, aber der Mittelwert sich stark ändert. Verwenden Sie dazu den `sort(...)`-Befehl.

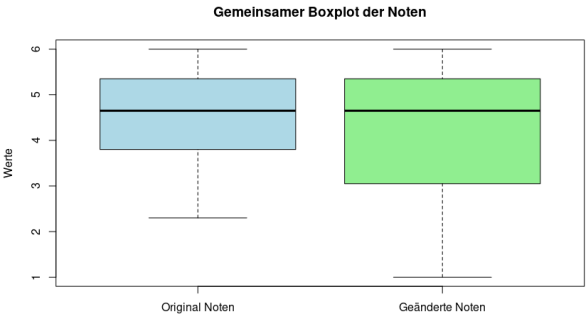
b) Erstellen Sie zu den beiden Datensätzen einen gemeinsamen Boxplot. Was erkennen Sie?

Lösung Hier ist der R-Code zur Lösung der Aufgabe:

Um den Mittelwert möglichst stark zu verändern müssen wir einen Wert möglichst stark ändern. Da hier die 12. und 13. Note für die Berechnung des Medians verwendet werden, können wir die vorangehenden Werte auf die Note 1 setzen. So erreichen wir die grösste Veränderung im Mittelwert ohne den Median zu verändern.

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

#a)

#neuen vektor erstellen
changedNoten <- sort(noten)

# 3 Werte abndern
changedNoten[11] <- 1
changedNoten[10] <- 1
changedNoten[9] <- 1

# kontrolle
medVorher <- median(noten)
medNachher <- median(changedNoten)
mitVorher <- mean(noten)
mitNachher <- mean(changedNoten)

medVorher == medNachher
mitVorher != mitNachher

print(paste("Median vorher: ", medVorher))
print(paste("Median nachher: ", medNachher))
print(paste("Mittw vorher: ", mitVorher))
print(paste("Mittw nachher: ", mitNachher))

#b)
boxplot(noten, changedNoten,
       names = c("Original Noten", "Genderte
            Noten"),
       col = c("lightblue", "lightgreen"),
       main = "Gemeinsamer Boxplot der Noten
            ",
       ylab = "Werte")
```

Dies sieht dann im Boxplot wie folgt aus:



2.4.2 A2.3 Interpretieren eines Boxplot

In einer Untersuchung in England wurden das Alter (in Jahren) und die Körpergrösse (in cm) von 170 Ehepaaren untersucht.

a) Datei `mannfrau.csv` einlesen

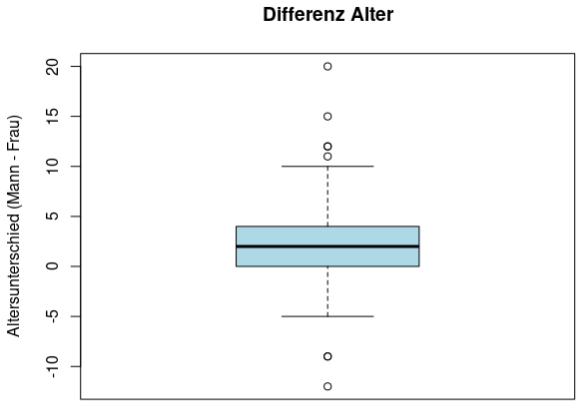
b) Erstellen eines Boxplot für die Differenz des Alters zwischen Ehemännern und Ehefrauen

c) Interpretieren Sie im Boxplot den Median und die Quartile. Was können Sie über Ausreisser sagen?

Lösung Aufgabe Mit folgendem R Skript kann man die Aufgabe lösen.

```
data <- read.csv("source/data/mannfrau.csv")

differenz <- data$alter.mann - data$alter.
    frau
boxplot(differenz,
       main = "Differenz Alter",
       ylab = "Altersunterschied (Mann - Frau
            )",
       col = "lightblue")
```

Mit diesem Skript generiert man dann folgenden Plot:



Kommen wir zur Analyse dieses Boxplots:

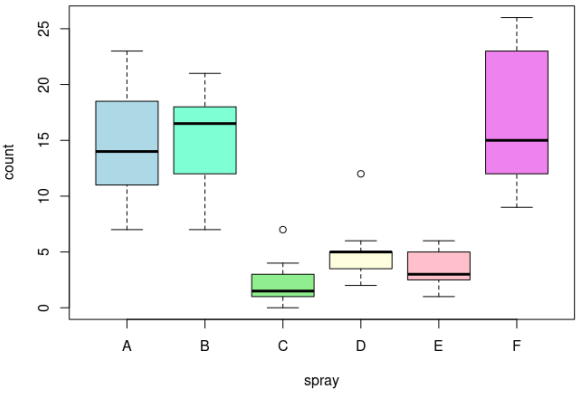
- Wir sehen, dass der Median sich bei etwa 2 befindet. Somit wissen wir, dass die Altersdifferenz bei der Hälfte der Ehepaaren kleiner als zwei und die andere hälfte grösser als 2 ist. Hier ist zu beachten, dass dies bedeutet das der Ehemann 2 Jahre älter ist, wenn die Ehefrau älter wäre, wäre der Wert bei `-2`.
- Das untere Quartil ist bei ungefähr 0. Das heisst, bei 25% aller untersuchten Ehepaare ist die Frau älter als der Mann.
- Das obere Quartil befindet sich bei ca. 5. Das heisst, bei 25% aller untersuchten Ehepaare ist der Mann mehr als 5 Jahre älter als die Frau.
- Die Hälfte der Ehepaare hat also einen Altersunteschied (Mann älter als Frau) zwischen 0 und 5 Jahren.
- Der maximale Unterschied ist `20` Jahre und das Minimum `-10` Jahre. In einem Fall also war die Frau mehr als 10 Jahre älter als der Mann.

2.4.3 A2.4 tapply

In dieser Aufgabe lernen wir die Funktion `tapply` zu verwenden. Wir nutzen das in R enthaltene `InsectSprays`-DataSet.

```
1 tapply(InsectSprays[, "count"], InsectSprays
      [, "spray"], FUN = mean)
2 tapply(InsectSprays$count, InsectSprays$spray
      , mean)
3
4 boxplot(count ~ spray,
5         data = InsectSprays,
6         col=c("lightblue", "aquamarine", "
              lightgreen", "lightyellow", "pink
              ", "violet"))
```

Wir sehen hier zwei verschiedene Schreibweisen für die Funktion `tapply`. Mit dem Plot wird ausserdem folgender Graph generiert.



Man sieht anhand vom Boxplot sehr gut, welche Spray's besser und welche schlechter sind. Ausserdem sieht man auch, welche Sprays teilweise Ausreisser haben und welche konstant überzeugen können.

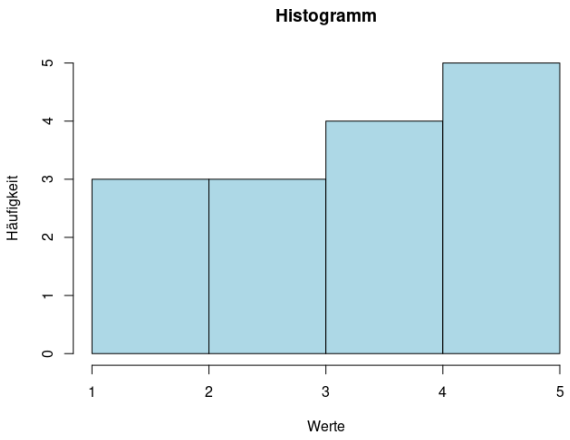
3 Histogramm, zweidimensionale deskriptive Statistik

3.1 Histogramme

Ein Histogramm ist eine grafische Darstellung der Verteilung eines Datensatzes. Es zeigt die Häufigkeit von Werten in verschiedenen Klassen (Bins) an. Die Höhe jedes Balkens im Histogramm entspricht der Anzahl der Datenpunkte innerhalb dieser Klasse.

```
daten <- c(1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5,
    5, 5, 5, 5)
hist(daten, main="Histogramm", xlab="Werte",
    ylab="Haeufigkeit", col="lightblue",
    border="black")
```

Dieses Histogramm zeigt die Häufigkeit der Werte 1 bis 5 in den Beispiel-Daten.

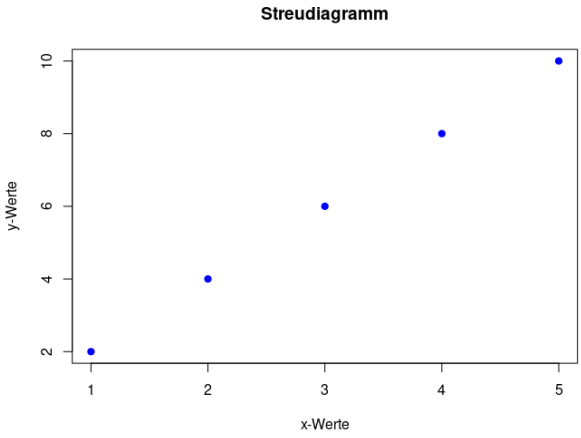


3.2 Zweidimensionale deskriptive Statistik

Die zweidimensionale deskriptive Statistik befasst sich mit der Analyse von zwei Variablen gleichzeitig. Ein häufig verwendetes Werkzeug ist das Streudiagramm (Scatterplot), das die Beziehung zwischen zwei Variablen grafisch darstellt.

```
# Beispiel-Daten
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 6, 8, 10)
# Streudiagramm erzeugen
plot(x, y, main="Streudiagramm", xlab="x-
    Werte", ylab="y-Werte", col="blue", pch
    =19)
```

Dieses Streudiagramm zeigt eine lineare Beziehung zwischen den Variablen x und y.



3.3 Korrelation

Die Korrelation misst die Stärke und Richtung der linearen Beziehung zwischen zwei Variablen. Der Korrelationskoeffizient r liegt zwischen -1 und 1.

```
# Beispiel-Daten
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 6, 8, 10)
# Korrelation berechnen
cor(x, y)
```

In diesem Beispiel beträgt der Korrelationskoeffizient $r = 1$, was auf eine perfekte positive lineare Beziehung hinweist.

3.4 Kontingenztabellen

Kontingenztabellen (Cross-Tabulations) werden verwendet, um die Häufigkeit von Beobachtungen in verschiedenen Kategorien für zwei oder mehr kategoriale Variablen darzustellen.

```
# Beispiel-Daten
geschlecht <- c("m", "w", "m", "w", "m")
kurs <- c("A", "B", "A", "B", "A")
# Kontingenztabelle erstellen
table(geschlecht, kurs)
```

Diese Tabelle zeigt die Häufigkeit von männlichen und weiblichen Teilnehmern in den Kursen A und B.

Geschlecht	Kurs A	Kurs B
m	3	0
w	0	2

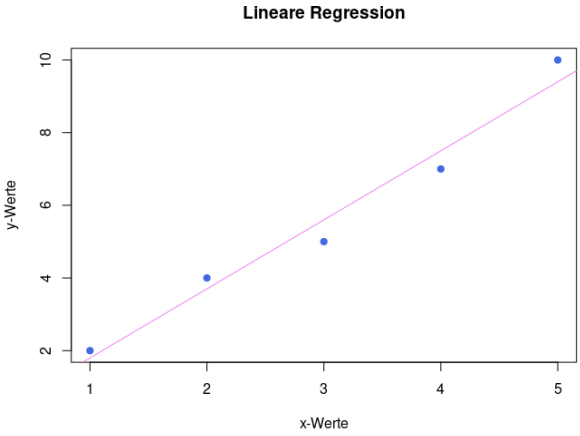
3.5 Die lineare Regression

Die lineare Regression wird verwendet, um die Beziehung zwischen einer abhängigen und einer unabhängigen Variablen zu modellieren. Das Ziel ist es, eine Gerade zu finden, die die Daten bestmöglich beschreibt.

```
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 5, 7, 10)
modell <- lm(y ~ x)
summary(modell)
```

Die Ausgabe eines Regressionsmodells liefert die geschätzten Koeffizienten der Regressionsgeraden, die anzeigen, wie stark und in welche Richtung die unabhängigen Variablen die abhängige Variable beeinflussen. Darüber hinaus enthält die Ausgabe statistische Kennzahlen, die die Qualität der Anpassung des Modells an die Daten bewerten. Diese Kennzahlen, wie das Bestimmtheitsmass R^2 und der p-Wert, geben Auskunft darüber, wie gut die Regressionsgerade die tatsächlichen Datenpunkte erklärt und ob die beobachteten Beziehungen statistisch signifikant sind.

```
plot(x, y, main="Lineare Regression", xlab="x
    -Werte", ylab="y-Werte", col="royalblue
    ", pch=19)
abline(modell, col="violet")
```



Dieses Diagramm zeigt die Datenpunkte sowie die Regressionsgerade, die die Beziehung zwischen x und y beschreibt.

3.6 Aufgaben

3.6.1 A3.2 Lineare Regression

In dieser Aufgabe geht es darum herauszufinden, ob grosse Frauen auch grosse Männer heiraten.

a) Datei `mannfrau.csv` einlesen

b) Streudiagramm erzeugen aus `groesse.mann` und `groesse.frau` mit Regressionsgerade. Das Streudiagramm soll dann interpretiert werden.

c) Bestimmen Sie die Koeffizienten der Regressionsgeraden

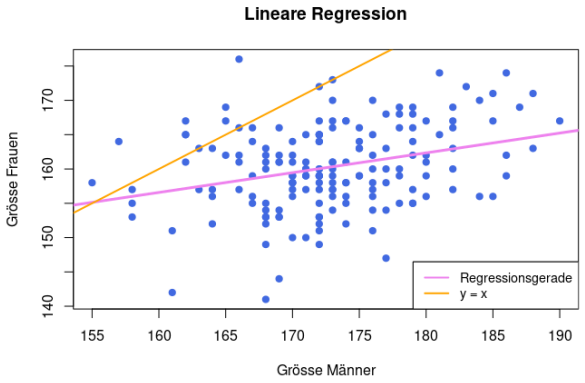
$$y = a + bx$$

d) Zeichnen Sie die gerade $y = x$ in den Plot ein. Wie ist der Plot mit Bezug zu dieser Geraden zu interpretieren?

Zur Lösung dieser Aufgabe können wir folgendes R Skript nutzen.

```r
data <- read.csv("source/data/mannfrau.csv")

groesseMaenner <- data$groesse.mann
groesseFrauen <- data$groesse.frau
modell <- lm(groesseFrauen ~ groesseMaenner)

plot(groesseMaenner, groesseFrauen, main="
    Lineare Regression", xlab="Groesse
    Maenner", ylab="Groesse Frauen", col="
    royalblue", pch=19)
abline(modell, col="violet", lwd = 3)
abline(a = 0, b = 1, col="orange", lwd = 2)

legend("bottomright", legend=c("
    Regressionsgerade", "y = x"), col=c("
    violet", "orange"), lty=1, lwd=2, cex
    =0.9)

# Koeffizienten bestimmen
koeffizienten <- coef(modell)
```

Wir sehen hier, dass wir das Streudiagramm erstellen und dann mithilfe linearer Regression eine Gerade durch diese bilden. Auch fügen wir noch eine Gerade $y = x$ hinzu. Der Plot sieht dann wie folgt aus.

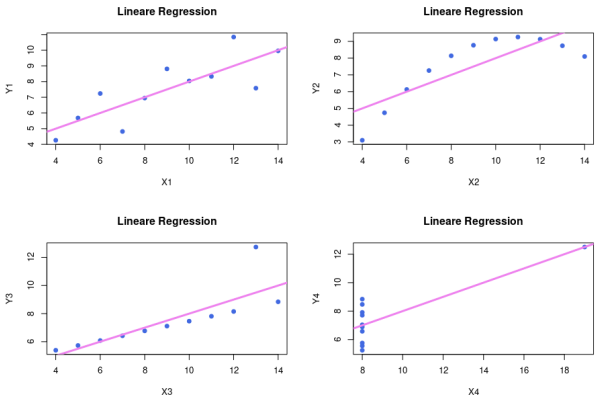


Erstens zur Regressionsgeraden. Die Regressionsgerade $y = 110.444 + 0.288x$ repräsentiert die Linie, die so durch das Streudiagramm verläuft, dass sie möglichst nahe an allen Punkten liegt. Je weiter oben ein Punkt liegt, desto grösser ist die Frau. Je weiter rechts ein Punkt liegt, desto grösser ist der Mann. Da die Regressionsgerade eine positive Steigung von links unten nach rechts oben aufweist, können wir daraus schliessen, dass grössere Männer tendenziell grössere Frauen haben.

Zusätzlich haben wir die Gerade $y = x$ (orange). Punkte auf dieser Linie repräsentieren Paare, bei denen beide Partner gleich gross sind. Punkte links von dieser Linie stehen für Paare, bei denen die Frau grösser ist als der Mann. Punkte rechts davon stehen für Paare, bei denen der Mann grösser ist als die Frau. Auch hier wird deutlich, dass in den meisten Ehen die Männer grösser als die Frauen sind.

3.6.2 Weitere Beispiele zu linearer Regression

```r
par(mfrow = c(2,2))

modell1 <- lm(anscombe$y1 ~ anscombe$x1)
modell2 <- lm(anscombe$y2 ~ anscombe$x2)
modell3 <- lm(anscombe$y3 ~ anscombe$x3)
modell4 <- lm(anscombe$y4 ~ anscombe$x4)

plot(anscombe$x1, anscombe$y1,
     main="Lineare Regression",
     xlab="X1",
     ylab="Y1",
     col="royalblue",
     pch=19)
abline(modell1, col="violet", lwd=3)

plot(anscombe$x2, anscombe$y2,
     main="Lineare Regression",
     xlab="X2",
     ylab="Y2",
     col="royalblue",
     pch=19)
abline(modell2, col="violet", lwd=3)

...
```



4 Korrelation, Wahrscheinlichkeitsmodelle

4.1 Korrelation

Die Korrelation misst die Stärke und Richtung der linearen Beziehung zwischen zwei Variablen. Der Korrelationskoeffizient r liegt zwischen -1 und 1. Ein Wert von $r = 1$ bedeutet eine perfekte positive lineare Beziehung, $r = -1$ eine perfekte negative lineare Beziehung und $r = 0$ keine lineare Beziehung.

- **Positive Korrelation**: Wenn $r > 0$, bedeutet dies, dass mit zunehmenden Werten einer Variable auch die Werte der anderen Variable zunehmen.
- **Negative Korrelation**: Wenn $r < 0$, bedeutet dies, dass mit zunehmenden Werten einer Variable die Werte der anderen Variable abnehmen.
- **Keine Korrelation**: Wenn $r \approx 0$, besteht keine lineare Beziehung zwischen den Variablen.

