# Statistik. Eine Einführung mit R

Bodo Sturm

GUC-Verlag, 2016

Lösungshinweise für die Übungsaufgaben, Stand: 2. Mai 2018

*Gliederung*

# Kapitel 1: Einführung

## Aufgabe 1.1: Merkmale

*Geben Sie zu den folgenden Merkmalen Beispiele für statistische Einheiten und Merkmalsausprägungen an. Nennen Sie Merkmalsart (qualitativ, quantitativ), Merkmalstyp (diskret, stetig) und Skalierung (nominal, ordinal, kardinal): a) Haarfarbe, b) Verdienst, c) Abiturnote in Mathe, d) Geschlecht, e) Beruf, f) Kontobewegungen in € pro Monat*

**Lösung:**

| **Merkmal** | **statistische Einheiten (Beispiel)** | **Merkmals-ausprägungen (Beispiel)** | **Merkmalsart Merkmalstyp** | **Skalierung** |
|---|---|---|---|---|
| Haarfarbe | Männer im Alter zwischen 60 und 65 | schwarz, braun, blond, grau | qualitativ | nominal |
| Verdienst | Studentische Hilfskräfte | 7 – 12 €/Stunde | quantitativ diskret | kardinal |
| Abiturnote in Mathe | Jahrgang 2000 | 0 – 15 Punkte | qualitativ | ordinal |
| Geschlecht | Studenten der HTWK | weiblich/ männlich | qualitativ | nominal |
| Beruf | Mitglieder der FDP | Arbeiter, Angestellte, Selbständiger | qualitativ | nominal |
| Kontobewegungen in € pro Monat | Girokonten der Sparkasse Leipzig | 1000€ | quantitativ diskret | kardinal |

Hinweise:

- Merkmal = Eigenschaft einer statistischen Einheit
- Statistische Einheit = Merkmalsträger, das zu untersuchende Einzelobjekt
- Merkmalsausprägung = verschiedene Abstufungen, Kategorien oder Werte eines Merkmals
- Qualitative Merkmale sind immer diskret, weil abzählbar viele Ausprägungen vorliegen. Daher wird nur bei quantitativen Merkmalen nach diskret und stetig unterschieden.
- Ordinal skalierte Merkmale sind grundsätzlich qualitativ, werden aber gerne als quantitative Merkmale verwendet, wenn z.B. Durchschnittsnoten berechnet werden.

## Aufgabe 1.2: Untersuchungen

*Betrachten Sie die folgenden statistischen Untersuchungen. Nennen Sie jeweils die statistische Einheit und mögliche Merkmalsausprägungen. Welcher Merkmalstyp und welche Skalierung liegen vor? Haben wir es mit qualitativen oder quantitativen Merkmalen zu tun? Im Fall von quantitativen Merkmalen, in welchen Einheiten wird die Variable gemessen?*

*a) Eine Studie untersucht die Autos von Mitarbeitern einer großen Firma. Erhoben werden Baujahr, Herstellerland und Typ.*

*b) Ein Bericht einer Verbraucherschutzorganisation listet 41 Kühlschränke auf mit den Merkmalen Marke, Kosten, Höhe, Breite, Tiefe, Typ, jährliche Energiekosten, Gesamteinschätzung (gut, ausgezeichnet usw.) und Reparaturanfälligkeit (Anzahl notwendiger Reparaturen je Kühlschrank in den letzten fünf Jahren).*

c) *Das Umweltbundesamt analysiert den Kraftstoffverbrauch von Pkw. Folgende Merkmale werden erhoben: Hersteller, Typ, Gewicht, PS, Verbrauch innerstädtisch, Verbrauch Autobahn.*

## Lösung a)

| Merkmal | statistische Einheiten | Merkmals-ausprägungen | Merkmalsart Merkmalstyp | Skalierung |
|---|---|---|---|---|
| Baujahr | Autos der Mitarbeiter | 2007, 2008, 2009 | quantitativ diskret | kardinal |
| Herstellerland | Autos der Mitarbeiter | Japan, USA, Frankreich | qualitativ | nominal |
| Typ | Autos der Mitarbeiter | VW Golf, Ford Fiesta | qualitativ | nominal |

## Lösung b)

| Merkmal | statistische Einheiten | Merkmals-ausprägungen | Merkmalsart Merkmalstyp | Skalierung | Maßeinheit des Merkmals |
|---|---|---|---|---|---|
| Marke | Kühlschränke | Miele, Siemens, AEG | qualitativ | nominal | - |
| Kosten | Kühlschränke | 200-1600 Euro | quantitativ diskret | kardinal | Euro |
| Höhe | Kühlschränke | 82-187 cm | quantitativ stetig | kardinal | cm |
| Breite | Kühlschränke | 55-70 cm | quantitativ stetig | kardinal | cm |
| Tiefe | Kühlschränke | 50-70 cm | quantitativ stetig | kardinal | cm |
| Typ | Kühlschränke | TL1400, TSE1423 | qualitativ | nominal | - |
| jährliche Energiekosten | Kühlschränke | 28-167 € / Jahr | quantitativ diskret | kardinal | Euro |
| Gesamtein-schätzung | Kühlschränke | sehr gut – ungenügend | qualitativ | ordinal | - |
| Reparatur-anfälligkeit | Kühlschränke | 1-5 Reparaturen / Kühlschrank | quantitativ diskret | kardinal | Anzahl Reparaturen |

## Lösung c)

| Merkmal | statistische Einheiten | Merkmals-ausprägungen | Merkmalsart Merkmalstyp | Skalierung | Maßeinheit des Merkmals |
|---|---|---|---|---|---|
| Hersteller | PKW | VW, Toyota, Ford | qualitativ | nominal | - |
| Typ | PKW | VW Golf, Ford Fiesta | qualitativ | nominal | - |
| Gewicht | PKW | 700-2200 kg | quantitativ | kardinal | kg |
| Motorstärke | PKW | 45-450 PS | quantitativ | kardinal | PS |
| Verbrauch innerstädtisch | PKW | 3.3-12.7 Liter | quantitativ | kardinal | Liter / 100km |
| Verbrauch Autobahn | PKW | 4.5-16.8 Liter | quantitativ | kardinal | Liter / 100km |

## Kapitel 2: Darstellung und Beschreibung qualitativer Daten

### Aufgabe 2.1: Kneipe

*Eine Befragung von 25 Gästen in einer Kneipe ergab folgende Daten*

| ID | Getränk | Frau | ID | Getränk | Frau | ID | Getränk | Frau |
|---|---|---|---|---|---|---|---|---|
| 1 | Bier | 0 | 10 | Wein | 0 | 19 | alkfrei | 1 |
| 2 | Mix | 0 | 11 | Wein | 1 | 20 | alkfrei | 1 |
| 3 | Wein | 0 | 12 | Wein | 1 | 21 | Mix | 1 |
| 4 | Bier | 0 | 13 | Wein | 1 | 22 | alkfrei | 1 |
| 5 | Bier | 0 | 14 | Mix | 1 | 23 | Wein | 1 |
| 6 | alkfrei | 0 | 15 | Mix | 1 | 24 | Wein | 1 |
| 7 | Bier | 0 | 16 | Bier | 1 | 25 | alkfrei | 1 |
| 8 | Bier | 0 | 17 | Mix | 1 | | | |
| 9 | Bier | 0 | 18 | Mix | 1 | | | |

*Dabei ist „ID“ die Nummer der/des Befragten. Erhoben wurden die Merkmale „Getränk“ mit den Ausprägungen {alkfrei, Bier, Mix, Wein} und „Geschlecht“. Letztere wurde zur Variable „Frau“ umcodiert mit den Ausprägungen {0, 1}, wobei 0 = Mann und 1 = Frau.*

*a) Ermitteln Sie die Häufigkeitsverteilung und den Modus für das Merkmal „Getränk“. Stellen Sie die Verteilung graphisch dar.*