```
# Beispiel-Daten
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 6, 8, 10)
# Korrelation berechnen
cor(x, y)
```

In diesem Beispiel beträgt der Korrelationskoeffizient $r = 1$, was auf eine perfekte positive lineare Beziehung hinweist.

4.2 Wahrscheinlichkeitsmodelle

Wahrscheinlichkeitsmodelle beschreiben, wie Daten sich unter bestimmten Bedingungen verteilen. Zwei wichtige Konzepte sind die bedingte Wahrscheinlichkeit und die Unabhängigkeit von Ereignissen.

4.2.1 Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit $P(A|B)$ ist die Wahrscheinlichkeit, dass das Ereignis A eintritt, gegeben, dass B bereits eingetreten ist. Sie wird berechnet durch:

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

- Beispiel: Die Wahrscheinlichkeit, dass es regnet, gegeben, dass es bewölkt ist.

4.2.2 Unabhängigkeit

Zwei Ereignisse A und B sind unabhängig, wenn das Eintreten von A keinen Einfluss auf die Wahrscheinlichkeit des Eintretens von B hat. Mathematisch ausgedrückt:

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

- Beispiel: Das Werfen eines Würfels und das Ziehen einer Karte aus einem Kartenspiel sind unabhängige Ereignisse.

4.3 Nützliche R-Funktionen

Am Ende der Zusammenfassung sind einige nützliche R-Funktionen aufgeführt, die in diesem Kapitel verwendet wurden.

`cor(x, y)`
: Berechnet den Korrelationskoeffizienten zwischen den Variablen x und y.

`prop.table(table)`
: Berechnet die bedingten Wahrscheinlichkeiten aus einer Kontingenztabelle.

`table(x, y)`
: Erstellt eine Kontingenztabelle für die Variablen x und y.

4.4 Aufgaben

4.4.1 A4.2 Korrelationskoeffizient

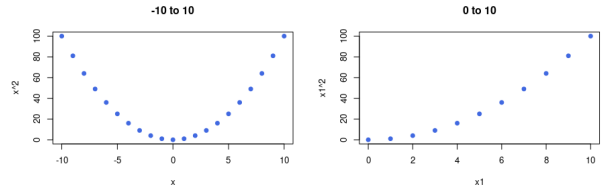
a) Erzeugen Sie den Vektor `t.x` mit den Werten `-10, -9, ..., 9, 10` und den Vektor `t.x1` mit den Werten `0, 1, ..., 9, 10`. Erzeugen Sie dann die Vektoren `t.y` und `t.y1`, deren Elemente die Quadratwerte der entsprechenden Elemente von `t.x` bzw. `t.x1` enthalten.

b) Zeichnen Sie die Streudiagramme `t.y` vs. `t.x` und `t.y1` vs. `t.x1`. Benützen Sie die R-Funktion `plot()`.

c) Berechnen Sie die Korrelationskoeffizienten zwischen `t.x` und `t.y` bzw. zwischen `t.x1` und `t.y1`. Benützen Sie die R-Funktion `cor()`. Warum sind die beiden Korrelationen so verschieden?

Um diese Aufgabe zu lösen, können wir folgendes R-Skript verwenden, welches uns den Plot erstellt sowie die Korrelationskoeffizienten berechnet.

```
1  t.x <- seq(from = -10, to = 10, by = 1)
2  t.x1 <- seq(from = 0, to = 10, by = 1)
3  t.y <- t.x^2
4  t.y1 <- t.x1^2
5
6  plot(t.x, t.y, main = "-10 to 10", xlab="x",
7       ylab="x^2",
8       col="royalblue",
9       pch=19)
10 plot(t.x1, t.y1, main = "0 to 10", xlab="x1",
11      ylab="x1^2",
12      col="royalblue",
13      pch=19)
14
15 cor(t.x, t.y)
16 cor(t.x1, t.y1)
```



Der Korrelationskoeffizient für den ersten Plot ist `0`, für die zweite lineare Beziehung ist der Korrelationskoeffizient `0.9631427`.

Dies bedeutet, dass die Beziehung zwischen `t.x` und `t.y` nicht linear ist, da deren Korrelationskoeffizient `0` beträgt. Dies erkennt man auch am Graphen. Während `t.x` stetig wächst, sinkt und steigt `t.y`.

Für die Beziehung zwischen `t.x1` und `t.y1` lässt sich sagen, dass beide Variablen nur zunehmen. Da jedoch `t.y1` quadratisch wächst und `t.x1` linear, beträgt deren Korrelationskoeffizient nicht ganz `1`, sondern `0.9631427`.

4.4.2 A4.4 Wahrscheinlichkeit 1

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

a) Beschreiben Sie den Ereignisraum in Form von Elementarereignissen.

Bei einem Zufallsexperiment, bei dem ein roter und ein blauer Würfel geworfen werden, besteht jedes Elementarereignis aus einem Paar (r, b), wobei r die Augenzahl des roten Würfels und b die Augenzahl des blauen Würfels ist. Es gibt insgesamt $6 \times 6 = 36$ Elementarereignisse.

b) Wie gross ist die Wahrscheinlichkeit eines einzelnen Elementarereignisses?

Da alle Elementarereignisse gleich wahrscheinlich sind, beträgt die Wahrscheinlichkeit eines einzelnen Elementarereignisses $\frac{1}{36}$.

c) Berechnen Sie die Wahrscheinlichkeit, dass das Ereignis E_1 „Die Augensumme ist 7“ eintritt.

Die möglichen Paare sind (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) und (6, 1). Es gibt 6 günstige Ereignisse. Die Wahrscheinlichkeit beträgt daher:

$$P(E_1) = \frac{6}{36} = \frac{1}{6}$$

d) Wie gross ist die Wahrscheinlichkeit, dass das Ereignis E_2 „Die Augensumme ist kleiner als 4“ eintritt.

Die möglichen Paare sind (1, 1), (1, 2), (2, 1). Es gibt 3 günstige Ereignisse. Die Wahrscheinlichkeit beträgt daher:

$$P(E_2) = \frac{3}{36} = \frac{1}{12}$$

e) Bestimmen Sie $P(E_3)$ für das Ereignis E_3 „Beide Augenzahlen sind ungerade“.

Die möglichen Paare sind (1, 1), (1, 3), (1, 5), (3, 1), (3, 3), (3, 5), (5, 1), (5, 3) und (5, 5). Es gibt 9 günstige Ereignisse. Die Wahrscheinlichkeit beträgt daher:

$$P(E_3) = \frac{9}{36} = \frac{1}{4}$$

f) Berechnen Sie $P(E_2 \cup E_3)$

Da die Ereignisse disjunkt sind, kann die Wahrscheinlichkeit einfach addiert werden:

$$P(E_2 \cup E_3) = P(E_2) + P(E_3)$$

$$= \frac{1}{12} + \frac{1}{4} = \frac{1}{12} + \frac{3}{12} = \frac{4}{12} = \frac{1}{3}$$

4.4.3 A4.5 Wahrscheinlichkeit 2

Die Ereignisse aus A und B seien unabhängig mit Wahrscheinlichkeiten $P(A) = 3/4$ und $P(B) = 2/3$. Berechnen Sie die Wahrscheinlichkeiten folgender Ereignisse:

Lösung

a) Beide Ereignisse treten ein.

$$P(A \cap B) = P(A) \times P(B) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$$

b) Mindestens eines von beiden Ereignissen tritt ein.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{3}{4} + \frac{2}{3} - \frac{1}{2} = \frac{17}{12} - \frac{1}{2} = \frac{17}{12} - \frac{6}{12} = \frac{11}{12}$$

c) Höchstens eines von beiden Ereignissen tritt ein.

$$P(A^c \cup B^c) = 1 - P(A \cap B) = 1 - \frac{1}{2} = \frac{1}{2}$$

d) Keines der beiden Ereignisse tritt ein.

$$P(A^c \cap B^c) = P(A^c) \times P(B^c)$$
$$= (1 - P(A)) \times (1 - P(B)) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$$

e) Genau eines der Ereignisse tritt ein.

$$P(A \Delta B) = P(A \cup B) - P(A \cap B)$$
$$= \frac{11}{12} - \frac{1}{2} = \frac{11}{12} - \frac{6}{12} = \frac{5}{12}$$

4.4.4 A4.6 Wahrscheinlichkeit 3

Ein Einsturz eines Gebäudes in Tokio kann durch zwei voneinander unabhängigen Ereignissen verursacht werden.

1. Gebäude kann durch zwei voneinander unabhängigen Ereignissen zum Einsturz gebracht werden.

$$P(E_1) = 0.04 \quad \text{und} \quad P(E_2) = 0.08$$

2. **Wahrscheinlichkeit für den Einsturz des Gebäudes:**

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$
$$= 0.04 + 0.08 - (0.04 \times 0.08) = 0.12 - 0.0032$$
$$= 0.1168$$

5 Zufallsvariable, Wahrscheinlichkeitsverteilung

5.1 Zufallsvariablen

Eine Zufallsvariable ist eine Funktion, die jedem möglichen Ergebnis eines Zufallsexperiments eine Zahl zuordnet. Beispiel: Ziehen einer Karte aus einem Stapel von 36 Karten (Schweizer Kartenspiel) und Zuordnung von Werten zu den Karten:

- 6, 7, 8, 9 haben Wert 0
- 10 hat den Wert 10
- Bube hat den Wert 2
- Dame hat den Wert 3
- König hat den Wert 4
- Ass hat den Wert 11

Durch diese Zuordnung können die Ergebnisse von Ziehungen miteinander verglichen werden:

```
1. Ziehung:
6, Dame, König -> 0 + 3 + 4 = 7

2. Ziehung:
8, Bube, Ass -> 0 + 2 + 11 = 13
```

Die zweite Ziehung ist somit besser.

Ein weiteres Beispiel ist das Werfen zweier Würfel. Die Augensumme kann als Zufallsvariable X betrachtet werden. Mögliche Werte sind die Zahlen von 2 bis 12. Die Wahrscheinlichkeit für jede Augensumme kann berechnet werden.

5.2 Wahrscheinlichkeitsverteilung

Die Wahrscheinlichkeitsverteilung einer Zufallsvariablen beschreibt, mit welcher Wahrscheinlichkeit die einzelnen möglichen Werte angenommen werden. Beispiel: Ziehen einer Karte aus einem Kartenspiel mit den oben definierten Werten:

- $P(X = 0) = \frac{16}{36} = \frac{4}{9}$
- $P(X = 2) = \frac{4}{36} = \frac{1}{9}$
- $P(X = 3) = \frac{4}{36} = \frac{1}{9}$
- $P(X = 4) = \frac{4}{36} = \frac{1}{9}$
- $P(X = 10) = \frac{4}{36} = \frac{1}{9}$
- $P(X = 11) = \frac{4}{36} = \frac{1}{9}$

Diese Wahrscheinlichkeiten summieren sich zu 1, was eine notwendige Eigenschaft einer Wahrscheinlichkeitsverteilung ist.

5.3 Erwartungswert und Standardabweichung

Der Erwartungswert einer Zufallsvariablen gibt die mittlere Lage der Verteilung an und wird berechnet als:

$$E(X) = \sum_i x_i \cdot P(X = x_i)$$

Beispiel für einen fairen Würfel:

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Die Standardabweichung misst die Streuung der Verteilung um den Erwartungswert und wird berechnet als:

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\sum_i (x_i - E(X))^2 \cdot P(X = x_i)}$$

Beispiel für einen fairen Würfel in R:

```
1 x <- 1:6
2 p <- rep(1/6, 6)
3 E_X <- sum(x * p)
4 var_X <- sum((x - E_X)^2 * p)
5 sd_X <- sqrt(var_X)
```

Für das Kartenspiel mit den vorher definierten Wahrscheinlichkeiten ergibt sich der Erwartungswert:

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

```
1 x <- c(0, 2, 3, 4, 10, 11)
2 p <- c(4/9, 1/9, 1/9, 1/9, 1/9, 1/9)
3 E_X <- sum(x * p)
4 var_X <- sum((x - E_X)^2 * p)
5 sd_X <- sqrt(var_X)
```

5.4 Beispiel zur Wahrscheinlichkeitsverteilung

Die Wahrscheinlichkeitsverteilung der Augensumme zweier Würfel:

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

Beispielaufgaben:

a) Wie gross ist die Wahrscheinlichkeit, die Augensumme 6 zu würfeln?

$$P(X=6) = \frac{5}{36}$$

b) Wie gross ist die Wahrscheinlichkeit, die Augensumme 6 oder 8 zu würfeln?

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

c) Wie gross ist die Wahrscheinlichkeit, höchstens die Augensumme 3 zu würfeln?

$$P(X \leq 3) = P(X=2) + P(X=3)$$
$$= \frac{1}{36} + \frac{2}{36} = \frac{3}{36} = \frac{1}{12}$$

d) Wie gross ist die Wahrscheinlichkeit, mindestens die Augensumme 3 zu würfeln?

$$P(X \geq 3) = 1 - P(X=2) = 1 - \frac{1}{36} = \frac{35}{36}$$

e) Wie gross ist die Wahrscheinlichkeit, eine Augensumme von 3 bis 5 zu würfeln?

$$P(3 \leq X \leq 5)$$
$$= P(X=3) + P(X=4) + P(X=5)$$
$$= \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{9}{36} = \frac{1}{4}$$

5.5 Aufgaben

5.5.1 A5.3 Wahrscheinlichkeitsverteilung

Man wirft eine Münze dreimal. Die Zufallsgrösse X gibt an, wie oft dabei `Zahl` geworfen wurde.

a) Stellen Sie die Verteilungsfunktion als Tabelle auf

b) Berechnen Sie die Wahrscheinlichkeit dafür, dass genau 2 mal Zahl geworfen wird.

c) Berechnen Sie die Wahrscheinlichkeit dafür, dass mindestens 2 mal Zahl geworfen wird.

d) Berechnen Sie die Wahrscheinlichkeit dafür, dass höchstens 1 mal Zahl geworfen wird.

x	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