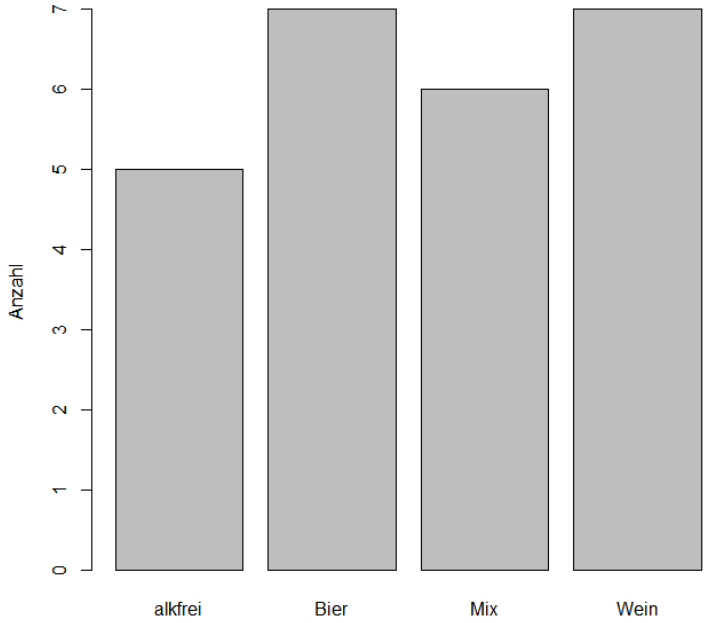
*b) Erstellen Sie die Kontingenztabelle für die gemeinsame Verteilung beider Merkmale.*

*c) Wie viel Prozent aller Frauen konsumieren alkoholfreie Getränke? Welcher Anteil der Befragten, die Wein konsumieren, sind Frauen?*

d) *Treffen Sie eine Aussage darüber, ob beide Merkmale statistisch unabhängig voneinander sind.*

**Lösung a)**

Häufigkeitsverteilung: Eine Häufigkeitsverteilung gibt an, wie oft ein vorkommender Wert (Merkmalsausprägung) auftritt. Dies kann beispielsweise anhand einer Tabelle oder einer Grafik veranschaulicht werden.



Modus: Lagemaß für qualitative Daten. Der Modus ist die Ausprägung des Merkmals mit der größten absoluten bzw. relativen Häufigkeit. In Aufgabe 1 sind die beiden häufigsten Getränke „Bier“ und „Wein“ mit einer Anzahl von 7. Die Verteilung hat somit zwei Modi und ist bimodal.

**Lösung b)**
Eine Kontingenztabelle stellt die gemeinsame Verteilung zweier (i.d.R. nominal skalierter) Merkmale dar.

| $n_{ij}$ | Frau | | |
|---|---|---|---|
| Getränk | 0(Mann) | 1(Frau) | $\sum$ |
| Alkfrei | 1 | 4 | 5 |
| Bier | 6 | 1 | 7 |
| Mix | 1 | 5 | 6 |
| Wein | 2 | 5 | 7 |
| $\sum$ | 10 | 15 | 25 |

**Lösung c)**
26.7% aller Frauen konsumieren alkoholfreie Getränke. Um diesen Wert zu berechnen, teilt man die Anzahl der Frauen, die alkoholfreie Getränke konsumieren, durch die Zahl der Frauen ($\frac{4}{15} = 0.267$).
71.4% der Befragten, die Wein konsumieren, sind Frauen ($\frac{5}{7} = 0.714$).

**Lösung d)**
Statistische Unabhängigkeit ist nicht gegeben, weil die bedingten Verteilungen des Merkmals „Getränk“ für jedes Geschlecht nicht gleich und nicht identisch der Randverteilungen sind. Vergleiche hierzu die Definition statistischer Unabhängigkeit.
Die folgende Tabelle zeigt die bedingten relativen Verteilungen ($h_{i|y_j}$) für das Merkmal Getränk und die relative Randverteilung ($h_i$) für Getränk. Die drei Verteilungen unterscheiden sich deutlich, die Abhängigkeit zwischen beiden Merkmalen ist recht groß.

| $h_{i\|y_j}$ | Frau | | |
|---|---|---|---|
| Getränk | 0(Mann) | 1(Frau) | $h_i$ |
| Alkfrei | 0.1 | 0.267 | 0.20 |
| Bier | 0.6 | 0.067 | 0.28 |
| Mix | 0.1 | 0.333 | 0.24 |
| Wein | 0.2 | 0.333 | 0.28 |
| $\sum$ | 1 | 1 | 1 |

Alternativ könnte KK* berechnet werden (vgl. Abschn. 4.1 im Buch und Lösung mit R: Option 2).

**Lösung mit R**
```
> # Generiere den Datensatz in R
> G    <- c("Bier", "Mix", "Wein", "Bier", "Bier", "alkfrei", "Bier",
+           "Bier", "Bier", "Wein", "Wein", "Wein", "Wein", "Mix", "Mix",
+           "Bier", "Mix", "Mix", "alkfrei", "alkfrei", "Mix", "alkfrei",
+           "Wein", "Wein", "alkfrei")
> F    <- c(rep(0, 10), rep(1, 15))
```

```
> data <- data.frame(cbind(G, F)); data[1:5, ]
     G F
1 Bier 0
2  Mix 0
3 Wein 0
4 Bier 0
5 Bier 0
> attach(data)
> # a) und b)
> # Säulendiagramm
> barplot(table(G), ylab = "Anzahl") # absolute Häufigkeit „Getränk“
> barplot(prop.table(table(G)), ylab = "Anteil") # relative Häufigkeit
>
> # Kontingenztabelle (addmargins(… für Randverteilungen))
> options(digits = 5)        # auf drei Stellen genau
> t(addmargins(table(F, G))) # absolute Verteilung (Tabelle s. Lösung ohne R)
        F
G         0  1 Sum
  alkfrei  1  4   5
  Bier     6  1   7
  Mix      1  5   6
  Wein     2  5   7
  Sum     10 15  25
> t(addmargins(prop.table(table(F, G))))  # relative Verteilung
        F
G           0    1  Sum
  alkfrei 0.04 0.16 0.20
  Bier    0.24 0.04 0.28
  Mix     0.04 0.20 0.24
  Wein    0.08 0.20 0.28
  Sum     0.40 0.60 1.00
>
> # d)
> # Option 1: Bedingte Verteilungen
> t(addmargins(prop.table(table(F, G), 1))) # bedingte Verteilung vertikal
        F
G                0        1      Sum
  alkfrei 0.100000 0.266667 0.366667
  Bier    0.600000 0.066667 0.666667
  Mix     0.100000 0.333333 0.433333
  Wein    0.200000 0.333333 0.533333
  Sum     1.000000 1.000000 2.000000
>
> # Die bedingten Verteilungen von G unterscheiden sich deutlich
> # von der Randverteilung => klares Indiz dafür, dass die Merkmale
> # G und F nicht unabhängig voneinander sind!
>
> t(addmargins(prop.table(table(F, G), 2))) # bedingte Verteilung horizontal
        F
G               0       1     Sum
  alkfrei 0.20000 0.80000 1.00000
  Bier    0.85714 0.14286 1.00000
  Mix     0.16667 0.83333 1.00000
  Wein    0.28571 0.71429 1.00000
  Sum     1.50952 2.49048 4.00000
>
> # Option 2: Berechnung von KK* (vgl. Kap. 4)
> chisq.test(table(F, G))  # Syntax, um Chi-squared = QK zu berechnen

        Pearson's Chi-squared test

data:  table(F, G)
X-squared = 8.67, df = 3, p-value = 0.034

Warning message:
In chisq.test(table(F, G)) : Chi-squared approximation may be incorrect
> QK  <- 8.67; m <- 2; n <- 25
> KK_star <-  sqrt(QK*m/((QK + n)*(m - 1)))
```

```
> KK_star # KK* = 0.718
[1] 0.71763

> # KK_star liegt im Intervall [0,1]. Bei 0 liegt perfekte Unabhängigkeit
> # vor, bei 1 perfekte Abhängigkeit. Daher ist die Abhängigkeit
> # zwischen beiden Merkmalen relativ groß.
> barplot(table(G), ylab = "Anzahl") # absolute Häufigkeit „Getränk“
```