```
p <- c(1/8, 3/8, 3/8, 1/8)
#b)
sum(p[3])
#c)
sum(p[3:4])
#d)
sum(p[1:2])
```

6 Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit beschreibt die Wahrscheinlichkeit eines Ereignisses, unter der Bedingung, dass ein anderes Ereignis bereits eingetreten ist. Die bedingte Wahrscheinlichkeit wird oft mit $P(A|B)$ bezeichnet, was gelesen wird als "die Wahrscheinlichkeit von A gegeben B". Diese Wahrscheinlichkeit wird berechnet mit der Formel:

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

Dabei ist $P(A \cap B)$ die Wahrscheinlichkeit, dass sowohl A als auch B eintreten, und $P(B)$ die Wahrscheinlichkeit, dass B eintritt.

6.1 Beispiele zur bedingten Wahrscheinlichkeit

6.1.1 Beispiel 1: Raucher und Geschlecht

Betrachten wir eine Gruppe von 20 Personen, die sich in Raucher und Nichtraucher sowie in Männer und Frauen aufteilen. Die folgende Tabelle zeigt die Anzahl der Personen in jeder Kategorie:

	Männer (M)	Frauen (F)	Total
Raucher (R)	3	1	4
Nichtraucher ($\neg R$)	9	7	16
Total	12	8	20

Um die Wahrscheinlichkeiten zu berechnen, teilen wir die Werte in der Tabelle durch die Gesamtzahl der Personen (20):

	Männer (M)	Frauen (F)	Total
Raucher (R)	0.15	0.05	0.2
Nichtraucher ($\neg R$)	0.45	0.35	0.8
Total	0.6	0.4	1

Die bedingte Wahrscheinlichkeit, dass eine zufällig ausgewählte Person ein Mann ist, gegeben dass sie Raucher ist, wird berechnet als:

$$P(M|R) = \frac{P(M \cap R)}{P(R)} = \frac{0.15}{0.2} = 0.75$$

Dies bedeutet, dass 75% der Raucher Männer sind.

6.1.2 Beispiel 2: Medizinischer Test

Ein medizinischer Test soll feststellen, ob eine Person eine bestimmte Krankheit hat. Die Tabelle zeigt die Wahrscheinlichkeiten für die Testergebnisse:

	Krankheit (D)	Keine Krankheit ($\neg D$)
Positiv (+)	0.009	0.099
Negativ (-)	0.001	0.891

Die bedingte Wahrscheinlichkeit, dass eine Person die Krankheit hat, gegeben dass der Test positiv ist, wird berechnet als:

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.08$$

Dies bedeutet, dass bei einem positiven Testergebnis die Wahrscheinlichkeit, dass die Person tatsächlich die Krankheit hat, nur 8% beträgt.

6.2 Das Bayes-Theorem

Das Bayes-Theorem liefert eine nützliche Methode zur Berechnung bedingter Wahrscheinlichkeiten. Es besagt, dass:

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

6.2.1 Beispiel: Spam-Filter

Ein Spam-Filter prüft E-Mails auf bestimmte Wörter wie "free"(gratis). Die Wahrscheinlichkeiten sind wie folgt:

- $P(\text{Spam}) = 0.7$
- $P(\text{Low Priority}) = 0.2$
- $P(\text{High Priority}) = 0.1$
- $P(\text{free}|\text{Spam}) = 0.9$
- $P(\text{free}|\text{Low Priority}) = 0.01$
- $P(\text{free}|\text{High Priority}) = 0.01$

Die bedingte Wahrscheinlichkeit, dass eine E-Mail Spam ist, gegeben dass das Wort "freeënthalten ist, wird berechnet als:

$$P(\text{Spam}|\text{free}) = \frac{P(\text{free}|\text{Spam}) \cdot P(\text{Spam})}{P(\text{free})}$$

$$= \frac{0.9 \cdot 0.7}{0.9 \cdot 0.7 + 0.01 \cdot 0.2 + 0.01 \cdot 0.1} = 0.995$$

6.3 Gesetz der totalen Wahrscheinlichkeit

Das Gesetz der totalen Wahrscheinlichkeit besagt, dass für eine Partitionierung $A_1, A_2, \ldots, A_k$ und ein Ereignis B gilt:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$$

$$+ \ldots + P(B|A_k)P(A_k)$$

6.3.1 Beispiel: Emails

Eine Email wird in die Kategorien Spam, niedrige Priorität und hohe Priorität eingeteilt. Die Wahrscheinlichkeiten sind wie folgt:

- $P(\text{Spam}) = 0.7$
- $P(\text{Low Prio}) = 0.2$
- $P(\text{High Prio}) = 0.1$
- $P(\text{free}|\text{Spam}) = 0.9$
- $P(\text{free}|\text{Low Prio}) = 0.01$
- $P(\text{free}|\text{High Prio}) = 0.01$

Die Wahrscheinlichkeit, dass das Wort “free” in einer Email erscheint, wird berechnet als:

$$P(\text{free}) = P(\text{free}|\text{Spam})P(\text{Spam})$$

$$+P(\text{free}|\text{Low Prio})P(\text{Low Prio})$$

$$+P(\text{free}|\text{High Prio})P(\text{High Prio})$$

$$= 0.9 \cdot 0.7 + 0.01 \cdot 0.2 + 0.01 \cdot 0.1$$

$$= 0.63 + 0.002 + 0.001 = 0.633$$

6.4 Zusammenfassung nützlicher R-Funktionen

Hier sind einige nützliche R-Funktionen zur Berechnung bedingter Wahrscheinlichkeiten und zur Visualisierung:

- `prop.table()` - Berechnet relative Häufigkeiten.
- `table()` - Erstellt eine Kreuztabelle der Häufigkeiten.
- `barplot()` - Erstellt ein Balkendiagramm.
- `mosaicplot()` - Erstellt ein Mosaikdiagramm für kategoriale Daten.
- `chisq.test()` - Führt einen Chi-Quadrat-Test zur Unabhängigkeit durch.

6.5 Aufgaben

6.5.1 A6.1 Satz von Bayes

In einem Land der Dritten Welt leiden 1% der Menschen an einer bestimmten Infektionskrankheit. Ein Test zeigt die Krankheit bei den tatsächlich Erkrankten zu 98% korrekt an. Leider zeigt der Test auch 3% der Gesunden als erkrankt an.

Wir bezeichnen mit K eine kranke Person und T eine positiv getestete Person.

a) Interpretieren (nicht berechnen!) Sie die W’keiten

$$P(K), P(\overline{T}), P(K \mid T), P(T \mid K), P(\overline{T} \mid \overline{K})$$

b) Bezeichnen Sie die W’keit, die in der Aufgabe gegeben sind, wie in a).

c) Berechnen Sie $P(\overline{K})$

d) Mit welcher W’keit zeigt der Test bei einer zufällig ausgewählten Person ein positives Ergebnis? Verwenden Sie das Gesetz der totale Wahrscheinlichkeit.

e) Mit welcher W’keit ist eine positiv getestete Person auch tatsächlich krank? Kommentieren Sie das Ergebnis und verwenden Sie das Bayes Theorem.

f) Mit welcher W'keit ist eine als negativ getestete Person gesund? Kommentieren Sie das Ergebnis und verwenden Sie das Bayes Theorem.

a)

$P(K)$: Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person krank ist.
$P(\overline{T})$: Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person ein negatives Testergebnis hat.
$P(K \mid \overline{T})$: Die Wahrscheinlichkeit, dass eine Person krank ist, gegeben, dass sie ein negatives Testergebnis hat.
$P(T \mid K)$: Die Wahrscheinlichkeit, dass eine Person ein positives Testergebnis hat, gegeben, dass sie krank ist.
$P(\overline{T} \mid \overline{K})$: Die Wahrscheinlichkeit, dass eine gesunde Person ein negatives Testergebnis hat.

b)

Gegebene Wahrscheinlichkeiten:

$$P(K) = 0.01$$

$$P(T \mid K) = 0.98$$

$$P(T \mid \overline{K}) = 0.03$$

c)

Berechnung von $P(\overline{K})$:

$$P(\overline{K}) = 1 - P(K) = 1 - 0.01 = 0.99$$

d)

Wahrscheinlichkeit für ein positives Testergebnis:

$$P(T) = P(T \mid K) \cdot P(K) + P(T \mid \overline{K}) \cdot P(\overline{K})$$

$$P(T) = 0.98 \cdot 0.01 + 0.03 \cdot 0.99$$
$$= 0.0098 + 0.0297 = 0.0395$$

e)

Wahrscheinlichkeit, dass eine positiv getestete Person krank ist (Bayes-Theorem):

$$P(K \mid T) = \frac{P(T \mid K) \cdot P(K)}{P(T)}$$

$$P(K \mid T) = \frac{0.98 \cdot 0.01}{0.0395} \approx 0.2481$$

Kommentar: Trotz des positiven Tests liegt die Wahrscheinlichkeit, dass die Person tatsächlich krank ist, nur bei etwa 24.81%. Dies liegt daran, dass die Krankheit sehr selten ist.

f)

Wahrscheinlichkeit, dass eine negativ getestete Person gesund ist (Bayes-Theorem):

$$P(\overline{K} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{K}) \cdot P(\overline{K})}{P(\overline{T})}$$

Dabei ist $P(\overline{T}) = 1 - P(T) = 1 - 0.0395 = 0.9605$

$$P(\overline{T} \mid \overline{K}) = 1 - P(T \mid \overline{K}) = 1 - 0.03 = 0.97$$

$$P(\overline{K} \mid \overline{T}) = \frac{0.97 \cdot 0.99}{0.9605} = 0.999792$$

Kommentar: Die Wahrscheinlichkeit, dass eine negativ getestete Person gesund ist, ist sehr hoch (etwa 99.95%), was zeigt, dass der Test sehr zuverlässig ist, wenn er ein negatives Ergebnis anzeigt.

7 Normalverteilung

7.1 Kontinuierliche Messdaten

In vielen Anwendungen haben wir es nicht mit diskreten Daten zu tun, sondern mit Messdaten. Messdaten können jeden Wert in einem bestimmten Bereich annehmen. Zum Beispiel können gemessene Körpergrössen (in cm) jeden Wert im Intervall $[0, 500]$ annehmen, wie etwa 145.325 cm oder 54 cm. Voraussetzung hierfür ist, dass eine beliebig genaue Messung möglich ist.

7.2 Definitionen

Der Wertebereich W_X einer Zufallsvariablen ist die Menge aller Werte, die X annehmen kann. Eine Zufallsvariable X ist stetig, wenn der Wertebereich W_X kontinuierlich ist. Eine kontinuierliche Menge ist ein zusammenhängender Ausschnitt aus der Zahlengeraden, im Gegensatz zu einer diskreten Menge wie $\{1, 2, 3\}$. Wichtige kontinuierliche Wertebereiche sind $W_X = \mathbb{R}$, $\mathbb{R}^+$ oder $[0, 1]$.

7.3 Intervalle

Intervalle können durch eckige und runde Klammern definiert werden, je nachdem, ob die Grenzen innerhalb oder ausserhalb des Intervalls sein sollen. Eine runde Klammer bedeutet, dass der Wert ausserhalb des Intervalls liegt, eine eckige Klammer bedeutet, dass der Wert innerhalb des Intervalls liegt. Beispielsweise enthält das Intervall $(1.2, 2.5]$ die Zahl 1.2 nicht, aber die Zahl 2.5.

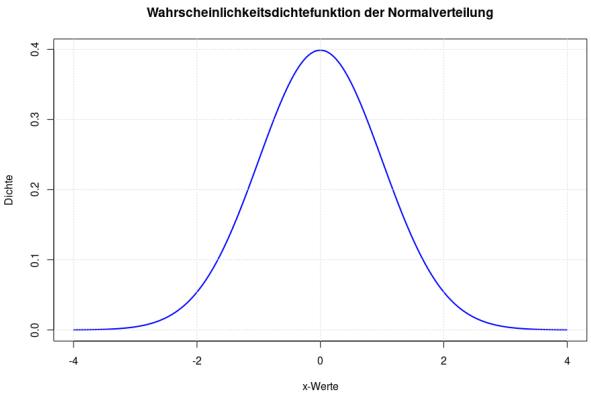
7.4 Punktwahrscheinlichkeit 0

Die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen hat "PunktWahrscheinlichkeiten $P(X = x)$ für alle möglichen x im Wertebereich. Bei einer stetigen Zufallsvariablen X gilt jedoch für alle $x \in W_X$, dass $P(X = x) = 0$. Das bedeutet, dass die Wahrscheinlichkeitsverteilung von X nicht mittels "PunktWahrscheinlichkeiten beschrieben werden kann.

7.5 Wahrscheinlichkeitsdichte

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

- $f(x) \geq 0$, das heisst, die Kurve liegt 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, was der Wahrscheinlichkeit entspricht, dass irgendein Wert gemessen wird.



Dies kann mit folgendem R Skript geplottet werden.

```
set.seed(42)
# Generate data for normal distribution
x <- seq(-4, 4, length = 1000)
y <- dnorm(x)

# Plot the PDF of the normal distribution
plot(x, y, type = "l", lwd = 2, col = "blue",
     main = "Wahrscheinlichkeitsdichtefunktion
           der Normalverteilung",
     xlab = "x-Werte", ylab = "Dichte")
```

7.6 Quantile

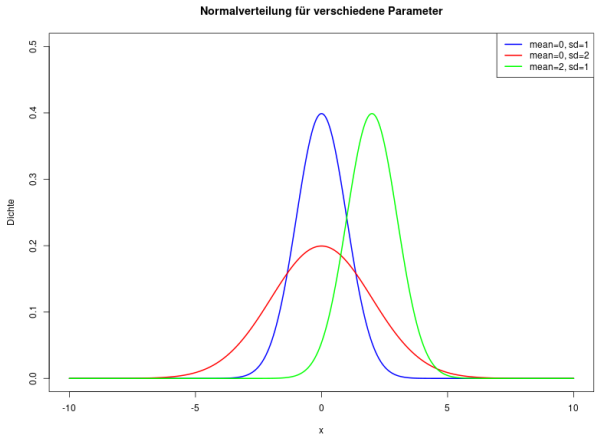
Bei stetigen Verteilungen ist das α-Quantil q_α derjenige Wert, bei dem die Fläche (Wahrscheinlichkeit) unter der Dichtefunktion von $-\infty$ bis q_α gerade α entspricht. Das 50%-Quantil ist der Median.

7.7 Normalverteilung

Die Normalverteilung (oder Gaussverteilung) einer Zufallsvariablen X wird als $X \sim N(\mu, \sigma^2)$ beschrieben. Ihre Dichtefunktion ist definiert als

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

Der Erwartungswert ist μ und die Varianz ist σ^2.




```r
x <- seq(-10, 10, length = 1000)
y1 <- dnorm(x, mean = 0, sd = 1)
y2 <- dnorm(x, mean = 0, sd = 2)
y3 <- dnorm(x, mean = 2, sd = 1)

plot(x, y1, type = "l", lwd = 2, col = "blue
     ", ylim = c(0, 0.5),
     main = "Normalverteilung fuer
          verschiedene Parameter",
     xlab = "x", ylab = "Dichte")
lines(x, y2, col = "red", lwd = 2)
lines(x, y3, col = "green", lwd = 2)
```

Eigenschaften der Normalverteilung

Die Dichtefunktionen der Normalverteilung sind "glockenförmig". Durch den Parameter μ wird die Kurve nach rechts (positiv) oder links (negativ) verschoben. Der Parameter σ beeinflusst die Breite und Höhe der Kurve: Ist σ klein, so ist die Kurve schmal und hoch; ist σ gross, so ist die Kurve breit und flach.

7.8 Beispiel: Verteilung von IQ

IQ-Tests folgen einer Normalverteilung mit einem Mittelwert von 100 und einer Standardabweichung von 15. Die Zufallsvariable X, die den IQ einer zufällig ausgewählten Person misst, ist normalverteilt mit $\mu = 100$ und $\sigma = 15$, also $X \sim N(100, 15^2)$.

Berechnung der Wahrscheinlichkeit für hohen IQ

Um die Wahrscheinlichkeit zu berechnen, dass jemand einen IQ von mehr als 130 hat, suchen wir $P(X \geq 130)$. Da die Funktion `pnorm` die Wahrscheinlichkeit für $P(X \geq 130)$ berechnet, verwenden wir $1 - pnorm(130, 100, 15)$.

```r
mean <- 100
sd <- 15

# Wahrscheinlichkeit fuer IQ > 130
prob_high_iq <- 1 - pnorm(130, mean, sd)

print(prob_high_iq)
```

Das Ergebnis ist etwa 0.02275. Das bedeutet, dass etwa 2% der Bevölkerung hochbegabt sind.

Berechnung eines Intervalls

Um das Intervall zu bestimmen, das 95% der IQ-Werte um den Mittelwert $\mu = 100$ enthält, verwenden wir die `qnorm`-Funktion in R. Wir berechnen die Quantile $q_{0.025}$ und $q_{0.975}$ wobei der Mittelwert 100 und die Standardabweichung 15 beträgt.

```r
mean <- 100
sd <- 15

# Quantile berechnen
lower_quantile <- qnorm(0.025, mean, sd)
upper_quantile <- qnorm(0.975, mean, sd)

print(lower_quantile)
print(upper_quantile)
```

Das Ergebnis ist etwa 70.6 für das 2.5%-Quantil und 129.4 für das 97.5%-Quantil. Das bedeutet dass 95% der Menschen einen IQ zwischen etwa 70 und 130 haben.

Berechnung der Wahrscheinlichkeit innerhalb einer Standardabweichung

Um zu bestimmen, wie viel Prozent der Bevölkerung innerhalb einer Standardabweichung vom Mittelwert liegen, verwenden wir die 'pnorm'. Wir berechnen die W'keit $P(85 \leq X \leq 115)$, wobei der Mittelwert 100 und die Standardabweichung 15 beträgt.

```
mean <- 100
sd <- 15
lower_bound <- 85
upper_bound <- 115

probability <- pnorm(upper_bound, mean, sd) -
      pnorm(lower_bound, mean, sd)
print(probability)
```

Das Ergebnis ist etwa 0.6827, was bedeutet, dass etwa 68% der Bevölkerung einen IQ zwischen 85 und 115 haben.

7.9 Zusammenfassung nützlicher R-Funktionen

- `pnorm(q, mean, sd)`: Berechnet die kumulative Verteilungsfunktion der Normalverteilung.
- `qnorm(p, mean, sd)`: Berechnet die Quantile der Normalverteilung.

7.10 Aufgaben

7.10.1 A7.2 W'keit, Intervalle und Standardabweichung

In einem Ort gibt es einige Karpfenteiche. Die Masse der Karpfen ist normalverteilt mit dem Erwartungswert $\mu = 4$kg und der Standardabweichung 1.25kg.

a) Wie gross ist die Wahrscheinlichkeit, einen Karpfen zu fangen, der höchstens 2.5kg bzw. mindestens 5kg wiegt?

b) Wie viel Prozent aller Karpfen wiegen zwischen 3kg und 4.5kg?

c) Der Fischereiverband will einen Preis für die schwersten Karpfen aussetzen. Welches Mindestgewicht muss man verlangen, damit die Wahrscheinlichkeit, den Preis zu bekommen, 2% beträgt?

a) Wahrscheinlichkeit für bestimmte Gewichte

Die Wahrscheinlichkeit, einen Karpfen zu fangen, der höchstens 2.5kg wiegt ist

$$P(X \leq 2.5) = pnorm(2.5, 4, 1.25)$$

Die Wahrscheinlichkeit, einen Karpfen zu fangen, der mindestens 5kg wiegt ist

$$P(X \geq 5) = pnorm(5, 4, 1.25)$$

```
mean <- 4
sd <- 1.25
# Wahrscheinlichkeit fuer Karpfen <= 2.5 kg
prob_less_2_5 <- pnorm(2.5, mean, sd)
# Wahrscheinlichkeit fuer Karpfen >= 5 kg
prob_greater_5 <- 1 - pnorm(5, mean, sd)

print(prob_less_2_5)
print(prob_greater_5)
```

Dies ergibt einen Wert für $P(X \leq 2.5) = 0.115$. Also ist die W'keit einen Karpfen mit höchstens 2.5kg zu fangen 11.5%. Die Wahrscheinlichkeit einen Karpfen mit mindestens 5kg zu fangen ist $P(X \geq 5) = 0.212$, also 21.2%.

b) Karpfen zwischen 3kg und 4.5kg

$$P(3 \leq X \leq 4.5) =$$
$$pnorm(4.5, 4.1.25) - pnorm(3, 4, 1.25)$$

```
# Wahrscheinlichkeit Karpfen zwischen 3 kg
    und 4.5 kg
prob_between_3_and_4_5 <- pnorm(4.5, mean, sd
    ) - pnorm(3, mean, sd)
print(prob_between_3_and_4_5)
```

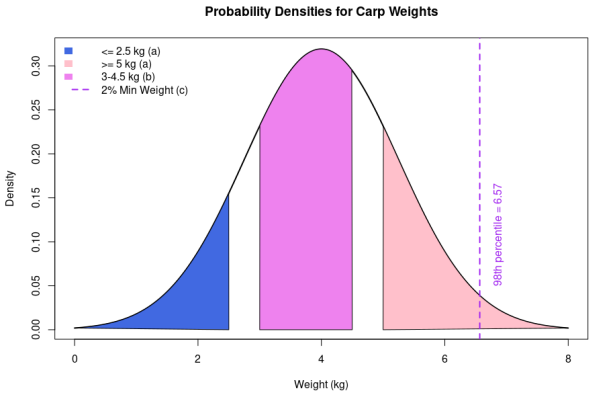
Dies ergibt $P(3 \leq X \leq 4.5) = 0.4436$. Also wiegen 44.36% der Karpfen zwischen 3kg und 4.5kg.

c) Mindestgewicht für Preis mit 2% Wahrscheinlichkeit

$$P(X \geq x) = 0.02 \Longrightarrow qnorm(0.98, 4, 1.25)$$

```
min_weight_for_prize <- qnorm(0.98, mean, sd)
print(min_weight_for_prize)
```

Der berechnete Wert ist 6.567186, also muss der Karpfen 6.57kg oder mehr wiegen, um den Preis zu gewinnen.



Hier sehen wir die Aufgabe nochmals visualisiert einem Plot mir R und das dazu genutzte Skript:

```
# Set parameters
mean <- 4
sd <- 1.25

# Create a sequence of weights for the x-axis
weights <- seq(0, 8, length.out = 1000)

# Calculate the normal density
density <- dnorm(weights, mean, sd)

# Calculate the probabilities
prob_less_2_5 <- pnorm(2.5, mean, sd)
prob_greater_5 <- 1 - pnorm(5, mean, sd)
prob_between_3_and_4_5 <- pnorm(4.5, mean, sd) - pnorm(3, mean, sd)
min_weight_for_prize <- qnorm(0.98, mean, sd)

# Create the plot
plot(weights, density, type = "l", lwd = 2, col = "black",
     main = "Probability Densities for Carp Weights",
     xlab = "Weight (kg)", ylab = "Density")

# Shade the area for weight <= 2.5 kg
polygon(c(weights[weights <= 2.5], 2.5), c(density[weights <= 2.5], 0), col = "royalblue")

# Shade the area for weight >= 5 kg
polygon(c(5, weights[weights >= 5]), c(0, density[weights >= 5]), col = "pink")

# Shade the area for weight between 3 kg and 4.5 kg
polygon(c(weights[weights >= 3 & weights <= 4.5], 4.5, 3), c(density[weights >= 3 & weights <= 4.5], 0, 0), col = "violet")

# Add a vertical line for the minimum weight for prize
abline(v = min_weight_for_prize, col = "purple", lwd = 2, lty = 2)

# Add text for the minimum weight for prize
text(min_weight_for_prize + 0.2, 0.05, paste("98th percentile =", round(min_weight_for_prize, 2)), col = "purple", pos = 4, srt = 90)

# Add legend
legend("topleft", legend = c("<= 2.5 kg (a)", ">= 5 kg (a)", "3-4.5 kg (b)", "2% Min Weight (c)"),
       fill = c("royalblue", "pink", "violet", NA),
       border = NA,
       lty = c(NA, NA, NA, 2), lwd = c(NA, NA, NA, 2),
       col = c(NA, NA, NA, "purple"),
       bty = "n")
```

8 Gesetz der grossen Zahlen, zentraler Grenzwertsatz

8.1 Gesetz der grossen Zahlen

Das Gesetz der grossen Zahlen besagt, dass der Durchschnitt einer grossen Anzahl von Zufallsvariablen (i.i.d., identisch und unabhängig verteilt) gegen den Erwartungswert der Einzelvariablen konvergiert. Dies bedeutet, dass bei einer grossen Anzahl von Wiederholungen eines Zufallsexperiments das arithmetische Mittel der Ergebnisse näher am Erwartungswert liegt.

Beispiel: Münzwurf

Angenommen, wir werfen eine faire Münze 1000 Mal. Der Erwartungswert für "Kopfist 0.5. Nach dem Gesetz der grossen Zahlen sollte der Anteil der "KopfErgebnisse nahe bei 0.5 liegen.

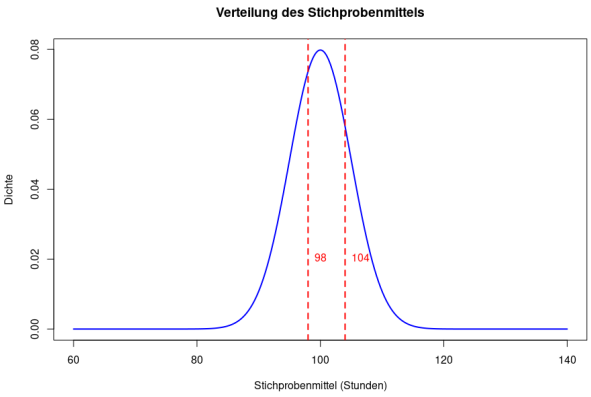
```
# Simulation des Muenzwurfs
set.seed(123)
ergebnisse <- rbinom(1000, 1, 0.5)
durchschnitt <- mean(ergebnisse)
print(durchschnitt) # Sollte nahe bei 0.5
    liegen
```

8.2 Zentraler Grenzwertsatz

Der zentrale Grenzwertsatz besagt, dass die Verteilung der Summe (oder des Durchschnitts) einer grossen Anzahl von i.i.d. Zufallsvariablen (unabhängig von der ursprünglichen Verteilung) näherungsweise normalverteilt ist, wenn die Anzahl der Variablen gross genug ist.

Beispiel: Lebensdauer eines elektrischen Teils

Angenommen, die Lebensdauer eines elektrischen Teils ist durchschnittlich 100 Stunden mit einer Standardabweichung von 20 Stunden. Wir testen 16 solcher Teile und berechnen das Stichprobenmittel.



Dieser Plot zeigt die Normalverteilung des Stichprobenmittels, wobei die vertikalen roten Linien die Grenzen von 98 und 104 Stunden markieren. Dies veranschaulicht die Wahrscheinlichkeiten, die im folgenden Beispiel berechnet wurden.

```
# Parameter
mean <- 100
sd <- 20
n <- 16

# Verteilung des Stichprobenmittels
mean_stichprobe <- mean
sd_stichprobe <- sd / sqrt(n)

# Wahrscheinlichkeit, dass das
    Stichprobenmittel unter 104 Stunden
    liegt
prob_unter_104 <- pnorm(104, mean =
    mean_stichprobe, sd = sd_stichprobe)
print(prob_unter_104)

# Wahrscheinlichkeit, dass das
    Stichprobenmittel zwischen 98 und 104
    Stunden liegt
prob_zwischen_98_104 <- pnorm(104, mean =
    mean_stichprobe, sd = sd_stichprobe) -
    pnorm(98, mean = mean_stichprobe, sd =
    sd_stichprobe)
print(prob_zwischen_98_104)
```

Das obige R-Skript berechnet die Verteilung des Stichprobenmittels der Lebensdauer eines elektrischen Teils.

Berechnung der Verteilung des Stichprobenmittels:
Der Mittelwert des Stichprobenmittels ($\mu_{\bar{X}}$) ist gleich dem Mittelwert der Population (μ). Die Standardabweichung des Stichprobenmittels ($\sigma_{\bar{X}}$) wird berechnet als $\sigma/\sqrt{n}$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{16}} = 5$$

Die Standardabweichung des Stichprobenmittels beträgt 5, was zeigt, dass die durchschnittlichen Stichprobenwerte im Durchschnitt um 5 Einheiten vom Mittelwert abweichen.

Berechnung der Wahrscheinlichkeiten:
Die Wahrscheinlichkeit, dass das Stichprobenmittel unter 104 Stunden liegt, wird mit der Funktion 'pnorm' berechnet. Die Wahrscheinlichkeit, dass das Stichprobenmittel zwischen 98 und 104 Stunden liegt, wird ebenfalls mit der Funktion 'pnorm' berechnet, indem die Wahrscheinlichkeiten für die Ober- und Untergrenze subtrahiert werden.

$$P(\bar{X} < 104) = P\left(Z < \tfrac{104-100}{\sigma_{\bar{X}}}\right) = P(Z < 0.8)$$

$$P(98 \leq \bar{X} \leq 104) = P\left(Z < \tfrac{104-100}{\sigma_{\bar{X}}}\right) - P\left(Z < \tfrac{98-100}{\sigma_{\bar{X}}}\right)$$

Diese Berechnungen zeigen, wie der zentrale Grenzwertsatz verwendet werden kann, um Wahrscheinlichkeiten für das Stichprobenmittel zu bestimmen. Das Verständnis dieser Konzepte ist entscheidend für statistische Analysen und Tests.

Analyse Resultat
Das Ergebnis der Berechnung ist folgendes.

```
sd_stichprobe num 5
prob_unter_104 num 0.7881446
prob_zwischen_98_104 num 0.4435663
```

Die Berechnung der Verteilung des Stichprobenmittels zeigt, dass die Standardabweichung des Stichprobenmittels 5 beträgt. Dies bedeutet, dass die durchschnittlichen Stichprobenwerte im Durchschnitt um 5 Einheiten vom Mittelwert abweichen.

Die Wahrscheinlichkeit, dass das Stichprobenmittel unter 104 Stunden liegt, beträgt etwa 78.8%. Dies zeigt, dass es sehr wahrscheinlich ist, dass der durchschnittliche Wert der Lebensdauer unter 104 Stunden liegt.

Die Wahrscheinlichkeit, dass das Stichprobenmittel zwischen 98 und 104 Stunden liegt, beträgt etwa 44.4%. Dies bedeutet, dass fast die Hälfte der getesteten Teile eine Lebensdauer innerhalb dieses Intervalls aufweisen wird.

8.2.1 Simulation des Zentralen Grenzwertsatzes

Wir werden den zentralen Grenzwertsatz mit einer Simulation verdeutlichen, bei der wir mehrfach Werte aus einer nicht-normalverteilten Zufallsvariable ziehen und das arithmetische Mittel dieser Werte berechnen.

Beispiel: Wir ziehen 1000-mal 10 Werte aus der Menge {0, 10, 11} und berechnen jeweils das arithmetische Mittel.

```
1  # Werte und Anzahl der Ziehungen
2  werte <- c(0, 10, 11)
3  n_ziehungen <- 10
4  n_wiederholungen <- 1000
5
6  # Simulation
7  set.seed(123)
8  mittelwerte <- replicate(n_wiederholungen,
       mean(sample(werte, n_ziehungen, replace
       = TRUE)))
9
10 # Histogramm der Mittelwerte
11 hist(mittelwerte, col = "darkseagreen3", main
        = "Histogramm der Mittelwerte", xlab =
       "Mittelwert", ylab = "Haeufigkeit",
       breaks = 20)
```

Diese Simulation zeigt, dass die Verteilung der Mittelwerte näherungsweise normalverteilt ist, obwohl die ursprünglichen Werte {0, 10, 11} nicht normalverteilt sind.

8.3 Aufgaben

8.3.1 A8.2 W'keit für Probe ausserhalb Standardabweichung

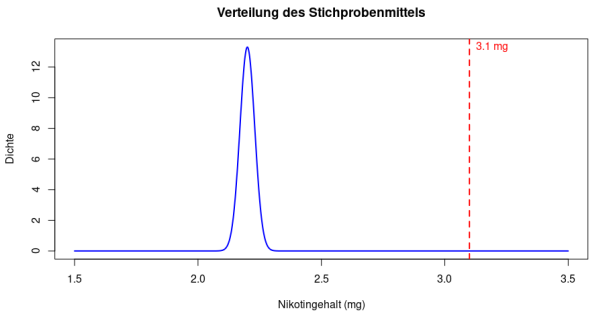
Ein Zigarettenhersteller gibt an, dass der Nikotingehalt in einer Zigarette durchschnittlich 2.2 mg mit einer Standardabweichung von 0.3 mg hat. Bei einer Stichprobe von 100 zufällig ausgewählten Zi-

garetten ist das Stichprobenmittel allerdings 3.1 mg.

Wie hoch ist die Wahrscheinlichkeit, dass das Stichprobenmittel einen Wert von 3.1 mg oder mehr erreicht, wenn die Aussage des Zigarettenherstellers wahr ist?

```
mean <- 2.2
sd <- 0.3
n <- 100
sample_mean <- 3.1

# Standardabweichung des Stichprobenmittels
sd_stichprobe <- sd / sqrt(n)

# Wahrscheinlichkeit, dass das
    Stichprobenmittel >= 3.1 ist
prob_ge_3_1 <- 1 - pnorm(sample_mean, mean =
    mean, sd = sd_stichprobe)
print(prob_ge_3_1)
```

Die W'keit, dass das Strichprobenmittel einen Wert von 3.1mg oder mehr erreicht, ist extrem gering ($P(\overline{X}_{100} \geq 3.1 \approx 0)$). Dies deutet darauf hin, dass die Stichprobe möglicherweise nicht repräsentativ ist oder die Aussage des Herstellers überprüft werden sollte. Wir sehen dies auch folgend visualisiert.



8.3.2 A8.3 Vergleich Verteilungen mit unterschiedlichem Stichprobenraum

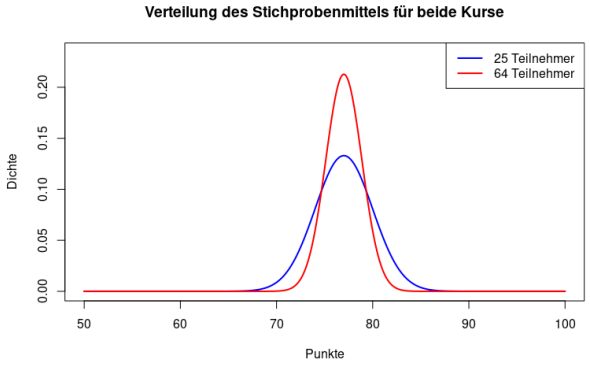
Ein Dozent weiss aus Erfahrung, dass bei einer Prüfung die Punktezahlen durchschnittlich bei 77 Punkte mit einer Standardabweichung von 15 Punkten liegen. In diesem Semester unterrichtet der Dozent zwei Kurse; der eine hat 25, der andere 64 Teilnehemer.

a) Wie gross ist die W'keit, dass das durchschnittliche Prüfungsergebnis im Kurs mit 25 Teilnehemer zwischen 72 und 82 Punkten liegt?

b) Wiederholen Sie die Rechnung aus Teil a) für den Kurs mit 64 Teilnerhmer.

```
mean <- 77
sd <- 15

n1 <- 25 # Kurs mit 25 TeilnehmerInnen
sd_stichprobe1 <- sd / sqrt(n1)

# W'keit, dass Stichprobenmittel zwischen 72
    und 82
prob_zwischen_72_82_25 <- pnorm(82, mean =
    mean, sd = sd_stichprobe1) - pnorm(72,
    mean = mean, sd_stichprobe1)
print(prob_zwischen_72_82_25)

n2 <- 64 # Kurs mit 64 TeilnehmerInnen
sd_stichprobe2 <- sd / sqrt(n2)

# W'keit, dass Stichpro. Mittel zwischen 72
    und 82
prob_zwischen_72_82_64 <- pnorm(82, mean =
    mean, sd = sd_stichprobe2) - pnorm(72,
    mean = mean, sd = sd_stichprobe2)
print(prob_zwischen_72_82_64)
```

Die W'keit, dass das durchschnittliche Prüfungsergebnis im Kurs mit 25 TeilnehmerInnen zwischen 72 und 82 Punkten liegt, beträgt etwa 0.682. Für den Kurs mit 64 TeilnehmerInnen beträgt diese W'keit etwa 0.841. Der Vergleich zeigt, dass die W'keit bei grösseren Stichproben enger um den Mittelwert liegt.



Der Plot zeigt die Verteilungen der Stichprobenmittelwerte: 64 Teilnehmer (rote Linie) haben eine engere Verteilung als 25 Teilnehmer (blaue Linie), was auf eine geringere Standardabweichung und höhere Präzision hinweist

9 Hypothesentest, z-Test, t-Test

9.1 Hypothesentest

Hypothesentests sind wichtige statistische Werkzeuge, um zu entscheiden, ob eine Messreihe zu einer bestimmten Grösse passt. Ein Beispiel ist die Prüfung, ob eine neue Abfüllmaschine, die angeblich Dosen mit 500 ml füllt, tatsächlich diesen Durchschnittswert einhält.

9.1.1 Beispiel: Abfüllmaschine

Angenommen, wir möchten prüfen, ob eine Abfüllmaschine Dosen mit durchschnittlich 500 ml füllt. Wir nehmen eine Stichprobe von 30 Dosen und messen die Füllmengen. Der angenommene Standardfehler beträgt 10 ml.