**Aufgabe 2.2: Kaufabsicht**

*Bei einer Befragung von 1443 Personen ergaben sich folgende Ergebnisse für die Merkmale „Altersgruppe“ mit den Ausprägungen {unter 20, 20 bis unter 60, 60 und darüber} und „Kaufabsicht“ mit den Ausprägungen {ja, nein}. 422 Teilnehmer sind unter 20, davon beabsichtigen 100 einen Kauf. 398 Teilnehmer sind im Alter von 20 bis unter 60, davon beabsichtigen 201 einen Kauf. 623 Teilnehmer sind 60 und darüber, davon wollen 500 kaufen.*

a) *Erstellen Sie mit diesen Angaben eine Kontingenztabelle der Merkmale Altersgruppe und Kaufabsicht.*
b) *Welcher Anteil der Befragten ist 60 Jahre und älter und hat eine Kaufabsicht? Welcher Anteil der mindestens 60jährigen hat eine Kaufabsicht? Wie viel Prozent der Befragten, die keine Kaufabsicht haben, sind unter 20?*
c) *Treffen Sie eine Aussage darüber, ob beide Merkmale statistisch unabhängig voneinander sind.*

**Lösung a)**

Folgende Werte für die Kontingenztabelle sind bereits gegeben:

| $n_{ij}$ | Kaufabsicht | | |
|---|---|---|---|
| Altersgruppe | Ja | Nein | $\sum$ |
| unter 20 | 100 | | 422 |
| 20 bis unter 60 | 201 | | 398 |
| 60 und darüber | 500 | | 623 |
| $\sum$ | | | 1443 |

Mit Hilfe dieser Angaben lassen sich dann die übrigen Werte ermitteln:

| $n_{ij}$ | Kaufabsicht | | |
|---|---|---|---|
| Altersgruppe | Ja | Nein | $\sum$ |
| unter 20 | 100 | 322 | 422 |
| 20 bis unter 60 | 201 | 197 | 398 |
| 60 und darüber | 500 | 123 | 623 |
| $\sum$ | 801 | 642 | 1443 |

In relativen Häufigkeiten (zur Gesamtanzahl) ergibt sich folgende Tabelle:

| $h_{ij}$ | Kaufabsicht | | |
|---|---|---|---|
| Altersgruppe | Ja | Nein | $\sum$ |
| unter 20 | 0.0693 | 0.2231 | 0.2924 |
| 20 bis unter 60 | 0.1393 | 0.1365 | 0.2758 |
| 60 und darüber | 0.3465 | 0.0852 | 0.4317 |
| $\sum$ | 0.5551 | 0.4449 | 1.0000 |

**Lösung b)**

- Der Anteil der Befragten, die 60 Jahre und älter sind und eine Kaufabsicht haben, errechnet sich, indem man die Anzahl der mindestens 60-jährigen mit Kaufabsicht durch die Gesamtanzahl der Befragten teilt: 500/1443 = 0.3465 = 34.65%.
- Der Anteil der mindestens 60-jährigen, der eine Kaufabsicht hat (also eine bedingte relative Häufigkeit), errechnet ausschließlich durch die Zeile „60 und darüber“: 500/623 = 0.8026 = 80.26%.
- Der Anteil der unter 20-jährigen von den Befragten ohne Kaufabsicht (also eine bedingte relative Häufigkeit) berechnet sich folgendermaßen: 322/642 = 0.5016 = 50.16%.

**Lösung c)**

Für diese Aufgabe ermitteln wir zunächst die bedingten relativen Häufigkeiten. Beispielhaft berechnen wir die bedingte Verteilung für das Merkmal „Altersgruppe“. Als Bedingungen kommen hier nur die Ausprägungen des Merkmals „Kaufabsicht“, {ja, nein}, in Frage. Die einzelnen Wahrscheinlichkeiten werden mithilfe der Kontingenztabelle wie folgt berechnet:

$h_{Altersgruppe|ja} = \left\{\frac{0.0693}{0.5551} = 0.1248,\ \frac{0.1393}{0.5551} = 0.2509,\ \frac{0.3465}{0.5551} = 0.6542\right\}$

Es ergibt sich folgende Tabelle:

| $h_{i\|y_j}$ | Kaufabsicht | |
|---|---|---|
| Altersgruppe | Ja | Nein |
| unter 20 | 0.1248 | 0.5016 |
| 20 bis unter 60 | 0.2509 | 0.3069 |
| 60 und darüber | 0.6242 | 0.1916 |
| $\sum$ | 1.0000 | 1.0000 |

Statistische Unabhängigkeit zweier Merkmale ist dann gegeben, wenn die bedingten Verteilungen eines Merkmals für alle Kategorien des anderen Merkmals identisch und gleich der Randverteilung sind. Dies ist in dieser Aufgabe nicht der Fall, denn wir sehen, dass $h_{Altersgruppe|ja} = \{0.1248, 0.2509, 0.6242\} \neq h_{Altersgruppe|nein} = \{0.5016, 0.3069, 0.1916\} \neq h_{Altersgruppe} = \{0.2924, 0.2759, 0.4317\}$

Die Verteilung des Merkmals „Altersgruppe“ variiert also, je nachdem, ob man die Gruppe mit Kaufabsicht, ohne Kaufabsicht oder eine Altersgruppe gesamt betrachtet. Das gleiche Resultat erhalten wir, wenn wir das Merkmal „Kaufabsicht“ betrachten:

| $h_{j\|x_i}$ | Kaufabsicht | | |
|---|---|---|---|
| Altersgruppe | Ja | Nein | $\sum$ |
| Unter 20 | 0.2370 | 0.7630 | 1.0000 |
| 20 bis unter 60 | 0.5050 | 0.4950 | 1.0000 |
| 60 und darüber | 0.8026 | 0.1974 | 1.0000 |

Allerdings besitzt die Abhängigkeit nur mittlere Stärke (vgl. Lösung mit R, Berechnung von KK*).

**Lösung mit R**

```
> # a) Erstellen der Kontingenztabelle
> Altersgruppe <- c(rep("01: unter 20",422), rep("02: 20 bis unter 60",398),
+                   rep("03: 60 und darüber",623))
```

```
> Kaufabsicht <- c(rep("01: ja",100), rep("02: nein",322),
+                  rep("01: ja",201), rep("02: nein",197),
+                  rep("01: ja",500), rep("02: nein",123))

> table(Altersgruppe, Kaufabsicht)             # absolute Häufigkeiten
                     Kaufabsicht
Altersgruppe          01: ja 02: nein
  01: unter 20           100      322
  02: 20 bis unter 60    201      197
  03: 60 und darüber     500      123

> addmargins(table(Altersgruppe, Kaufabsicht)) # mit Randverteilungen
                     Kaufabsicht
Altersgruppe          01: ja 02: nein  Sum
  01: unter 20           100      322  422
  02: 20 bis unter 60    201      197  398
  03: 60 und darüber     500      123  623
  Sum                    801      642 1443

> options(digits = 3)                          # Begrenzung Stellen
> prop.table(table(Altersgruppe, Kaufabsicht)) # relative Häufigkeiten
                     Kaufabsicht
Altersgruppe          01: ja 02: nein
  01: unter 20        0.0693   0.2231
  02: 20 bis unter 60 0.1393   0.1365
  03: 60 und darüber  0.3465   0.0852

> addmargins(prop.table(table(Altersgruppe, Kaufabsicht))) # mit Randvert.
                     Kaufabsicht
Altersgruppe          01: ja 02: nein    Sum
  01: unter 20        0.0693   0.2231 0.2924
  02: 20 bis unter 60 0.1393   0.1365 0.2758
  03: 60 und darüber  0.3465   0.0852 0.4317
  Sum                 0.5551   0.4449 1.0000
>
> # c) Berechnung von KK* (vgl. Kap. 4)
> chisq.test(table(Altersgruppe, Kaufabsicht))

        Pearson's Chi-squared test

data:  table(Altersgruppe, Kaufabsicht)
X-squared = 300, df = 2, p-value <2e-16

> QK  <- 300; m <- 2; n <- 1443
> KK_star <- sqrt(QK*m/((QK + n)*(m - 1)))
> KK_star
[1] 0.587
# => Mittelstarke Abhängigkeit
```

**Aufgabe 2.3: HTWK Kalender**

*Beim Tag der offenen Tür an der HTWK Leipzig wurden Teilnehmer danach befragt, ob sie den neuen HTWK-Kalender kaufen würden. Bei der Befragung von 1415 Personen ergab sich folgende Kontingenztabelle für die Merkmale „Campus-Gruppe“ und „Kaufwahrscheinlichkeit“.*

| $n_{ij}$ | Kaufwahrscheinlichkeit | | | |
|---|---|---|---|---|
| Campus-Gruppe | gering | mittel | hoch | $\sum$ |
| Student | 197 | 388 | 320 | 905 |
| Mitarbeiter | 103 | 137 | 98 | 338 |
| Alumni | 20 | 18 | 18 | 56 |
| Anwohner | 13 | 58 | 45 | 116 |
| $\sum$ | 333 | 601 | 481 | 1415 |

*a) Welcher Anteil der Befragten sind Alumni?*
*b) Welcher Anteil der Befragten wird mit hoher Wahrscheinlichkeit kaufen?*
*c) Welcher Anteil der Befragten, die mit hoher Wahrscheinlichkeit kaufen, sind Alumni?*
*d) Welcher Anteil der Alumni wird mit hoher Wahrscheinlichkeit kaufen?*
*e) Ermitteln Sie die Randverteilung für das Merkmal Campus-Gruppe.*
*f) Ermitteln Sie die Verteilung für das Merkmal Campus-Gruppe unter der Bedingung, dass eine hohe Kaufwahrscheinlichkeit vorliegt.*
g) *Gibt es aus der Studie Evidenz dafür, dass man sich beim Verkauf des Kalenders auf eine bestimmte Gruppe konzentrieren sollte?*

**Lösung a)**
Der Anteil der Alumnis an den Befragten errechnet sich über die Randverteilung für das Merkmal Campus-Gruppe:
$\frac{56}{1415} = 0.0396 = 3.96\%$

**Lösung b)**
Der Anteil der Befragten, die mit hoher Wahrscheinlichkeit kaufen, errechnet sich über die Randverteilung für das Merkmal Kaufwahrscheinlichkeit:
$\frac{481}{1415} = 0.3399 = 33.99\%$

**Lösung c)**
Hier ist eine bedingte relative Häufigkeit gesucht. Die Bedingung ist „hohe Kaufwahrscheinlichkeit". Dazu betrachtet man nur die Spalte der hohen Kaufwahrscheinlichkeiten: $\frac{18}{481} = 0.0374 = 3.74\%$

**Lösung d)**
Auch hier liegt eine bedingte relative Häufigkeit vor. Die Bedingung ist „Alumni". Dazu betrachtet man nur die Zeile der Alumnis:
$\frac{18}{56} = 0.3214 = 32.14\ \%$

**Lösung e)**
Die Randverteilung für das Merkmal Campus-Gruppe berechnet sich über die Zahl der Beobachtungen für die Ausprägungen von Campus-Gruppe in Relation zum Stichprobenumfang:
Student: $\frac{905}{1415} = 0.6396 = 63.96\%$
Mitarbeiter: $\frac{338}{1415} = 0.2389 = 23.89\%$
Alumni: $\frac{56}{1415} = 0.0369 = 3.69\%$
Anwohner: $\frac{116}{1415} = 0.0820 = 8.20\%$

**Lösung f)**
Die Verteilung für das Merkmal Campus-Gruppe bei hoher Kaufwahrscheinlichkeit berechnet sich aus der Anzahl der Befragten der Campus-Gruppen bei hoher Kaufwahrscheinlichkeit in Relation zur Gesamtzahl an Beobachtungen mit hoher Kaufwahrscheinlichkeit:
Student: $\frac{320}{481} = 0.665 = 66.5\ \%$
Mitarbeiter: $\frac{98}{481} = 0.204 = 20.4\ \%$
Alumni: $\frac{18}{481} = 0.037 = 3.7\ \%$

Anwohner: $\frac{45}{481} = 0.094 = 9.4\,\%$

**Lösung g)**
Um die Frage beantworten zu können, bietet sich ein Vergleich der bedingten Verteilung von Campus-Gruppe bei hoher Kaufwahrscheinlichkeit aus Aufgabe f) und der (unbedingten) Randverteilung aus Aufgabe e) an. Die bedingte Verteilung für Campus-Gruppe ist praktisch gleich der Randverteilung (dies gilt auch für die Bedingungen mittlere und geringe Kaufwahrscheinlichkeit). Dies lässt auf statistische Unabhängigkeit schließen. Danach ist es weitgehend unerheblich, welcher Campus-Gruppe eine befragte Person angehört – die Bereitschaft den Kalender zu kaufen ist die gleiche. Das Potential, Käufer zu gewinnen, ist jedoch von Gruppe zu Gruppe unterschiedlich. Legt man die Daten der Stichprobe zugrunde, ist dieses bei den Studenten am größten, gefolgt von den Mitarbeitern. Darüber hinaus wäre zu klären, welche Grundgesamtheiten vorhanden sind.

**Lösung mit R**

```
> Gruppe <- c(rep("01: Student",905),rep("02: Mitarbeiter",338),
+             rep("03: Alumni",56),rep("04: Anwohner",116))
>
> W_keit <- c(rep("01: gering",197),rep("02: mittel",388),
+             rep("03: hoch",320),  rep("01: gering",103),
+             rep("02: mittel",137),rep("03: hoch",98),
+             rep("01: gering",20), rep("02: mittel",18),
+             rep("03: hoch",18),   rep("01: gering",13),
+             rep("02: mittel",58), rep("03: hoch",45))
>
> addmargins(table(Gruppe, W_keit))   # abs. Häufigkeiten mit Randvert.
                 W_keit
Gruppe            01: gering 02: mittel 03: hoch  Sum
  01: Student            197        388      320  905
  02: Mitarbeiter        103        137       98  338
  03: Alumni              20         18       18   56
  04: Anwohner            13         58       45  116
  Sum                    333        601      481 1415
> options(digits = 2)                           # Kontrolle der Ausgabe
> addmargins(prop.table(table(Gruppe, W_keit)))  # rel. Häufigkeiten mit Randvert.
                 W_keit
Gruppe            01: gering 02: mittel 03: hoch    Sum
  01: Student         0.1392     0.2742   0.2261 0.6396
  02: Mitarbeiter     0.0728     0.0968   0.0693 0.2389
  03: Alumni          0.0141     0.0127   0.0127 0.0396
  04: Anwohner        0.0092     0.0410   0.0318 0.0820
  Sum                 0.2353     0.4247   0.3399 1.0000

> prop.table(table(Gruppe, W_keit), 2)      # bedingte rel. Verteilung für Gruppe
                 W_keit
Gruppe            01: gering 02: mittel 03: hoch
  01: Student          0.592      0.646    0.665
  02: Mitarbeiter      0.309      0.228    0.204
  03: Alumni           0.060      0.030    0.037
  04: Anwohner         0.039      0.097    0.094
```

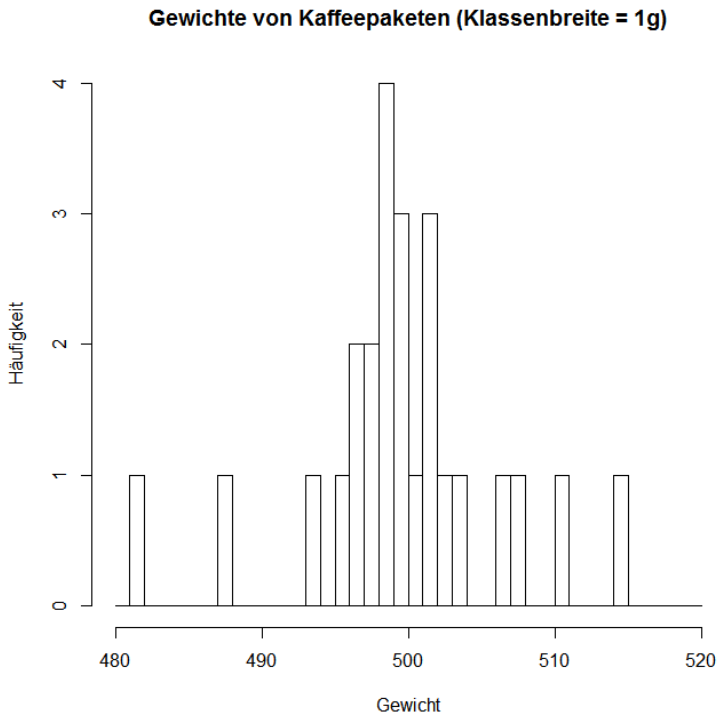
## Kapitel 3: Darstellung und Beschreibung quantitativer Daten