```
# Parameter
mu <- 500
sigma <- 10
n <- 30

# Stichprobendaten
set.seed(123)
stichprobe <- rnorm(n, mean = mu, sd = sigma)
mean_stichprobe <- mean(stichprobe)
sd_stichprobe <- sd(stichprobe)

# z-Test
z_wert <- (mean_stichprobe - mu) / (sigma /
    sqrt(n))
p_wert <- 2 * pnorm(-abs(z_wert))
```

Der berechnete p-Wert beträgt 0.79640, was bedeutet, dass die Wahrscheinlichkeit, dass das Stichprobenmittel von 30 Dosen zufällig 499.53ml oder weiter von 500ml abweicht, ungefähr 79.64% ist. Da der p-Wert grösser als 0.05 ist, können wir die Nullhypothese nicht ablehnen. Das bedeutet, es gibt keinen ausreichenden Beweis dafür, dass die Abfüllmaschine falsch eingestellt ist.

9.2 z-Test

Der z-Test wird verwendet, wenn die Standardabweichung der Grundgesamtheit bekannt ist und die Stichprobengrösse gross (n $>$ 30) ist. Der z-Test prüft, ob der Mittelwert der Stichprobe signifikant vom angenommenen Mittelwert abweicht.

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Der p-Wert gibt die Wahrscheinlichkeit an, dass das beobachtete Ergebnis (oder ein extremeres) unter der Nullhypothese auftritt.

9.3 t-Test

Der t-Test wird verwendet, wenn die Standardabweichung der Grundgesamtheit unbekannt ist und die Stichprobengrösse klein ($n \leq 30$) ist. Der t-Test prüft ebenfalls, ob der Mittelwert der Stichprobe signifikant vom angenommenen Mittelwert abweicht.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

9.3.1 Beispiel: Durchschnittsgewicht

Angenommen, wir möchten das Durchschnittsgewicht von Äpfeln prüfen. Wir nehmen eine Stichprobe von 15 Äpfeln und messen deren Gewichte. Der angenommene Mittelwert ist 150g pro 15 Äpfel.

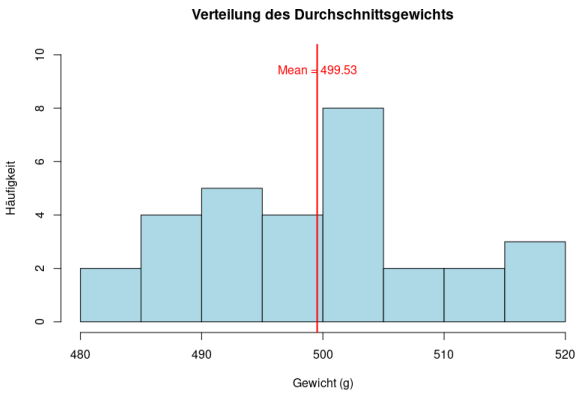
```
# Parameter
mu <- 150
n <- 15

# Stichprobendaten
set.seed(123)
stichprobe <- rnorm(n, mean = mu, sd = 15)
mean_stichprobe <- mean(stichprobe)
sd_stichprobe <- sd(stichprobe)

# t-Test
t_wert <- (mean_stichprobe - mu) / (
    sd_stichprobe / sqrt(n))
p_wert <- 2 * pt(-abs(t_wert), df = n - 1)
```

Der berechnete p-Wert beträgt 0.49651, was bedeutet, dass die Wahrscheinlichkeit, dass das Stichprobenmittel von 15 Äpfeln zufällig 152.29g oder weiter von 150g abweicht ungefähr 49.65% beträgt. Da der p-Wert grösser als 0.05 ist, können wir die Nullhypothese nicht ablehnen. Das bedeutet, es gibt keinen ausreichenden Beweis dafür,

dass das Durchschnittsgewicht der Äpfel signifikant von 150g abweicht.

Visualisierung

Um die Verteilung der Stichprobenergebnisse zu visualisieren, können wir ein Histogramm oder eine Dichtefunktion erstellen. Hier ein Beispiel für die Visualisierung der Verteilung des Durchschnittsgewichts der Äpfel.



9.4 Nullhypothese und Signifikanzniveau

Die Nullhypothese H_0 ist die Annahme, dass es keinen Unterschied oder Effekt gibt. Sie dienst als Ausgangspunkt für Hypothesentests. Zum Beispiel:

- Abfüllmaschine: H_0, die Maschine füllt Dosen mit durchschnittlich 500ml.
- Durchschnittsgewicht: H_0, das Durchschnittsgewicht der Äpfel beträgt 150g.

Das Signifikanzniveau α ist die Schwelle, bei der wir die Nullhypothese ablehnen. Typische Werte für α sind $0.05, 0.01$ und 0.10. Ein $\alpha = 0.05$ bedeutet, dass wir bereit sind, in 5% der Fälle einen Fehler 1. Art zu akzeptieren.

9.4.1 Interpretation des p-Werts

Der p-Wert gibt die W'keit an, unter der Annahme, dass H_0 wahr ist, ein Ergebnis zu erhalten, das genauso extrem oder extremer ist als das beobachtete Ergebnis.

- Wenn $p \leq \alpha$, lehnen wir H_0 ab (statistisch signifikant).
- Wenn $p > \alpha$, können wir H_0 nicht ablehnen (nicht signifikant).

9.4.2 Interpretation von p-Wert und α

Wenn $p \leq 0.05$
Angenommen ein Bauer möchte wissen, ob seine Äpfel im Durchschnitt schwerer sind als 150g. Da $p \leq \alpha$ können wir ihm sagen: Es ist sehr wahrscheinlich, dass die Äpfel tatsächlich schwerer sind als 150g.

Wenn $p > 0.05$
Ein Hersteller möchte überprüfen ob seine Abfüllmaschine Dosen mit durchschnittlich 500ml füllt. Da $p > 0.05$ können wir sagen: Es gibt keine ausreichenden Beweise dafür, dass die Maschine die Dosen im Durchschnitt mit 500ml füllt.

9.5 Aufgaben

9.5.1 A9.3 Hypothesentest

Eine Bäckerei gibt an, dass die von ihr hergestellten Brötchen ein Mindestgewicht von 50 g bei bekannter Standardabweichung $\sigma = 3$ g haben. Die Gewichte sind normalverteilt.

Ein Statistikstudent, der misstrauisch ist und vermutet, dass die Brötchen ein zu geringes Gewicht haben, kauft in der Bäckerei $n = 16$ Brötchen und wiegt alle Brötchen. Er erhält folgende Werte (in g):

```
46, 48, 52, 49, 46, 51, 52, 47, 49,
    44, 48, 51, 49, 50, 53, 47
```

a) Null- und Alternativhypothese und Hypothesentest

Nullhypothese (H_0): $\mu \geq 50$ g

Alternativhypothese (H_A): $\mu < 50$ g

Wir führen einen z-Test auf dem 5%-Signifikanzniveau durch.

```r
data <- c(46, 48, 52, 49, 46, 51, 52, 47, 49,
      44, 48, 51, 49, 50, 53, 47)

mu_0 <- 50
sigma <- 3
n <- length(data)
mean_sample <- mean(data)
sd_sample <- sd(data)

# z-Test
z_wert <- (mean_sample - mu_0) / (sigma /
     sqrt(n))
p_wert <- pnorm(z_wert)

print(p_wert)
```

Ergebnis: Der p-Wert ist knapp grösser als 0.05 (0.067), daher lehnen wir die Nullhypothese nicht ab. Es gibt keine Hinweise darauf, dass das durchschnittliche Gewicht der Brötchen weniger als 50 g beträgt.

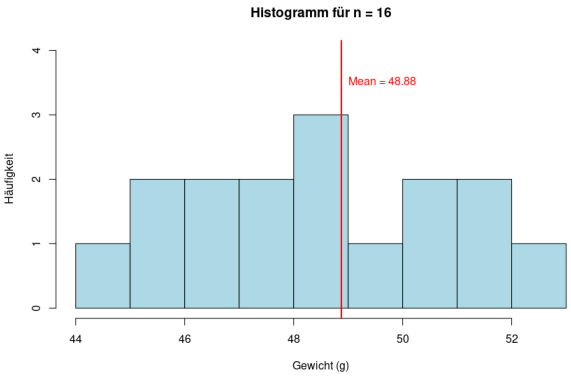
b) Grössere Stichprobe

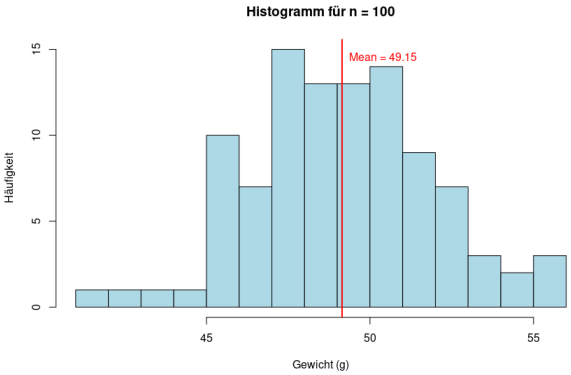
Dem Studenten kommen bei seiner Auswertung Bedenken wegen des kleinen Stichprobenumfangs von $n = 16$. Er untersucht deshalb noch einmal das Brötchengewicht, diesmal für $n = 100$ Brötchen. Er erhält denselben Mittelwert in der Stichprobe wie bei den $n = 16$ Brötchen.

Wir führen denselben Test mit $n = 100$ durch.

```r
n_large <- 100

z_wert_large <- (mean_sample - mu_0) / (sigma
      / sqrt(n_large))
p_wert_large <- pnorm(z_wert_large)

print(p_wert_large)
```

Ergebnis: Der p-Wert von 8.841729×10^{-5} ist deutlich kleiner als das Signifikanzniveau von 0.05, wodurch die Nullhypothese H_0 abgelehnt wird. Dies bedeutet, dass der Durchschnitt der gemessenen Brötchengewichte signifikant kleiner als 50 g ist.





Bei $n = 100$ sind die Daten zufällig generiert, daher kann der Mittelwert leicht abweichen. Der Plot dient rein dazu, den Unterschied zu visualisieren.

c) Verlassen auf gegebene Daten

Der Student ist nun auch misstrauisch gegenüber der bekannten Standardabweichung und möchte sich nur auf die gegebenen Daten verlassen. Wie geht er vor? Wir führen den Hypothesentest mit den gegebenen Daten durch.

```r
t_wert <- (mean_sample - mu_0) / (sd_sample /
     sqrt(n))
p_wert_t <- pt(t_wert, df = n - 1)

print(p_wert_t)
```

Ergebnis: Der p-Wert aus dem t-Test ist kleiner als 0.05 (0.047), daher wird die Nullhypothese verworfen. In a) wurde sie nicht verworfen, es kann also durchaus einen Unterschied machen, welchen Test wir anwenden.

10 Vertrauensintervall, Zwei-Stichprobentest und Wilcoxon-Test

10.1 Vertrauensintervall

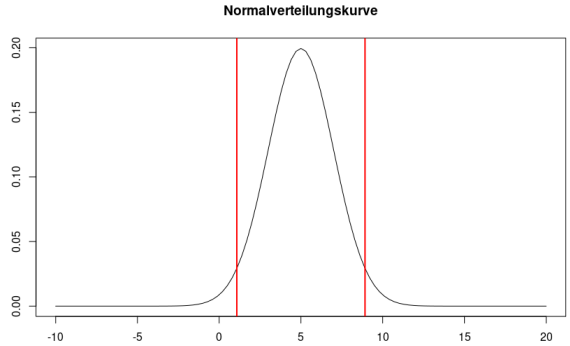
Bei der Punktschätzung für den Mittelwert μ einer Messreihe erhalten wir nur einen einzigen Schätzwert. Allerdings wissen wir nicht, wie nahe dieser geschätzte Mittelwert beim wahren Mittelwert der Verteilung der Messreihe liegt. Ein Vertrauensintervall gibt an, in welchem Bereich der wahre Mittelwert mit einer bestimmten vorgegebenen Wahrscheinlichkeit liegt.

Beispiel: Angenommen, wir haben eine Normalverteilung $X \sim N(5, 2^2)$ und möchten das Vertrauensintervall für den Mittelwert bestimmen.

```
# Bestimmen der Quartile
qnorm(p = c(0.025, 0.975), mean = 5, sd = 2)
```

```
[1] 1.080072 8.919928
```

Die Quantile $q_{0.025}$ und $q_{0.975}$ bestimmen den Verwerfungsbereich für einen zweiseitigen Test mit $\alpha = 0.05$. Wenn ein beobachteter Wert x_n im Verwerfungsbereich liegt, wird die Nullhypothese H_0 verworfen.



Dieser Plot zeigt die Normalverteilung mit dem Verwerfungsbereich (rote Linien).

- Wenn der beobachtete Wert innerhalb der roten Linien liegt (also nicht im Verwerfungsbereich), wird die Nullhypothese H_0 nicht verworfen.
- Wenn der beobachtete Wert ausserhalb der roten Linien liegt (also im Verwerfungsbereich), wird die Nullhypothese H_0 nicht verworfen.

10.2 Zweistichprobentest

Ein Zweistichprobentest vergleicht die Mittelwerte zweier unabhängiger Stichproben. Er kann entweder gepaart oder ungepaart sein.

10.2.1 Gepaarte Stichproben

Gepaarte Stichproben liegen vor, wenn jede Beobachtung in einer Gruppe eindeutig einer Beobachtung in der anderen Gruppe zugeordnet werden kann.

Beispiel: Untersuchung des Unterschieds im Augeninnendruck vor und nach einer Behandlung.

```
# Daten vor und nach der Behandlung
vorher <- c(25, 25, 27, 44, 30, 67, 53, 53,
     52, 60, 28)
nachher <- c(27, 29, 37, 56, 46, 82, 57, 80,
     61, 59, 43)

# t-Test fuer gepaarte Stichproben
t.test(nachher, vorher, paired = TRUE)
```

```
t = 4.2716, df = 10, p-value = 0.001633
```

- **t-Statistik:** beschreibt die Differenz zwischen den Mittelwerten der zwei Gruppen zur Standardabweichung dieser Differenz. Ein hoher Absolutwert deutet darauf hin, dass die Differenz zwischen den Mittelwerten "vorher" und "nachher" Daten sehr gross ist im Vergleich zur Streuung der Daten.

- **p-Wert**: Der p-Wert gibt die W'keit an, dass die beobachtete Differenz der Mittelwerte oder eine noch extremere Differenz auftritt, wenn die Nullhypothese wahr ist. Ein p-Wert von 0.001633 ist sehr klein, was darauf hinweist, dass die W'keit, dass die Differenz zufällig ist, extrem gering ist. Daher können wir die Nullhypothese ablehnen und sagen, dass die Mittelwerte signifikant unterschiedlich sind.

Durch den niedrigen p-Wert können wir also sagen, dass es unwahrscheinlich ist, dass der beobachtete Unterschied zwischen den Messungen rein zufällig sind. Mit anderen Worte: Wir können mit hoher Sicherheit sagen, dass die Veränderungen auf die Behandlung zurückzuführen sind.

10.2.2 Ungepaarte Stichproben

Ungepaarte Stichproben sind unabhängig voneinander und haben keine Zuordnung zwischen den Gruppen.

Beispiel: Vergleich der Mittelwerte von zwei Messgeräten A und B.

```
# Daten von Messgeraet A und B
A <- c(79.98, 80.04, 80.02, 80.04, 80.03,
    80.03, 80.04, 79.97, 80.05, 80.03,
    80.02, 80, 80.02)
B <- c(80.02, 79.94, 79.98, 79.97, 80.03,
    79.95, 79.97)

# t-Test fuer ungepaarte Stichproben
t.test(A, B, paired = FALSE)
```

```
t = 2.8399, df = 9.3725, p-value = 0.01866
```

Der p-Wert von 0.01866 zeigt einen signifikanten Unterschied zwischen den beiden Messgeräten. Dies bedeutet, dass die W'keit dass der beobachtete Unterschied zwischen den Messwerten der beiden Geräte rein zufällig ist. Mit anderen Worten: WIr können mit hoher Sicherheit sagen, dass die Unterschiede in den Messwerte tatsächlich auf einen echten Unterschied zwischen den Messgeräten zurückzuführen sind.

10.3 Wilcoxon-Test

Der Wilcoxon-Test ist eine nicht-parametrische Alternative zum t-Test und wird verwendet, wenn die Daten nicht normalverteilt sind. Er prüft, ob die Medianwerte zweier Stichproben unterschiedlich sind.

Beispiel: Untersuchung des Unterschieds zwischen zwei Messgeräten.