**Aufgabe 3.1: Kaffee**

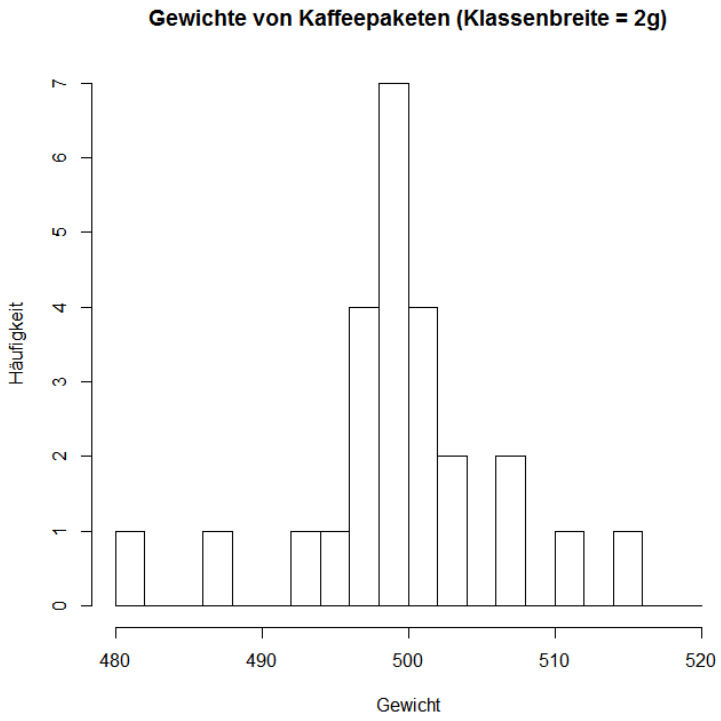
*Beim Nachwiegen von 25 verpackten Ein-Pfund-Paketen Kaffee ergaben sich folgende Werte (in g) für die Variable Gewicht:*
*494, 497, 497, 488, 482, 498, 498, 499, 503, 511, 499, 504, 508, 496, 502, 500, 499, 500, 507, 502, 500, 499, 515, 501, 502.*

*a) Zeichnen Sie ein Histogramm mit (i) der Klassenbreite 1g und (ii) der Klassenbreite 2g.*
*b) Berechnen Sie den Mittelwert und den Median der Variablen Gewicht.*

**Lösung a)**
(i)




(ii)



Beim Histogramm ist zu beachten, dass die obere Klassengrenze in das Intervall eingeschlossen ist. Es wird also für das Merkmal $X$ die absolute Häufigkeit für $(a, b]$ bzw. $a < X \leq b$ ausgegeben. Zum Beispiel gibt es nur mit 482 eine Beobachtung im Intervall (480,482].

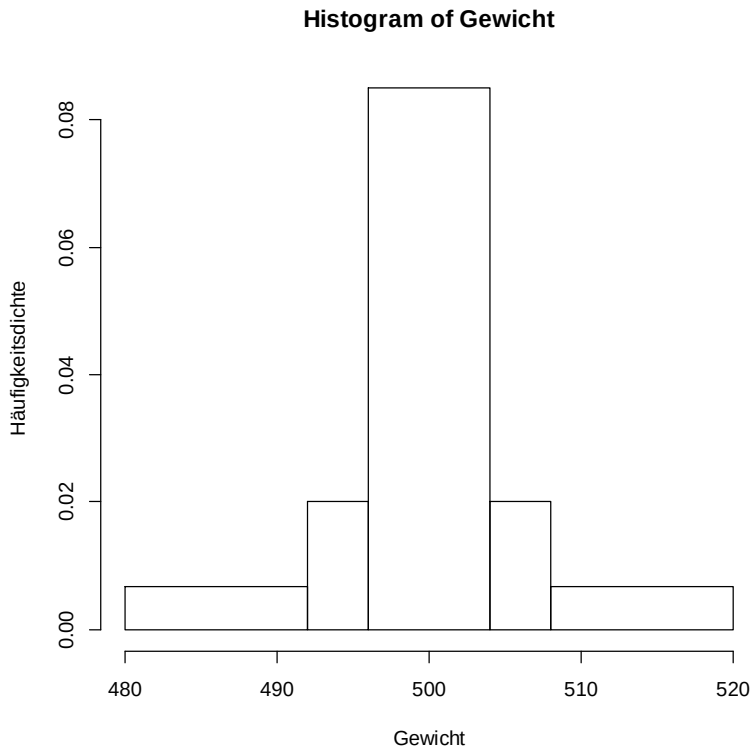
**Lösung b)**

$$\bar{x} = \frac{\sum_1^{25} x_{Gewicht}}{n} = \frac{12501}{25} = 500.04\ g$$

Man summiert die 25 Werte auf und teilt sie durch die Anzahl der Pakete $n$ ($n = 25$).
Das mittlere Gewicht einer Kaffeepackung beträgt 500.04g.

$$n\ ungerade => x_{[0.5]} = x_{\frac{n+1}{2}} = x_{\frac{25+1}{2}} = x_{13}$$

Der Median wird hier durch den 13. von 25 Werten dargestellt in der Reihenfolge vom kleinsten zum größten Wert. Er beträgt 500g.
Da die Verteilung annähernd symmetrisch ist, sind Median und arithmetisches Mittel fast identisch.
In dieser Aufgabe wurde bisher nur der Spezialfall eines Histogramms, nämlich gleiche Klassenbreiten betrachtet. Angenommen nun, wir wollen ein Histogramm mit folgenden Klassenbreiten zeichnen: (480, 492], (492, 496], (496, 504], (504, 508], (508, 520].
In diesem Fall ergäbe sich dieses Histogramm




Die Berechnung der Werte für die Häufigkeitsdichte kann aus der R-Lösung nachvollzogen werden.
Man beachte: Bei absoluten Häufigkeiten je Klasse ist ein Vergleich mit den Häufigkeiten anderer Klassen nur dann möglich, wenn gleiche Klassenbreiten vorliegen. Sind die Klassen nicht gleich breit, wird im Histogramm die Häufigkeitsdichte dargestellt und ist entsprechend zu interpretieren.

## Lösung mit R

```
> # a) (i)
> # Einlesen des Datenvektors
> Gewicht <- c (494, 497, 497, 488, 482, 498, 498, 499,
+               503, 511, 499, 504, 508, 496, 502, 500, 499,
+               500, 507, 502, 500, 499, 515, 501, 502)
> # Zusammenfassung der wichtigsten Werte für die Variable Gewicht.
> summary(Gewicht)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    482     498     500     500     502     515
>
> # Histogramm mit der Klassenbreite = 1g (i)
> hist(Gewicht, breaks = seq(480, 520, 1), xlab = "Gewicht in g",
+ main = "Kaffeepakete (Klassenbreite = 1g)", ylab = "Häufigkeit")
>

> # a) (ii)
> # Histogramm mit Klassenbreite = 2g (ii)
> hist(Gewicht, breaks = seq(480, 520, 2), xlab = "Gewicht in g",
+ main = "Kaffeepakete (Klassenbreite = 2g)", ylab = "Häufigkeit")
>
> # Um die beiden Histogramme gegenüberzustellen und vergleichen zu können
> # benutzen wir den Befehl „par“.
> # Alle Zeilen müssen gemeinsam ausgeführt werden.
> par(mfrow = c(1,2))              # eine Zeile, zwei Spalten
> hist(Gewicht, breaks = seq(480, 520, 1), xlab = "Gewicht in g",
+ main = "Klassenbreite = 1g", ylab = "Häufigkeit")
> hist(Gewicht, breaks = seq(480, 520, 2), xlab = "Gewicht in g",
+ main = "Klassenbreite = 2g", ylab = "Häufigkeit")
> par(mfrow = c(1,1))              # Standardeinstellung
>
> # b)
> # Der Median und das arithmetische Mittel der Variablen können
> # entweder gemeinsam mit der Funktion summary() ausgegeben werden oder
> # einzeln mit mean() oder median().
> summary(Gewicht)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    482     498     500     500     502     515
> mean(Gewicht)
[1] 500
> median(Gewicht)
[1] 500

> # Ergänzung zur Aufgabe
> # Histogramm mit unterschiedlichen Klassenbreiten
> hist(Gewicht, breaks = c(480, 492, 496, 504, 508, 520),
+ xlab = "Gewicht in g", ylab = "Häufigkeitsdichte")
>
> # Ausgabe absolute Klassenhäufigkeit n_K, relative Klassenhäufigkeit h_k
> # und der Häufigkeitsdichte density
> n <- length(Gewicht)
> Gewicht_int <- cut(Gewicht, breaks = c(480, 492, 496, 504, 508, 520))
> n_K <- table(Gewicht_int); n_K
Gewicht_int
(480,492] (492,496] (496,504] (504,508] (508,520]
        2         2        17         2         2
>
> h_K <- n_K/n
> delta <- c(12, 4, 8, 4, 12)
> density <- h_K/delta
> data.frame(n_K)     # Ausgabe absolute Klassenhäufigkeit
  Gewicht_int Freq
1   (480,492]    2
2   (492,496]    2
3   (496,504]   17
4   (504,508]    2
5   (508,520]    2
> data.frame(h_K)     # Ausgabe relative Klassenhäufigkeit
```

```
  Gewicht_int Freq
1   (480,492] 0.08
2   (492,496] 0.08
3   (496,504] 0.68
4   (504,508] 0.08
5   (508,520] 0.08
> data.frame(density) # Ausgabe Häufigkeitsdichte
  Gewicht_int        Freq
1   (480,492] 0.006666667
2   (492,496] 0.020000000
3   (496,504] 0.085000000
4   (504,508] 0.020000000
5   (508,520] 0.006666667
>
> # Flächen des Histogramms sind = 1
> 12*0.00666667 + 4*0.02 + 8*0.085 + 4*0.02 + 12*0.00666667
[1] 1
```