```
C <- c(79.98, 80.01, 80.02, 80.04, 80.05,
    80.06, 80.07, 80.08, 80.09, 80.10)
D <- c(80.15, 80.16, 80.17, 80.18, 80.19,
    80.20, 80.21, 80.22, 80.23, 80.24)
wilcox.test(A, B, paired = FALSE)
```

```
W = 0, p-value = 1.083e-05
```

Der Wilcoxon-Test zeigt einen signifikanten Unterschied mit einem p-Wert von 0.01454. Auch hier wieder bedeutet dies, dass der Unterschied in den Messungen nicht rein zufällig ist.

10.4 Gepaarte vs. Ungepaarte Daten

Gepaarte Daten:

Definition: Bei gepaarten Daten gibt es eine natürliche Paarung der Datenpunkte zwischen zwei Gruppen, oft in Vorher-Nachher-Situationen oder bei zwei Messungen an denselben Einheiten.

Beispiele:

- Vorher-Nachher-Messungen, z.B. Blutdruck vor und nach einem Medikament bei denselben Patienten.
- Zwei Messgeräte an denselben Orten.
- Zwei Methoden zur Messung derselben Variable an denselben Probanden.

Ungepaarte Daten:

Definition: Ungepaarte Daten stammen aus zwei unabhängigen Gruppen ohne Verbindung zwischen den Datenpunkten.

Beispiele:

- Zwei unabhängige Gruppen von Patienten, die mit verschiedenen Medikamenten behandelt werden.
- Gewichte von Fischen in zwei verschiedenen Seen.
- Zwei unabhängige Stichproben aus einer Bevölkerung.

Erkennung in Aufgaben:

- **Fragestellung lesen**: Hinweise wie "Vorher" und "Nachher" oder "zwei Messungen an denselben Einheiten" deuten auf gepaarte Daten hin.
- **Datenstruktur prüfen**: Wenn die Daten als Paare vorliegen (z.B. Vorher-Nachher-Werte), sind sie gepaart.

10.5 Aufgaben

10.5.1 A10.2 Zwei Tiefen-Messgeräte

Zwei Tiefen-Messgeräte messen für die Tiefe einer Gesteins-Schicht an 9 verschiedenen Orten die folgenden Werte:

Messgerät A	120	265	157	187	219	288	156	205	163
Messgerät B	127	281	160	185	220	298	167	203	171
Differenz d_i	-7	-16	-3	2	-1	-10	-11	2	-8

Die Kennzahl für die Differenz $\bar{d_n}$ beträgt -5.78, die Standardabweichung $\sigma_D = 6.2$. Es wird vermutet, dass Gerät B systematisch grössere Werte misst. Bestätigen die Messwerte diese Vermutung oder ist eine zufällige Schwankung als Erklärung plausibel?

a) Handelt es sich um verbundene (gepaarte) oder um unabhängige Stichproben?

Da die Messungen von zwei Geräten an denselben Orten durchgeführt wurden, handelt es sich um gepaarte Stichproben.

b) Führen Sie einen t-Test auf dem Niveau $\alpha = 0.05$ durch. Formulieren Sie explizit: Modellannahmen, Nullhypothese, Alternative, und Testergebnis.

Modellannahmen: Die Differenzen der Messwerte sind normalverteilt.

Nullhypothese (H_0): Die mittlere Differenz der Messwerte beträgt 0 ($\mu_D = 0$).

Alternative Hypothese (H_A): Die mittlere Differenz der Messwerte ist kleiner als 0 ($\mu_D < 0$).

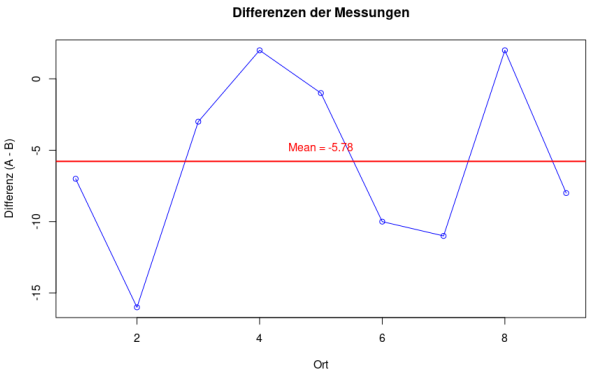
```
# Daten
A <- c(120, 265, 157, 187, 219, 288, 156,
    205, 163)
B <- c(127, 281, 160, 185, 220, 298, 167,
    203, 171)
d <- A - B

# t-Test
t.test(d, alternative = "less", mu = 0)
```

```
t = -2.773, df = 8, p-value = 0.01304
```

Der p-Wert von 0.01304 liegt unter dem Signifikanzniveau von 0.05, somit wird die Nullhypothese H_0 verworfen. Dies bedeutet, dass die Messwerte der beiden Geräte systematisch unterschiedlich sind, und Gerät B tatsächlich grössere Werte misst.

Um die Unterschiede zwischen den Messgeräten zu visualisieren, können wir die Differenzen der Messwerte plotten.

```
# Differenzen plotten
plot(d, type = "o", col = "blue", xlab = "Ort
    ", ylab = "Differenz (A - B)", main = "
    Differenzen der Messungen")
abline(h = mean(d), col = "red", lwd = 2)
text(x = 5, y = mean(d) + 1, labels = paste("
    Mean =", round(mean(d), 2)), col = "red
    ")
```



Dieser Plot zeigt die Differenzen der Messwerte an den verschiedenen Orten sowie den Mittelwert der Differenzen als rote Linie. Der negative Mittelwert bestätigt die systematisch höheren Messwerte von Gerät B.

10.5.2 A10.7 Fieber-Medikament

Die Körpertemperatur von 10 Patienten wird zum Zeitpunkt der Verabreichung eines Medikaments T_1 und 2 Stunden später T_2 gemessen. Es soll überprüft werden, ob dieses Medikament eine fiebersenkende Wirkung hat.

Patient Nr.	1	2	3	4	5	6	7	8	9	10
Temp.1 in C	39.1	39.3	38.9	40.6	39.5	38.4	38.6	39.0	38.6	39.2
Temp.2 in C	38.1	38.3	38.8	37.8	38.2	37.3	37.6	37.8	37.4	38.1

a) Handelt es sich um einen gepaarten oder ungepaarten Test? Begründen Sie Ihre Antwort.
b) Handelt es sich um einen ein- oder zweiseitigen Test? Begründen Sie Ihre Antwort
c) Formulieren Sie Null- und Alternativhypothese
d) Wir nehmen an, die Daten seien normalverteilt. Welchen Test wählen Sie? Führen Sie den Test mit R auf Signifikanzniveau 5% durch.
e) Wenn wir nicht davon ausgehen können, dass die Daten normalverteilt sind, welchen Test wählen Sie? Führen Sie diesen auf Signifikanzniveau 5% durch.
f) Erklären Sie den Unterschied der p-Werte in Teilaufgaben d) und e)

a) Gepaarter oder ungepaarter Test:

Es handelt sich um einen gepaarten Test, da die Temperaturen der gleichen Patienten vor und nach der Verabreichung des Medikaments gemessen wurden.

b) Ein- oder zweiseitiger Test:

Da geprüft werden soll, ob das Medikament eine fiebersenkende Wirkung hat, handelt es sich um einen einseitigen Test.

c) Null- und Alternativhypothese:

$$H_0 : \mu_D = 0 \text{ (keine Wirkung)}$$

$$H_A : \mu_D > 0 \text{ (senkt die Temperatur)}$$

d) Test bei Normalverteilung:

Wir verwenden einen gepaarten t-Test, um zu überprüfen, ob die Differenz der Temperaturen vor und nach der Medikamentenverabreichung signifikant ist. Der folgende R-Code führt diesen Test durch:

```r
temp_vorher <- c(39.1, 39.3, 38.9, 40.6,
    39.5, 38.4, 38.6, 39.0, 38.6, 39.2)
temp_nachher <- c(38.1, 38.3, 38.8, 37.8,
    38.2, 37.3, 37.6, 37.8, 37.4, 38.1)

t_test_result <- t.test(temp_nachher,
    temp_vorher, paired = TRUE, alternative
    = "less")
print(t_test_result)
```

```
t = -6.7693, df = 9, p-value = 0.0001554
```

Der p-Wert von 0.0001554 zeigt, dass die Temperatur nach der Medikamentenverabreichung signifikant gesunken ist. Die Nullhypothese wird verworfen, und es kann angenommen werden, dass das Medikament eine fiebersenkende Wirkung hat.

e) Test bei nicht-normalverteilten Daten:

Wenn wir nicht davon ausgehen können, dass die Daten normalverteilt sind, verwenden wir den Wilcoxon-Vorzeichen-Rang-Test:

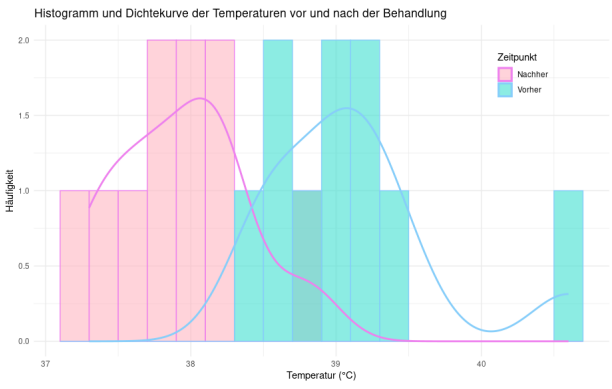
```r
wilcox_test_result <- wilcox.test(temp_vorher
    , temp_nachher, paired = TRUE,
    alternative = "greater")
print(wilcox_test_result)
```

```
V = 55, p-value = 0.002865
```

Der p-Wert von 0.002865 ist hier kleiner als 0.05. Somit ist die Differenz statistisch signifikant. Wir können also davon ausgehen, dass das Medikament fiebersenkend ist.

f) Unterschied der p-Werte in d) und e):

Der p-Wert des Wilcoxon-Tests ist grösser als der p-Wert des t-Tests. Da der Wilcoxon-Test keine Normalverteilung voraussetzt, kommt eine zusätzliche Unsicherheit hinzu, was den p-Wert erhöht. Der t-Test ist präziser, wenn die Daten normalverteilt sind, jedoch ist der Wilcoxon-Test oft vorzuziehen, wenn die Verteilung der Daten unbekannt ist.



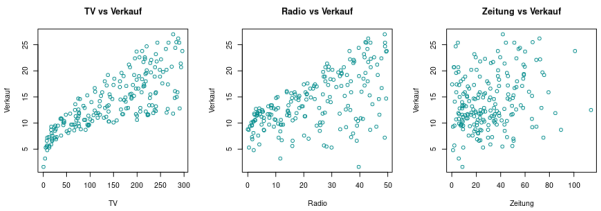
11 Lineare Regression

11.1 Einführung

In der linearen Regression geht es darum, den Zusammenhang zwischen einer Zielgrösse Y und einer oder mehreren Prädiktorvariablen $X_1, X_2, \ldots, X_p$ zu modellieren. Dies ist ein wichtiger Ausgangspunkt im Machine Learning. Ein Beispiel ist die Analyse der Werbebudgets und Verkaufszahlen einer Firma, um eine Strategie zur Verkaufssteigerung zu entwickeln.

11.2 Beispiel - Werbedaten

Die Werbedaten enthalten Verkaufszahlen und die entsprechenden Werbebudgets für TV, Radio und Zeitung in verschiedenen Märkten.



Diese Streudiagramme zeigen den Zusammenhang zwischen den Werbebudgets und den Verkaufszahlen. Besonders deutlich wird der Zusammenhang zwischen TV-Werbung und Verkauf.

11.3 Mathematische Modellierung

Die Beziehung zwischen der Zielgrösse Y und den Prädiktoren $X_1, X_2, \ldots, X_p$ wird durch die Funktion f und einen zufälligen Fehlerterm ε modelliert:

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

Dabei ist ε der zufällige Fehlerterm mit Mittelwert 0, der nicht durch die Prädiktoren erklärt werden kann.

Prädiktoren sind Variablen, die verwendet werden, um ein Ergebnis oder eine abhängige Variable vorherzusagen. In der statistischen Modellierung und speziell in der linearen Regression sind Prädiktoren die unabhängigen Variablen, die als Eingaben in das Modell eingehen, um die Zielgrösse oder die abhängige Variable zu erklären.

11.3.1 Lineares Modell

In der linearen Regression wird angenommen, dass es eine lineare Beziehung zwischen den Prädiktoren (unabhängigen Variablen) und der Zielgrösse (abhängigen Variable) gibt. Die Beziehung kann durch die folgende Gleichung beschrieben werden.

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

Die Koeffizienten $\beta_0, \beta_1, \ldots, \beta_p$ werden so geschätzt, dass die Summe der quadrierten Abstände der beobachteten Werte Y von den vorhergesagten Werten $\hat{Y}$ minimiert wird. Diese Methode nennt sich Methode der kleinsten Quadrate.

11.4 Schätzung der Parameter

Bei der Schätzung der Parameter in der linearen Regression geht es darum, die Koeffizienten zu bestimmen, die die Beziehung zwischen den Prädiktoren und der Zielgrösse am besten beschreiben. Das Ziel ist es, ein Modell zu erstellen, das die Daten so gut wie möglich erklärt.

```
1 lm_model <- lm(Verkauf ~ TV + Radio + Zeitung
      , data = Werbung)
2 summary(lm_model)
```

```
Residuals:
    Min 1Q Median 3Q Max
-8.8277 -0.8908 0.2418 1.1893 2.8292

Coefficients:
            Estimate Std. Error t value Pr(>|
                t|)
(Intercept) 2.938889 0.311908 9.422 <2e-16
     ***
TV 0.045765 0.001395 32.809 <2e-16 ***
Radio 0.188530 0.008611 21.893 <2e-16 ***
Zeitung -0.001037 0.005871 -0.177 0.86
---
```

```
Signif. codes: 0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘
     0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 1.686 on 196 degrees
      of freedom
Multiple R-squared: 0.8972, Adjusted R-
     squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value:
     < 2.2e-16
```

Die Ausgabe zeigt die geschätzten Koeffizienten für das lineare Regressionsmodell, das den Verkauf als Funktion der Werbeausgaben für TV, Radio und Zeitung beschreibt. Der geschätzte Wert ist 2.938889. Dies ist der erwartete Verkaufswert wenn keine Ausgaben für TV, Radio oder Zeitung gemacht werden. Beim TV zeigt der Koeffizient (0.045765) an, dass eine ERhöhung der TV-Werbeausgaben um eine Einheit (z.B. 1000 Dollar) den Verkauf um etwa 0.046 Einheiten erhöht, was statistisch signifikant ist, da $p < 2 - 16$.

Der Residual standard error gibt an, wie weit die tatsächlichen Verkaufszahlen im Durchschnitt von den durch das Modell vorhergesagten Verkaufszahlen abweichen. Der Multiple R-squared (0.8972) zeigt uns, dass etwa 89.72% der Variabilität in den Verkaufszahlen durch das Modell erklärt wird. Der Adjusted R-squared Wert ist eine bereinigte Version des R-squared, die die Anzahl der Prädiktoren im Modell berücksichtigt.

11.5 Vertrauensintervalle und Hypothesentests

Vertrauensintervalle geben an, in welchem Bereich die wahren Werte der Parameter mit einer bestimmten Wahrscheinlichkeit liegen. Hypothesentests überprüfen, ob ein Zusammenhang zwischen den Prädiktoren und der Zielgrösse besteht.

```
1 confint(lm_model, level = 0.95)
```

```
                2.5 % 97.5 %
(Intercept) 2.32376228 3.55401646
TV 0.04301371 0.04851558
Radio 0.17154745 0.20551259
Zeitung -0.01261595 0.01054097
```

Hier zeigt das Vertrauensintervall für den Koeffizienten der TV-Werbung (0.043 bis 0.049), dass wir zu 95% sicher sind, dass der wahre Wert des Koeffizienten in diesem Bereich liegt. Da das Intervall keine Null enthält, ist TV-Werbung ein signifikanter Prädiktor für den Verkauf. Im Gegensatz dazu umfasst das Intervall für die Zeitungswerbung (-0.013 bis 0.011) die Null (da $0 \in (-0.013, 0.011)$), was darauf hinweist, dass dieser Prädiktor keinen signifikanten Einfluss auf den Verkauf hat.

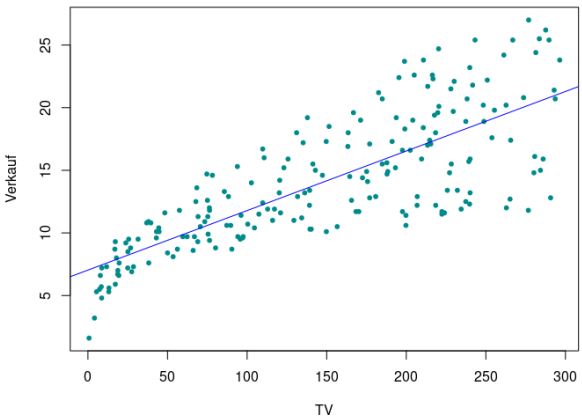
11.6 R^2-Statistik

Die R^2-Statistik misst den Anteil der Variabilität in der Zielgrösse, der durch das Modell erklärt wird. Ein Wert nahe bei 1 zeigt ein gutes Modell an, ein Wert nahe bei 0 ein schlechtes Modell.

```
1 summary(lm_model)$r.squared
```

11.7 Beispiel - Einfache lineare Regression

```
1 lm_tv <- lm(Verkauf ~ TV, data = Werbung)
2 summary(lm_tv)
3
4 plot(Werbung$TV, Werbung$Verkauf, col = "
      darkcyan", xlab = "TV", ylab = "Verkauf
      ", pch = 20)
5 abline(lm_tv, col = "blue")
```



Die resultierende Regressionsgerade zeigt, wie sich die Verkaufszahlen in Abhängigkeit von den TV-Werbebudgets verhalten.

11.8 Überanpassung und Modellwahl

Ein zu kompliziertes Modell kann zu Überanpassung führen, wobei Fehler und Ausreisser zu stark berücksichtigt werden. In vielen Fällen ist ein einfaches lineares Modell ausreichend und bevorzugt wegen seiner Interpretierbarkeit.

```
# Vergleich verschiedener Modelle
lm_quad <- lm(Verkauf ~ poly(TV, 2), data = Werbung) # Quadratisches Modell
summary(lm_quad)

# Plot der quadratischen Anpassung
plot(Werbung$TV, Werbung$Verkauf, col = "darkcyan", xlab = "TV", ylab = "Verkauf", pch = 20)
points(Werbung$TV, predict(lm_quad), col = "red", pch = 4)
```

11.9 Aufgaben

11.9.1 A11.1

a) Untersuchen Sie den Datensatz `Auto` (in Bibliothek ISLR enthalten) und `?Auto`

b) Stellen Sie das Modell für eine ewinfache lineare Regression mit mpg als Zielvariable und horsepower als Prädiktor auf.

c) Verwenden Sie die `summary()`-Funktion um die Resulate zu printen. Kommentieren Sie folgendes:

   i) Gibt es einen Zusammenhang zwischen der Zielgrösse und dem Prädiktor?

   ii) Wie interpretieren Sie die Koeffizienten für `(intercept)` und `horsepower`? Ist der Zusammenhang positiv oder negativ?

   iii) Bestimmen Sie die Vertrauensintervalle (mit `confint()`) und interpretieren Sie diese.

   iv) Interpretieren Sie den R^2-Wert.

d) Plotten Sie die Zielvariable und den Prädiktor mit der Regressionsgeraden (`abline`). Wie interpretieren Sie diesen Plot im Vergleich zum `summary()`-Output.

a) Untersuchung des Datensatzes:

Der Datensatz `Auto` wird mit den folgenden Befehlen untersucht:

```
# Bibliothek laden und Datensatz untersuchen
library(ISLR)
data(Auto)
head(Auto)
?Auto
```

Die ersten Zeilen des Datensatzes zeigen die Variablen und einige Beobachtungen. Mit `?Auto` wird eine detaillierte Beschreibung des Datensatzes angezeigt.

b) Einfaches lineares Modell erstellen:

Wir stellen das Modell für eine einfache lineare Regression mit `mpg` als Zielvariable und `horsepower` als Prädiktor auf:

```
# Einfaches lineares Modell erstellen
lm_model <- lm(mpg ~ horsepower, data = Auto)
summary(lm_model)
```

c) Regression durchführen und Ergebnisse interpretieren:

i) **Zusammenhang zwischen Zielgrösse und Prädiktor**:

Der `summary()`-Befehl liefert die folgenden Ergebnisse:

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
    Min 1Q Median 3Q Max
-13.571 -3.259 -0.343 2.763 16.924

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861 0.717499 55.66 <2e-16 ***
horsepower -0.157845 0.006446 -24.49 <2e-16 ***
---
Signif. codes: 0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 4.906 on 390 degrees of freedom
```

```
Multiple R-squared: 0.6059, Adjusted
      R-squared: 0.6049
F-statistic: 599.7 on 1 and 390 DF,
     p-value: < 2.2e-16
```

Es gibt einen signifikanten negativen Zusammenhang zwischen der Zielgrösse `mpg` und dem Prädiktor `horsepower`. Der p-Wert für `horsepower` ist kleiner als 2e-16, was auf einen starken Zusammenhang hinweist.

ii) **Koeffizienten interpretieren**:

Der Intercept (β_0) beträgt 39.935861 und der Koeffizient für `horsepower` (β_1) beträgt -0.157845. Das bedeutet, dass mit jeder Einheit Zunahme von `horsepower` die `mpg` um etwa 0.157845 Einheiten abnimmt. Der Zusammenhang ist negativ.

iii) **Vertrauensintervalle**:

Die Vertrauensintervalle für die Koeffizienten werden mit `confint()` bestimmt:

```
# Vertrauensintervalle der
    Koeffizienten
confint(lm_model, level = 0.95)
```

```
                2.5 % 97.5 %
(Intercept) 38.522500 41.349222
horsepower -0.170517 -0.145173
```

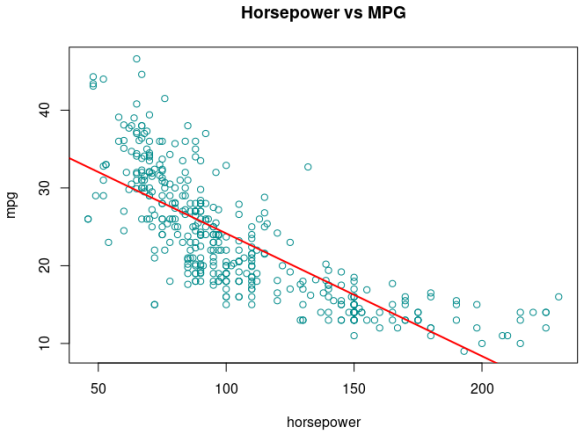
Das 95%-Vertrauensintervall für `horsepower` ist (-0.170517, -0.145173), was bedeutet, dass wir mit 95%-iger Sicherheit sagen können, dass der wahre Koeffizient in diesem Bereich liegt. Da das Intervall keine Null umfasst, ist der Prädiktor signifikant.

iv) **R^2-Wert**:

Der Multiple R-squared-Wert beträgt 0.6059. Das bedeutet, dass etwa 60.59% der Variabilität in `mpg` durch das lineare Modell erklärt wird. Ein höherer R^2-Wert weist auf eine bessere Modellanpassung hin.

d) Plot der Zielvariable und des Prädiktors mit Regressionsgerade:

```
plot(Auto$horsepower, Auto$mpg, col = "
    darkcyan", xlab = "horsepower", ylab = "
    mpg", main = "Horsepower vs MPG")
abline(lm_model, col = "red", lwd = 2)
```



Der Plot zeigt die Beziehung zwischen `horsepower` und `mpg`. Die rote Linie stellt die Regressionsgerade dar, die die Richtung und Stärke des Zusammenhangs zeigt. Im Vergleich zum Summary-Output bestätigt der Plot den negativen Zusammenhang zwischen `horsepower` und `mpg`, wobei die Punkte entlang der Regressionsgeraden streuen.

```
# Bibliothek laden und Datensatz untersuchen
library(ISLR)
data(Auto)
head(Auto)
?Auto

# Einfaches lineares Modell erstellen
lm_model <- lm(mpg ~ horsepower, data = Auto)
summary(lm_model)

# Vertrauensintervalle der Koeffizienten
confint(lm_model, level = 0.95)

# Plot der Zielvariable und des Praediktors
    mit Regressionsgerade
plot(Auto$horsepower, Auto$mpg, col = "
    darkcyan", xlab = "horsepower", ylab = "
    mpg", main = "Horsepower vs MPG")
abline(lm_model, col = "red", lwd = 2)
```

12 Multiple lineare Regression

12.1 Einleitung

Die einfache lineare Regression ist ein nützliches Verfahren, um einen Output aufgrund einer einzelnen erklärenden Variablen vorherzusagen. In der Praxis hängt der Output jedoch oft von mehreren erklärenden Variablen ab. Ein typisches Beispiel ist der Datensatz Werbung, in dem der Zusammenhang zwischen TV-Werbung und Verkauf untersucht wird. Zusätzlich gibt es Daten für Werbeausgaben für Radio und Zeitung. Die Frage ist, ob und wie sich diese zusätzlichen Werbeausgaben auf den Verkauf auswirken.

12.2 Beispiel: Datensatz Werbung

Für die Werbedaten können separate einfache Regressionen für jede Werbeausgabe durchgeführt werden:

```
# Einfache Regression von Verkauf auf TV
lm(TV ~ Verkauf, data = Werbung)

# Einfache Regression von Verkauf auf Radio
lm(Radio ~ Verkauf, data = Werbung)

# Einfache Regression von Verkauf auf Zeitung
lm(Zeitung ~ Verkauf, data = Werbung)
```

Die Ergebnisse der einfachen Regressionen sind jedoch nicht zufriedenstellend, da nicht klar ist, wie man für gegebene Werte der drei erklärenden Variablen eine Vorhersage für den Verkauf machen soll. Jede Regressionsgleichung ignoriert die anderen Variablen, was zu irreführenden Schätzungen führen kann, insbesondere wenn die erklärenden Variablen korrelieren.

12.3 Multiple lineare Regression

Eine bessere Methode ist die Verwendung aller erklärenden Variablen in einem Modell. Das multiple lineare Regressionsmodell für den Datensatz Werbung ist:

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

wobei:

- β_0: Achsenabschnitt (Intercept)
- β_1: Koeffizient für TV-Werbung
- β_2: Koeffizient für Radio-Werbung
- β_3: Koeffizient für Zeitungswerbung

Die Berechnungen und Interpretationen für das multiple Modell sind ähnlich wie beim einfachen linearen Modell, jedoch meist komplexer. Grafische Methoden entfallen für Systeme mit mehr als zwei erklärenden Variablen.

12.3.1 Beispiel: Einkommen

Ein weiteres Beispiel für eine multiple lineare Regression ist der Datensatz Einkommen, bei dem das Einkommen sowohl von der Ausbildung als auch von der Erfahrung abhängt:

$$\text{Einkommen} = \beta_0 + \beta_1 \cdot \text{Ausbildung} + \beta_2 \cdot \text{Erfahrung} + \epsilon$$

12.4 Schätzung der Parameter

Bei der Schätzung der Parameter in der linearen Regression geht es darum, die Koeffizienten zu bestimmen, die die Beziehung zwischen den Prädiktoren und der Zielgrösse am besten beschreiben. Das Ziel ist es, ein Modell zu erstellen, das die Daten so gut wie möglich erklärt.

```
# Lineares Modell fuer die Werbedaten
lm_model <- lm(Verkauf ~ TV + Radio + Zeitung
     , data = Werbung)
summary(lm_model)
```

Die Ausgabe liefert die geschätzten Koeffizienten, Standardfehler, t-Werte und p-Werte, die zur Beurteilung der Signifikanz der Prädiktoren verwendet werden.

```
# Residuals:
# Min 1Q Median 3Q Max
# -8.8277 -0.8908 0.2418 1.1893 2.8292
#
# Coefficients:
# Estimate Std. Error t value Pr(>|t|)
```

```
# (Intercept) 2.938889 0.311908 9.422 <2e-16
    ***
# TV 0.045765 0.001395 32.809 <2e-16 ***
# Radio 0.188530 0.008611 21.893 <2e-16 ***
# Zeitung -0.001037 0.005871 -0.177 0.860
#
# Residual standard error: 1.686 on 196
    degrees of freedom
# Multiple R-squared: 0.8972, Adjusted R-
    squared: 0.8956
# F-statistic: 570.3 on 3 and 196 DF, p-value
    : < 2.2e-16
```

Analyse der Ausgabe:

- Der p-Wert für die Prädiktoren TV und Radio ist nahezu null, was darauf hinweist, dass sie signifikant zur Vorhersage des Verkaufs beitragen.
- Der Koeffizient für Zeitung ist nicht signifikant, was darauf hinweist, dass Zeitungswerbung keinen signifikanten Einfluss auf den Verkauf hat.
- Der R^2-Wert von 0.8972 zeigt an, dass das Modell etwa 89.72% der Variabilität der Verkaufsdaten erklärt.

12.5 Interpretation der Koeffizienten

- **TV**: Ein Anstieg der TV-Werbeausgaben um 1000 CHF führt zu einem durchschnittlichen Anstieg des Verkaufs um 45.76 Einheiten.
- **Radio**: Ein Anstieg der Radio-Werbeausgaben um 1000 CHF führt zu einem durchschnittlichen Anstieg des Verkaufs um 188.53 Einheiten.
- **Zeitung**: Die Zeitungswerbung hat keinen signifikanten Einfluss auf den Verkauf.

12.5.1 Vertrauensintervalle

Vertrauensintervalle geben an, in welchem Bereich die wahren Werte der Parameter mit einer bestimmten Wahrscheinlichkeit liegen. Hypothesentests überprüfen, ob ein Zusammenhang zwischen den Prädiktoren und der Zielgrösse besteht.

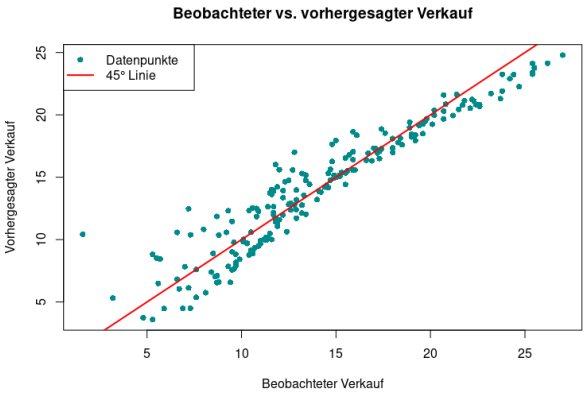
```
# Vertrauensintervalle der Koeffizienten
confint(lm_model, level = 0.95)
```

Die Ausgabe zeigt, dass die wahren Werte für Intercept, TV und Radio innerhalb der angegebenen Intervalle liegen. Das Intervall für Zeitungswerbung umfasst die Null, was darauf hinweist, dass dieser Prädiktor keinen signifikanten Einfluss auf den Verkauf hat.

```
2.5 % 97.5 %
(Intercept) 6.129719 7.935468
TV 0.042231 0.052843
Radio 0.206765 0.085768
Zeitung -0.001518 0.001798
```

Wir können bei diesem Output davon ausgehen, dass TV der einzige signifikante Prädiktor ist, da das 95%-Vertrauensintervall$(0.042, 0.052)$ Null nicht enthält und alle Werte im Intervall kleiner oder gleich dem Signifikanzniveau 0.05 sind.

12.6 Visualisierung



Der Plot zeigt die Beziehung zwischen den beobachteten Verkaufswerten und den vorhergesagten Verkaufswerten aus einem multiplen linearen Regressionsmodell. Jeder Punkt repräsentiert ein Datenpaar aus beobachtetem und vorhergesagtem Wert. Die rote Linie repräsentiert die ideale 45 deg-Linie, bei der die vorhergesagten Werte genau den beobachteten Werten entsprechen würden.

Die einfache lineare Regression verwendet nur einen Prädiktor

12.7 Aufgaben

12.7.1 A12.1 Boston

a) Definieren Sie ein multiples lineares Regressionsmodell mit der Zielvariable `medv` und den Prädiktoren `lstat` und `age`. Definieren Sie das Modell und interpretieren Sie alle Werte in der Ausgabe `summary()` (Koeffizienten, p-Werte, R^2-Wert, p-Wert und F-Statistik).

b) Der Boston-Datensatz enthält 13 Variablen, und es wäre also umständlich, dies alles eingeben zu müssen, um eine Regression mit allen Prädiktoren zu erstellen. Stattdessen können wir folgende Kurzschreibweise nutzen: `lm(medv ~ ., data = Boston)`.

Interpretieren Sie in der `summary()` Ausgabe den Koeffizienten von `age` und den entsprechenden p-Wert, vergleichen Sie diesen mit der Ausgabe in a) und erklären Sie den Unterschied.

c) Der Wert von R^2 ist grösser als der in a) berechnete Wert. Erläutern Sie.

d) Mit Hilfe der Funktion `lm()` ist es einfach, Interaktionsterme in ein lineares Modell aufzunehmen. Die Syntax `lstat:black` weist R an, einen Interaktionsterm zwischen `lstat` und `black` zu berücksichtigen.

Die Syntax `lstat * age` beinhaltet gleichzeitig `lstat`, `age` und den Interaktions-Begriff `lstat:age` als Prädiktoren; es ist eine Abkürzung für `lstat + age + lstat:age`.

Diskutieren Sie nochmals alle Werte in der `summary()` von `lstat * age` wie in a).

Wir untersuchen den Datensatz `Boston` aus dem letzten Übungsblatt weiter.

Um ein multiples lineares Regressionsmodell unter Verwendung der kleinsten Quadrate anzupassen, verwenden wir wieder die Funktion `lm()`. Die Syntax `lm(y ~ x1 + x2 + x3)` wird verwendet, um ein Modell mit drei Prädiktoren, `x1`, `x2` und `x3`, zu erstellen. Die Funktion `summary()` gibt jetzt die Regressionskoeffizienten für alle Prädiktoren aus.

a) Modell mit `lstat` und `age`:

```
# Bibliotheken laden und Datensatz untersuchen
library(MASS)
data(Boston)
head(Boston)
?Boston

# Modell erstellen
lm_model_a <- lm(medv ~ lstat + age, data = Boston)
summary(lm_model_a)
```

```
Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min 1Q Median 3Q Max
-15.981 -3.978 -1.283 1.968 23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276 0.73085 45.458 < 2e-16 ***
lstat -1.03207 0.04819 -21.416 < 2e-16 ***
age 0.03454 0.01223 2.826 0.00491 **
---
Signif. codes: 0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared: 0.5513, Adjusted R-squared: 0.5495
F-statistic: 309 on 2 and 503 DF, p-value: < 2.2e-16
```

Da $\hat{\beta}_0 = 33.22$ können wir sagen: In Vierteln, in denen es keine Bevölkerung mit niedrigerem Status und keine vor 1940 gebauten Einheiten gibt, liegt der mittlere Wert der Häuser bei \$33 220.

$\hat{\beta}_1 = -1.03$: Für jedes zusätzliche Porzent der Bevölkerung mit niedrigerem Status sinkt der mittlere Wert um \$1030.

$\hat{\beta}_2 = 0.03$: Für jedes zusätzliche Prozent der Einheiten, die vor 1949 gebaut wurden, erhöht sich der mittlere Wert um \$30.