**Aufgabe 3.2: Anleihen der öffentlichen Hand**

*Der Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung veröffentlicht die jährliche Verzinsung von Anleihen der öffentlichen Hand in der Periode 2002 bis 2011.*

| 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|
| 4.6 | 3.8 | 3.7 | 3.2 | 3.7 | 4.3 | 4.0 | 3.1 | 2.4 | 2.4 |

*Berechnen Sie die durchschnittliche jährliche Verzinsung der Anleihen von 2002 bis 2011.*

**Lösung**

Die Merkmalsausprägungen (hier die Wachstumsfaktoren) sind multiplikativ miteinander verknüpft. Daher erfolgt die Berechnung durch das geometrische Mittel.

$G_X = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

$G_X = \sqrt[10]{1.046 * 1.038 * 1.037 * 1.032 * 1.037 * 1.043 * 1.040 * 1.031 * 1.024 * 1.024}$

$G_X = 1.0352$

Die durchschnittliche jährliche Verzinsung beträgt daher 1.0352 – 1 = 0.0352 oder 3.52%. Man beachte, dass die Wachstumfaktoren in der Formel für $G_X$ eingesetzt werden müssen und nicht die Wachstumsraten.

**Lösung mit R**

```
> (1.046*1.038*1.037*1.032*1.037*1.043*1.040*1.031*1.024*1.024)^(1/10) - 1
[1] 0.0352
```

**Aufgabe 3.3: Umsatz**

*Der Umsatz eines Unternehmens entwickelte sich in den Jahren 2011 bis 2014 jeweils mit folgenden jährlichen Wachstumsraten*

| Jahr | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|
| Umsatzänderung in % | 8 | 15 | -4 | 12 |

*Berechnen Sie sowohl das geometrische als auch das arithmetische Mittel der Wachstumsraten. Vergleich und Interpretation.*

**Lösung**

Der durchschnittliche Wachstumsfaktor, berechnet mit dem arithmetischen Mittel, ist

$\overline{r_t} = \frac{1.08+1.15+0.96+1.12}{4} = 1.0775$

Mit dem geometrischen Mittel ergibt sich dagegen

$G_{r_t} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
$G_{r_t} = \sqrt[4]{1.08 * 1.15 * 0.96 * 1.12}$
$G_{r_t} = 1.075$
Das arithmetische Mittel (durchschnittliche Wachstumsrate 7.75%) würde in diesem Fall zu höheren Werten/Umsätzen führen als das geometrische Mittel (7.5%). Dies ist kein Zufall. Es gilt, dass das geometrische Mittel für jede Variable mit nur positiven und sich unterscheidenden Werten stets kleiner ist als das arithmetische Mittel.

**Lösung mit R**

```
> # arithmisches Mittel der Wachstumsraten
> (1.08 + 1.15 + 0.96 + 1.12)/4 - 1
[1] 0.0775
>
> # geometrisches Mittel der Wachstumsraten
> (1.08*1.15*0.96*1.12)^(1/4) - 1
[1] 0.075
```

**Aufgabe 3.4: Theorie**

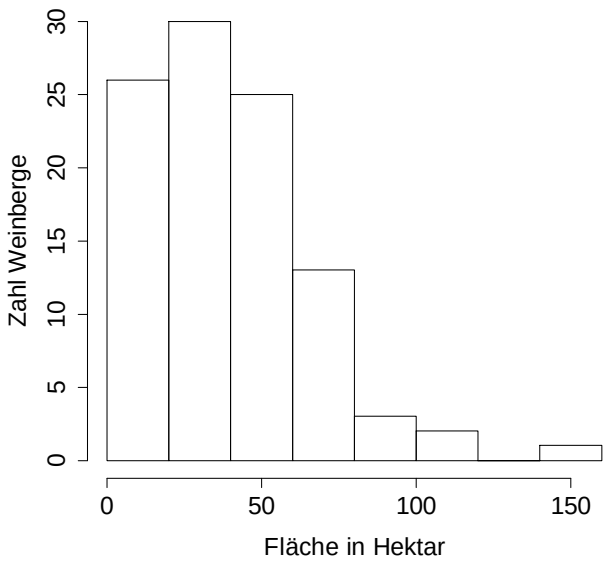
*a) Beweisen Sie den folgenden Satz: Keine andere Zahl hat eine kleinere Summe quadrierter Abweichungen von vorgegebenen Ausgangsdaten als deren arithmetisches Mittel.*
*Hinweis: Suchen Sie die Zahl d, die die Summe der quadrierten Abweichungen von den Ausgangsdaten minimiert.*
*b) Zeigen Sie für $n = 2$ (z.B. für die Variablen $x_1$ und $x_2$ mit $x_1 > 0$, $x_2 > 0$), dass folgender Zusammenhang gilt: $G_x \leq \bar{x}$.*

**Lösung a)**
Wir suchen die Zahl d, die die Summe der quadrierten Abweichungen von den Ausgangsdaten $x_i$ minimiert:
$\min_d \sum(x_i - d)^2$
$\frac{\partial \sum(x_i-d)^2}{\partial d} = 2\sum(x_i - d)(-1) = 0$
$\sum x_i - \sum d = 0 \;\Rightarrow\; \sum x_i = nd \;\Rightarrow\; d = \frac{1}{n}\sum x_i$
Wie man sieht: Die Zahl *d* ist das arithmetische Mittel.

**Lösung b)**
Das geometrische Mittel ist für jede Variable mit nur positiven Werten stets kleiner als das arithmetische Mittel, es sei denn, alle Werte sind gleich. Wir stellen diese Ungleichheit für die Werte $x_1$ und $x_2$ auf:
$\sqrt{x_1 x_2} \leq \frac{x_1+x_2}{2} \Longrightarrow 2\sqrt{x_1 x_2} \leq x_1 + x_2 \Longrightarrow 4x_1x_2 \leq (x_1+x_2)^2$
$(x_1+x_2)^2 - 4x_1x_2 \geq 0 \Longrightarrow x_1^2 + 2x_1x_2 + x_2^2 - 4x_1x_2 \geq 0$
$x_1^2 - 2x_1x_2 + x_2^2 \geq 0 \Longrightarrow (x_1 - x_2)^2 \geq 0$
Der Term $(x_1 - x_2)^2$ ist nie negativ, damit ist die Behauptung für beliebige Werte $x_1 > 0$, $x_2 > 0$ bewiesen.

**Aufgabe 3.5: Weinanbau**

***Weinanbau.*** *Das abgebildete Histogramm zeigt die Fläche von 100 Weinbergen in einer spanischen Weinanbauprovinz.*




*a) Beschreiben Sie die Verteilung der Fläche.*
*b) Welches Lagemaß würden Sie zur Beschreibung der Fläche heranziehen? Begründung.*

**Lösung a)**
Das Histogramm zeigt die Verteilung der Weinberge nach Größenklassen der Breite 20ha. Die Verteilung kann als linkssteil bzw. rechtsschief beschrieben werden, da die Klassen mit den relativ kleinen Werten stärker besetzt sind als die Klassen mit relativ großen Werten.