Alle p-Werte sind signifikant. Der R^2-Wert beträgt 0.5513, daher werden etwa 55% der Variation durch das Modell erklärt. Der p-Wert des F-

Wertes liegt unterhalb des Signifikanzniveaus und ist daher signifikant. Die Nullhypothese wird abgelehnt.

b) Modell mit allen Prädiktoren:

```
# Modell mit allen Praediktoren erstellen
lm_model_b <- lm(medv ~ ., data = Boston)
summary(lm_model_b)
```

```
Call:
lm(formula = medv ~ ., data = Boston)

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
... ... ... ... ...
age 6.922e-04 1.321e-02 0.052 0.958229
... ... ... ... ...
Multiple R-squared: 0.7406, Adjusted R-
     squared: 0.7338
F-statistic: 108.1 on 13 and 492 DF, p-value:
      < 2.2e-16
```

`age` ist hier fast 1 also gar nicht signifikant. In a) aber ist dieser signifikant, was bedeutet, dass die Variable stark mit anderen Variablen korrelieren muss (siehe `d`)).

c) Vergleich der R^2-Werte:

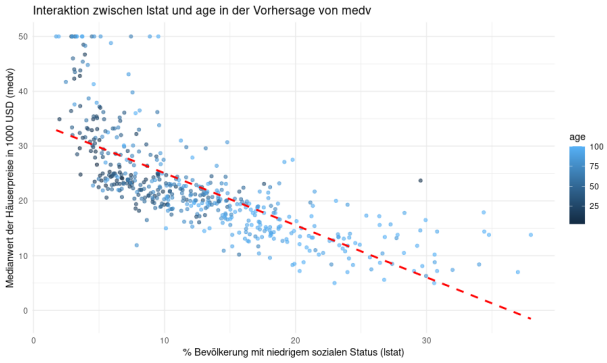
Der höhere R^2-Wert im Modell b) zeigt, dass die zusätzliche Einbeziehung weiterer Prädiktoren zu einer besseren Erklärung der Variabilität von `medv` führt. Der R^2-Wert ist höher als im Modell aus a) und beträgt 0.7406, was bedeutet, dass etwa 74% der Variabilität von `medv` durch dieses Modell erklärt werden kann.

d) Interaktionsterme:

```
# Modell mit Interaktionsterm
lm_interaction <- lm(medv ~ lstat * age, data
      = Boston)
summary(lm_interaction)
```

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

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.34170 1.56220 23.259 < 2e-16
     ***
lstat -1.39212 0.16754 -8.310 < 2e-16 ***
age -0.00072 0.01983 -0.036 0.971
lstat:age 0.00414 0.00222 1.865 0.063 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'
     0.05 '.' 0.1 ' ' 1

F-statistic: 209 on 3 and 502 DF, p-value: <
     2.2e-16
```

Hier sehen wir, dass der Interaktionsterm `lstat:age` fast signifikant ist ($p = 0.063$), was darauf hindeutet, dass es eine mögliche Wechselwirkung zwischen den Variablen `lstat` und `age` gibt.



a) $\hat{\beta}_0 = 36.10$: In Vierteln ohne niedrigen Status und keine vor 1940 gebauten Einheiten liegt der mittlere Wert der Häuser bei \$36,100.
b) $\hat{\beta}_1 = -1.39$: Jedes zusätzliche Prozent Bevölkerung mit niedrigem Status senkt den mittleren Wert um \$1390.
c) $\hat{\beta}_2 = -0.00072$: Jedes zusätzliche Prozent der vor 1940 gebauten Einheiten senkt den mittleren Wert um \$0.72. Dieser Wert ist nicht signifikant ($p = 0.97$).
d) $\hat{\beta}_3 = 0.004$: Der Interaktionsterm ($p = 0.0252$) deutet auf eine signifikante Wechselwirkung zwischen `lstat` und `age` hin.
e) Der p-Wert für `age` ist 0.97, also nicht signifikant, während er es ohne Interaktion war. Der p-Wert des Interaktionsterms beträgt 0.0252 und liegt unter dem Signifikanzniveau von 5%. Dies deutet auf eine signifikante Interaktion hin.
f) Der Wert von R^2 beträgt 0.56, daher wird etwa 56% der Variation durch das Modell erklärt.
g) Der p-Wert des F-Wertes liegt unter dem Signifikanzniveau und ist daher signifikant. Mindestens eine Variable trägt signifikant zum Modell bei.

13 Qualitative Variablen, Variablenselektion

13.1 Qualitative erklärende Variablen

Bisher haben wir angenommen, dass alle Variablen in unserem linearen Regressionsmodell quantitativ sind. Oftmals sind jedoch einige der erklärenden Variablen qualitativ. Diese qualitativen Variablen werden auch als Faktoren bezeichnet und nehmen diskrete Werte oder Levels an. Beispiele:

- **Zielvariable:** Balance (monatliche Kreditkartenrechnung, quantitativ)
- **Erklärende Variablen:** Age (Alter), Cards (Anzahl Kreditkarten), Education (Anzahl Jahre Ausbildung), Income (Einkommen), Limit (Kreditkartenlimite), Rating (Kreditwürdigkeit)
- **Qualitative Variablen:** Gender (Geschlecht), Student (Studentenstatus), Married (Familienstand), Ethnicity (Ethnie)

13.1.1 Beispiel: Geschlecht als qualitative Variable

Für die Variable `Gender` kodieren wir `Female` als 1 und `Male` als 0.

```
Credit <- read.csv("../Data/Credit.csv")
gender <- Credit$Gender == "Female"
balance <- Credit$Balance

summary(lm(balance ~ gender))
```

```
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 509.80 33.13 15.389 < 2e-16
    ***
## genderTRUE 19.73 46.05 0.429 0.669
```

- $\hat{\beta}_0 = 509.80$: Durchschnittliche Kreditkartenrechnungen für Männer.
- $\hat{\beta}_1 = 19.73$: Unterschied der durchschnittlichen Rechnungen zwischen Frauen und Männern.
- Der hohe p-Wert (0.669) zeigt, dass der Unterschied statistisch nicht signifikant ist.

13.2 Qualitative erklärende Variablen mit mehr als zwei Levels

Wenn eine qualitative Variable mehr als zwei Levels hat, benötigen wir für jedes Level eine Indikatorvariable (Dummy-Variable), ausser für eines, das als Referenzkategorie dient.

- Beispiel: Variable `Ethnicity` mit den Levels `Asian`, `Caucasian`, `African American`
- Kodierung:
  - $x_{i1} = 1$ für `Asian`, sonst 0
  - $x_{i2} = 1$ für `Caucasian`, sonst 0

```
ethnicity <- Credit$Ethnicity
summary(lm(balance ~ ethnicity))
```

```
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 531.00 46.32 11.464 <2e-16 ***
## ethnicityAsian -18.69 65.02 -0.287 0.774
## ethnicityCaucasian -12.50 56.68 -0.221
    0.826
```

- $\hat{\beta}_0 = 531.00$: Durchschnittliche Kreditkartenrechnungen für Afroamerikaner (Referenzkategorie).
- $\hat{\beta}_1 = -18.69$: Unterschied der durchschnittlichen Rechnungen zwischen Asiaten und Afroamerikanern.
- $\hat{\beta}_2 = -12.50$: Unterschied der durchschnittlichen Rechnungen zwischen Kaukasiern und Afroamerikanern.
- Beide Unterschiede sind statistisch nicht signifikant (hohe p-Werte).

13.3 Variable Selection

Die Auswahl der relevanten Variablen ist ein wichtiger Schritt in der Modellentwicklung. Es gibt verschiedene Methoden, um die besten erklärenden Variablen auszuwählen.

13.3.1 Schrittweise Vorwärtsselektion

Bei der schrittweisen Vorwärtsselektion beginnen wir mit einem Modell, das keine erklärenden Variablen enthält, und fügen eine Variable nach der anderen hinzu, die die grösste Verbesserung bringt.

```
f.full <- lm(Balance ~ ., data = Credit)
f.empty <- lm(Balance ~ NULL, data = Credit)
add1(f.empty, scope = f.full)
```

```
## Single term additions
## Model: Balance ~ NULL
## Df Sum of Sq RSS AIC
## <none> 84339912 4905.6
## Income 1 18131167 66208745 4810.7
## Limit 1 62624255 21715657 4364.8
## Rating 1 62904790 21435122 4359.6
```

Die Variable `Rating` hat den kleinsten RSS-Wert und wird daher als erste hinzugefügt.

13.3.2 Schrittweise Rückwärtsselektion

Bei der schrittweisen Rückwärtsselektion beginnen wir mit einem vollen Modell und entfernen schrittweise die am wenigsten nützliche Variable.

```
reg <- regsubsets(Balance ~ ., data = Credit,
      method = "backward", nvmax = 11)
summary(reg)$which
```

```
## (Intercept) Income Limit Rating Cards Age
## 1 TRUE FALSE FALSE TRUE FALSE FALSE
## 2 TRUE FALSE TRUE TRUE FALSE FALSE
```

Das beste Modell mit drei Variablen enthält `Income`, `Limit` und `Student`.

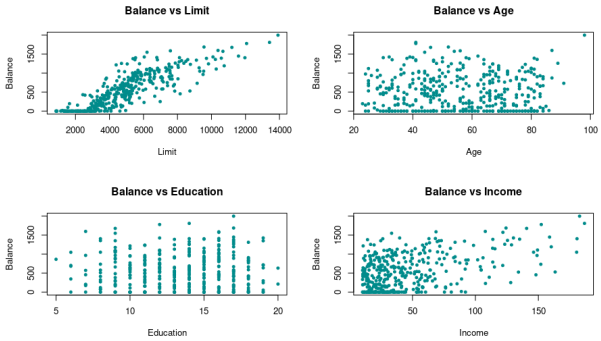
13.3.3 AIC als Kriterium

Ein alternatives Kriterium zur Auswahl der besten Variablen ist der Akaike Information Criterion (AIC).

```
add1(f.empty, scope = f.full, test = "AIC")
```

```
## Single term additions
## Model: Balance ~ NULL
## Df Sum of Sq RSS AIC
## <none> 84339912 4905.6
## Income 1 18131167 66208745 4810.7
```

Auch hier wird die Variable `Rating` als erste hinzugefügt, da sie den kleinsten AIC-Wert hat.



Die Abbildung zeigt Scatterplots der Variablen `Balance`, `Age`, `Cards`, `Education`, `Income`, `Limit` und `Rating`. Es gibt einen starken Zusammenhang zwischen `Limit` und `Balance`, während `Age` und `Education` keinen deutlichen Zusammenhang mit `Balance` aufweisen.

13.4 Aufgaben

13.4.1 A13.1 Autositze

In der Bibliothek ISLR hat es den Datensatz Carseats. Wir möchten Sales (Anzahl Kinderautositze) aufgrund von verschiedenen Prädiktoren in 400 verschiedenen Standorten vorhersagen.

Der Datensatz enthält qualitative Prädiktoren, wie ShelveLoc als Indikator der Lage im Gestell, das heisst der Platz in einem Geschäft, wo der Autositz ausgestellt ist. Der Prädiktor nimmt die drei Werte `Bad, Medium` und `Good` an. Für qualitative Variablen generiert R Dummy-Variablen automatisch.

a) Untersuchen Sie den Datensatz mit `head(Carseat) und ?Carseat`

b) Finden Sie mit `lm()` ein multiples Regressionsmodell um Sales aus Price, Urban und US vorherzusagen.

c) Interpretieren Sie die Koeffizienten in diesem Modell. Achten Sie darauf, dass einige Variablen qualitativ sind.

d) Schreiben Sie das Modell in Gleichungsform. Achten Sie darauf, dass Sie die qualitativen Variablen richtig behandeln.

e) Für welche Prädiktoren kann die Nullhypothese $H_0 : \beta_j = 0$ verworfen werden?

f) Auf der Basis der vorhergehenden Frage, finden Sie ein kleineres Modell, das nur Prädiktoren verwendet für die es Hinweise auf einen Zusammenhang mit der Zielvariablen gibt.

g) Wie genau passen die Modelle in a) und e) die Daten an?

a)

Der Datensatz besteht aus folgenden Spalten: Sales, CompPrice, Income, Advertising, Population, Price, ShelveLoc, Age, Education, Urban, US. ShelveLoc kann die Werte `Bad, Medium, Good` tragen und `Urban, US` sind Booleans (`Yes / No`). Die restlichen Spalten sind nummerische Werte.

b)

```
1  library(ISLR) # Bibliothek ISLR laden
2  data("Carseats") # Datensatz Carseats laden
3
4  #a)
5  head(Carseats)
6
7  #b)
8  modell <- lm(Carseats$Sales ~ Carseats$Price
       + Carseats$Urban + Carseats$US)
9  summary(modell)
```

```
Call:
lm(formula = Carseats$Sales ~ Carseats$Price
    + Carseats$Urban +
   Carseats$US)

Residuals:
    Min 1Q Median 3Q Max
-6.9206 -1.6220 -0.0564 1.5786 7.0581

Coefficients:
                  Estimate Std. Error t value
                      Pr(>|t|)
(Intercept) 13.043469 0.651012 20.036 < 2e-16
     ***
Carseats$Price -0.054459 0.005242 -10.389 < 2
    e-16 ***
Carseats$UrbanYes -0.021916 0.271650 -0.081
    0.936
Carseats$USYes 1.200573 0.259042 4.635 4.86e
    -06 ***
---
Signif. codes: 0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘
    0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 2.472 on 396 degrees
     of freedom
Multiple R-squared: 0.2393, Adjusted R-
    squared: 0.2335
F-statistic: 41.52 on 3 and 396 DF, p-value:
    < 2.2e-16
```

c)

- Der Koeffizient 13.04 ist schwer zu interpretieren. Wir sehen aber unter d), dass dies die mittleren Verkaufszahlen in Geschäften die in ländlichen Gegenden ausserhalb der USA erreicht werden. Wobei aber der Preis der Kindersitze nicht \$0 ist (nicht sehr realistisch).
- Der Koeffizient -0.05 zeigt uns, dass für eine Zunahme von einem Dollar durchschnittlich 0.05 Einheiten Kindersitze weniger verkauft werden.
- Der Koeffizient -0.021 besagt, dass verglichen zu ländlichen Regionen durchschnittlich 0.021 Einheiten weniger verkauft werden. Der p-Wert ist aber sehr hoch, so dass dies eher eine zufällige Abweichung ist.
- Der Koeffizient 1.2 besagt, dass verglichen zu Geschäften ausserhalb der USA, 1.2 Einheiten mehr verkauft werden.

d)

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

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

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

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

Das Modell lautet dann

$$= \beta_0 + \beta_1 \cdot \textbf{Price} + \begin{cases} \beta_2 + \beta_3 + \epsilon_i & 1 \\ \beta_2 + \epsilon_i & 2 \\ \beta_3 + \epsilon_i & 3 \\ \epsilon_i & 4 \end{cases}$$

1. falls i-te Person urban in den USA lebt
2. falls i-te Person urban nicht in den USA lebt
3. falls i-te Person ländlich in den USA lebt
4. falls i-te Person ländlich nicht in den USA lebt

e)

1. `(Intercept)`: Der p-Wert ist sehr klein, daher ist der Intercept signifikant
2. `Carseats$Price`: Der p-Wert ist ebenfalls sehr klein, daher ist dieser Prädiktor signifikant.
3. `Carseats$UrbanYes`: Der p-Wert ist hoch (0.936), daher ist dieser Prädiktor nicht signifikant.
4. `Carseats$USYes`: Der p-Wert ist sehr klein, daher ist dieser Prädiktor signifikant.

Die Nullhypothese kann für alle ausser Urban verworfen werden. Nur Urban hat einen p-Wert von > 0.05.

f)

Um ein Modell zu finden, welches nur die signifikanten Prädiktoren trägt. Nutzen wir folgendes Skript.

```
1 #f)
2 modell2 <- lm(Carseats$Sales ~ Carseats$Price
      + Carseats$US)
3 summary(modell2)
```

```
Call:
lm(formula = Carseats$Sales ~ Carseats$Price
    + Carseats$US)

Residuals:
    Min 1Q Median 3Q Max
-6.9269 -1.6286 -0.0574 1.5766 7.0515

Coefficients:
             Estimate Std. Error t value Pr
                  (>|t|)
(Intercept) 13.03079 0.63098 20.652 < 2e-16
    ***
Carseats$Price -0.05448 0.00523 -10.416 < 2e
    -16 ***
Carseats$USYes 1.19964 0.25846 4.641 4.71e-06
     ***
---
Signif. codes: 0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘
    0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 2.469 on 397 degrees
     of freedom
Multiple R-squared: 0.2393, Adjusted R-
    squared: 0.2354
F-statistic: 62.43 on 2 and 397 DF, p-value:
    < 2.2e-16
```

g)

Bei beiden Modellen ist der Zusammenhang belegt (p-Wert für F-Wert praktisch 0), aber wenn wir die R^2-Werte betrachten, so ist der mit 0.2393 relativ schlecht. Nur 23% der Variabilität der Sales können durch das Modell erklärt werden.

14 R Glossary

`abline(a, b, ...)`
Fügt eine Gerade zu einem Plot hinzu. Die Parameter `a` und `b` sind der Schnittpunkt mit der y-Achse und die Steigung der Linie.

```
plot(x, y)
abline(a = 1, b = 2, col = "red")
```

`add1(object, scope, ...)`
Fügt eine Variable zu einem Regressionsmodell hinzu und bewertet die Änderung der Anpassung.

```
add1(f.empty, scope = f.full)
```

`boxplot(x, ...)`
Erzeugt einen Boxplot zur grafischen Darstellung der Verteilung eines numerischen Vektors.

```
boxplot(x,
  main = "Boxplot Beispiel",
  xlab = "Kategorie",
  ylab = "Werte",
  col = "lightblue")
```

`coef(object)`
Extrahiert die Koeffizienten eines linearen Modells.

```
model <- lm(y ~ x)
coef(model)
```

`cor(x, y)`
Berechnet den Korrelationskoeffizienten zwischen den Variablen `x` und `y`.

```
cor(x, y)
```

`IQR(x)`
Berechnet die Interquartilsdifferenz (Quartilsabstand) eines numerischen Vektors `x`.

```
iqr_value <- IQR(x)
```

`lm(formula, data, ...)`
Erstellt ein lineares Modell basierend auf der Formel und den Daten.

```
lm_model <- lm(y ~ x, data = dataset)
```

`mean(x, na.rm = TRUE)`
Berechnet das arithmetische Mittel eines numerischen Vektors `x`. Der Parameter `na.rm` entfernt NA-Werte, bevor das Mittel berechnet wird.

```
mean_value <- mean(x, na.rm = TRUE)
```

`order(x)`
Gibt die Indizes der sortierten Elemente eines Vektors zurück.

```
order_indices <- order(x)
```

`plot(x, y, ...)`
Erstellt ein Streudiagramm der Variablen `x` und `y`.

```
plot(x, y, main="Streudiagramm", xlab="
     X-Werte", ylab="Y-Werte", col="blue
     ", pch=19)
```

`read.csv(file, ...)`
Lädt Daten aus einer CSV-Datei und speichert sie in einem Dataframe.

```
data <- read.csv("data.csv")
```

`regsubsets(formula, data, method, ...)`
Führt eine schrittweise Vorwärts- oder Rückwärtsselektion zur Auswahl der besten erklärenden Variablen durch.

```
reg <- regsubsets(Balance ~ ., data =
    Credit, method = "backward", nvmax
    = 11)
```

`summary(object)`
: Gibt eine detaillierte Zusammenfassung eines linearen Modells zurück, einschließlich der Koeffizienten und ihrer statistischen Signifikanz.

```
summary(lm_model)
```

`tapply(X, INDEX, FUN)`
: Wendet eine Funktion auf Teildatensätze eines Vektors an, die durch einen Index-Vektor definiert sind.

```
mean_values <- tapply(data$value,
    data$group, mean)
```

`write.csv(x, file, row.names)`
: Schreibt ein Dataframe in eine CSV-Datei. Der Parameter `row.names` gibt an, ob Zeilennamen gespeichert werden sollen.

```
write.csv(data, "output.csv", row.names
     = FALSE)
```