**Lösung b)**
Der Median $x_{[0.5]}$ ist das besser geeignete Lagemaß. Das arithmetische Mittel $\bar{x}$ würde im Falle einer schiefen Verteilung zu stark verzerrt werden, der Median erweist sich diesen gegenüber als weniger anfällig. Hier ist $\bar{x} > x_{[0.5]}$.

**Aufgabe 3.6: Varianz**

*Beweisen Sie die Formel für die Varianzberechnung für Daten aus einer Häufigkeits-verteilung:*
$s_X^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \bar{x}^2$.

**Lösung**
Herleitung:
$s_X^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i \; - \frac{1}{n} 2\bar{x} \sum_{i=1}^{k} x_i \, n_i + \frac{1}{n}\sum_{i=1}^{k} n_i \, \bar{x}^2$
$= \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i \; - 2\bar{x} \sum_{i=1}^{k} x_i \, h_i + \bar{x}^2 \frac{1}{n}\sum_{i=1}^{k} n_i$
$= \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i \; - 2\bar{x}^2 + \bar{x}^2 \sum_{i=1}^{k} h_i \; = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i \; - 2\bar{x}^2 + \bar{x}^2 1$
$= \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i \; - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i \; - \bar{x}^2$
Man beachte, dass $\frac{1}{n}\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} h_i = 1$ gilt.

**Aufgabe 3.7: Training**

*Eine Studie untersucht den Effekt regelmäßigen Trainings auf die Herzfrequenz von Mitarbeitern einer Firma. An der Studie nehmen 20 Mitarbeiter im Alter von 30 bis 40 Jahre teil, wobei sich die Hälfte als „trainiert" bezeichnet und die übrigen als „untrainiert". Nach einer Belastung (15-minütiger lockerer Lauf) wird die Herzfrequenz (in Anzahl Herzschläge pro Minute) der Mitarbeiter gemessen. Es ergeben sich die folgenden Daten:*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| trainiert: | 120 | 134 | 106 | 157 | 144 | 133 | 116 | 128 | 140 | 123 |
| untrainiert: | 135 | 140 | 136 | 147 | 154 | 153 | 136 | 138 | 165 | 155 |

*Vergleichen Sie beide Gruppen mit Hilfe zweier Boxplots. Interpretieren Sie die Angaben aus den Boxplots.*

**Lösung**

Die fünf Zahlen, die für die Boxplots nötig sind, sind $min$, $Q_1$, $x_{Med}$, $Q_3$ und $max$. Wir ermitteln für „trainiert" die Quartile folgendermaßen: Die geordnete Reihe ist
106, 116, 120, 123, 128, 133, 134, 140, 144, 157
Der Median ist $x_{Med} = \frac{128+133}{2} = 130.5$
Dieser Wert teilt die oben und unteren 50% in zwei Hälften. Für jede Hälfte ist wieder der Median zu ermitteln, so ergibt sich $Q_1 = 120$ und $Q_3 = 140$.
Wir erhalten also die Werte:
*„trainiert":* $min = 106, Q_1 = 120, x_{Med} = 130.5, Q_3 = 140, \max = 157$
*„untrainiert":* $min = 135, Q_1 = 136, x_{Med} = 143.5, Q_3 = 154, \max = 165.$

Die Erstellung der Boxplots erfolgt in fünf Schritten:

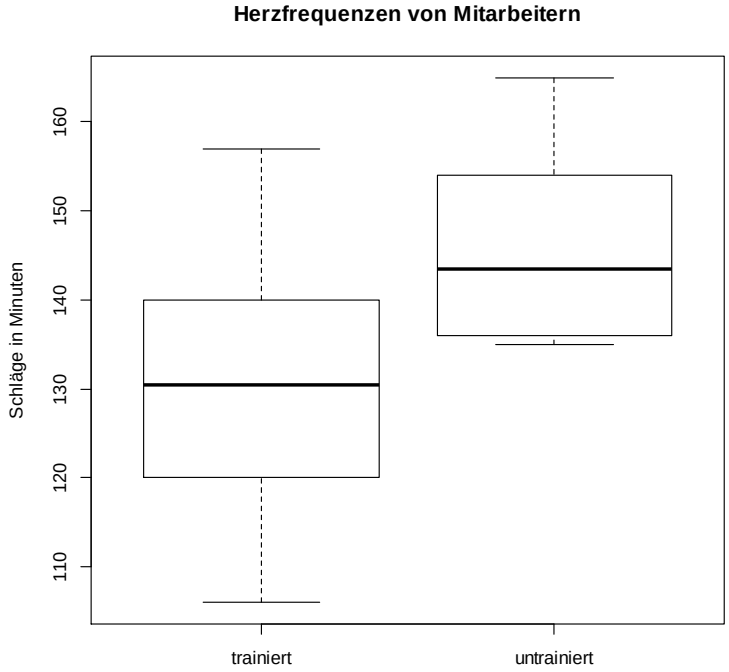
(i) Das Diagramm eines Boxplots hat nur eine Ausrichtung – entweder vertikal oder horizontal. In dieser Ausrichtung wird zunächst eine Achse über den gesamten Datenbereich eingezeichnet.

(ii) Es wird ein Viereck („Box") eingezeichnet, welches sich vom 1. Quartil bis zum 3. Quartil ausdehnt und vom Median durch eine Linie geteilt wird.

(iii) Zwei Begrenzungen werden markiert (aber *nicht* im späteren Boxplot eingezeichnet); eine Begrenzung liegt $1.5 \cdot IQA$ oberhalb des 3. Quartils, die andere liegt $1.5 \cdot IQA$ unterhalb des 1. Quartils.

(iv) Kleine Begrenzungslinien (sog. Whisker), die vom Ende der Box bis zum größten bzw. kleinsten Wert *innerhalb* der Begrenzung gehen, werden dann eingezeichnet. Datenpunkte außerhalb der Begrenzung werden *nicht* mit den Whiskern verbunden.

(v) Alle Werte, die jenseits der Whisker liegen, werden als *Ausreißer* individuell gekennzeichnet, z.B. durch einen Punkt.




Der Median der untrainierten Gruppe liegt deutlich höher als der Median der trainierten Gruppe. Die mittleren 50% der trainierten (untrainierten) Mitarbeiter haben eine Herzfrequenz zwischen 120 und 140 (136 und 154) Herzschlägen pro Minute.
Die Verteilung der Frequenzen der trainierten Mitarbeiter ist relativ symmetrisch, wohingegen die Frequenzen der untrainierten Gruppe ungleichmäßiger verteilt ist (rechtsschiefe Verteilung). Für Trainierte ist die gesamte Streuung größer als für Untrainierte (größere Distanz zwischen den Whisker).
Interpretation: Bei manchen wirkt das Training, bei anderen wirkt es nicht. Im Mittel (Median) hat Training aber einen senkenden Effekt auf die Herzfrequenz.

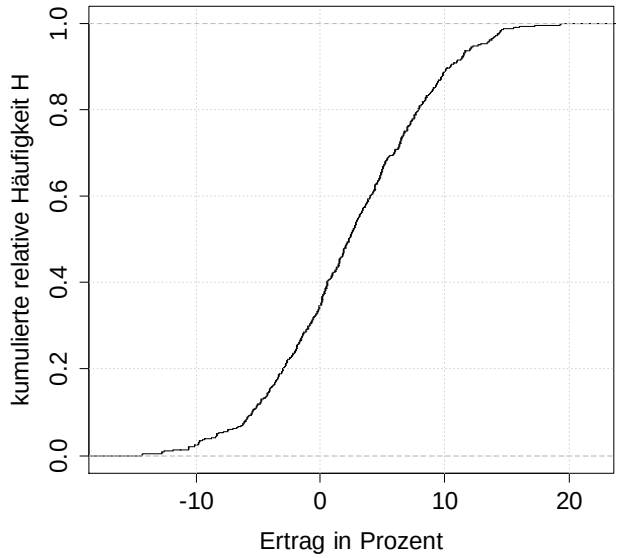
**Lösung mit R**

```
> # Eingabe der Werte für trainierte und untrainierte Personen.
> trainiert   <- c(120, 134, 106, 157, 144, 133, 116, 128, 140, 123)
> untrainiert <- c(135, 140, 136, 147, 154, 153, 136, 138, 165, 155)
>
> boxplot(trainiert, untrainiert, names = c("trainiert", "untrainiert"),
+ ylab = "Schläge in Minuten", main = "Herzfrequenzen von Mitarbeitern")
>
> # Die wesentlichen Angaben für den Boxplot
> boxplot.stats(trainiert, do.conf = FALSE, do.out = TRUE)
$stats
[1] 106.0 120.0 130.5 140.0 157.0

$n
[1] 10

$conf
NULL

$out
numeric(0)

> boxplot.stats(untrainiert, do.conf = FALSE, do.out = TRUE)
$stats
[1] 135.0 136.0 143.5 154.0 165.0
```

```
$n
[1] 10

$conf
NULL

$out
numeric(0)

> quantile(trainiert, type = 2)   # Quantile für den Boxplot
   0%   25%   50%   75%  100%
106.0 120.0 130.5 140.0 157.0
> quantile(untrainiert, type = 2) # Quantile für den Boxplot
   0%   25%   50%   75%  100%
135.0 136.0 143.5 154.0 165.0
>
> # Bestimmung der Quantile über die Verteilungsfunktion F
> plot.ecdf(trainiert)
> abline(h = c(0.25, 0.5, 0.75), lty = 2)
> abline(v = c(120.0, 130.5, 140.0), lty = 2)
>
> plot.ecdf(untrainiert)
> abline(h = c(0.25, 0.5, 0.75), lty = 2)
> abline(v = c(136.0, 143.5, 154.0), lty = 2)
```

**Aufgabe 3.8 Verzinsung**

*Abgebildet ist die empirische Verteilungsfunktion H für den monatlichen Ertrag von ausgewählten Investmentfonds für die Periode 1975 bis 2010.*




*Schätzen und interpretieren Sie a) den Median; b) das 1. und 3. Quartil; c) die Spannweite und d) den IQA.*

**Lösung a)**

Der Median $x_{[0.5]} = x_{Med}$, bei dem mindestens 50% kleiner oder gleich $x_{Med}$ sind und mindestens 50% größer oder gleich $x_{Med}$ sind, beträgt rund 3%.

**Lösung b)**

Der Wert für das erste Quartil $Q_1$, bei dem mindestens 25% kleiner oder gleich $Q_1$ sind, beträgt ca. -1%. Der Wert für das dritte Quartil $Q_3$, bei dem mindestens 25% größer oder gleich $Q_3$ sind, beträgt ca. 8%.

**Lösung c)**
Die Spannweite beträgt ca. 35% (20 - (-15) = 35). Dies ist die Differenz zwischen dem größten und dem kleinsten Wert.

**Lösung d)**
Der Interquartilsabstand beträgt ca. 9% ($Q_1 - Q_3$). In einem Intervall dieser Breite liegen die mittleren 50% der Beobachtungen.

# Kapitel 4: Assoziation und Korrelation

**Aufgabe 4.1: Korrigierter Kontingenzkoeffizient**

*Berechnen Sie für die Kontingenztabellen in Bsp. 2.2 und Aufgabe 2.1:*
*a) die Quadratische Kontingenz QK und den korrigierten Kontingenzkoeffizienten KK*.*
*b) Interpretieren Sie KK*.*

**Lösung für Bsp. 2.2)**
Die Kontingenztabelle mit absoluten und relativen Häufigkeiten sind im Folgenden dargestellt:

| $n_{ij}$ | Geschlecht | | |
|---|---|---|---|
| Journal | Frau | Mann | $\sum$ |
| Cosmopolitan | 45 | 5 | 50 |
| Economist | 5 | 10 | 15 |
| Sports Illustrated | 10 | 25 | 35 |
| $\sum$ | 60 | 40 | 100 |

| $h_{ij}$ | Geschlecht | | |
|---|---|---|---|
| Journal | Frau | Mann | $\sum$ |
| Cosmopolitan | 0.45 | 0.05 | 0.50 |
| Economist | 0.05 | 0.10 | 0.15 |
| Sports Illustrated | 0.10 | 0.25 | 0.35 |
| $\sum$ | 0.60 | 0.40 | 1.00 |

Für die Berechnung der quadratischen Kontingenz ermitteln wir zunächst die erwarteten absoluten Werte bei statistischer Unabhängigkeit. Hierzu wird die absolute Zeilenhäufigkeit mit der absoluten Spaltenhäufigkeit multipliziert und durch die Gesamtzahl dividiert, d.h. $E_{ij} = \frac{n_{i\circ} n_{\circ j}}{n}$.

$Cosmopolitan \cap Frau = 50 * 60/100 = 30$
$Cosmopolitan \cap Mann = 50 * 40/100 = 20$
$Economist \cap Frau = 15 * 60/100 = 9$
$Economist \cap Mann = 15 * 40/100 = 6$
$Sports\ Illustrated \cap Frau = 15 * 60/100 = 21$
$Sports\ Illustrated \cap Mann = 15 * 40/100 = 14$
Es entsteht folgende Tabelle:

| $E_{ij}$ | Geschlecht | | |
|---|---|---|---|
| Journal | Frau | Mann | $\sum$ |
| Cosmopolitan | 30 | 20 | 50 |
| Economist | 9 | 6 | 15 |
| Sports Illustrated | 21 | 14 | 35 |
| $\sum$ | 60 | 40 | 100 |

Nun können wir die Quadratische Kontingenz $QK$ ermitteln mithilfe folgender Formel:

$QK = \frac{\sum(n_{ij} - E_{ij})^2}{E_{ij}}$

| Zeile $i$,Spalte $j$ | $n_{ij}$ | $E_{ij}$ | $(n_{ij} - E_{ij})^2/E_{ij}$ |
|---|---|---|---|
| 1,1 | 45 | 30 | 7.50 |
| 1,2 | 5 | 20 | 11.25 |
| 2,1 | 5 | 9 | 1.78 |
| 2,2 | 10 | 6 | 2.67 |
| 3,1 | 10 | 21 | 5.76 |
| 3,2 | 25 | 14 | 8.64 |
| | | $QK$ = | 37.60 |

Zur Berechnung von $KK^*$ benötigen wir noch $m$, die kleinere Zahl der Zeilen- und Spaltenzahl in der Kontingenztabelle und $n$, die Anzahl der Beobachtungen. Im Beispiel haben wir drei Zeilen und zwei Spalten, d.h. $m = 2$. Damit lässt sich der korrigierte Kontingenzkoeffizient berechnen mit

$$KK^* = \sqrt{\frac{QK \cdot m}{(QK+n)(m-1)}} = \sqrt{\frac{37.60 \cdot 2}{(37.60+100)(2-1)}} = 0.74$$

**Lösung b)**

KK* gibt ein auf das Intervall [0,1] normiertes Maß für die Stärke des Zusammenhangs von zwei kategorial skalierten Merkmalen an. Im vorliegenden Fall kann man von einem relativ starken Zusammenhang zwischen den Merkmalen Geschlecht und Zeitschriftenpräferenz ausgehen. Aus KK* lässt sich im Allgemeinen keine Aussage über die Richtung des Zusammenhangs sagen – in diesem Fall aber schon, denn wir können davon ausgehen, dass das Geschlecht die Zeitschriftenpräferenz beeinflusst und nicht umgekehrt.

**Lösung mit R**

```
> # Einlesen des Datensatzes von Bsp. 2.2
> # Kontingenztabelle erstellen:
> # Daten aus Bsp. 2.1 sind identisch zu Bsp. 2.2
> data <- read.csv("Bsp._2.1.csv") # Einlesen des Datensatzes
> attach(data)                # Objekt "data" an den Suchpfad binden
> addmargins(prop.table(table(Journal, Geschlecht)))
                    Geschlecht
Journal              Frau Mann  Sum
  Cosmopolitan       0.45 0.05 0.50
  Economist          0.05 0.10 0.15
  Sports Illustrated 0.10 0.25 0.35
  Sum                0.60 0.40 1.00

> # Quadratische Kontingenz QK
> # Alternative 1:
> # Berechnung der erwarteten Werte bei statistischer Unabhängigkeit
> EXP <- chisq.test(table(Journal, Geschlecht)) # hier absolut
> EXP$expected
                    Geschlecht
Journal              Frau Mann
  Cosmopolitan         30   20
  Economist             9    6
  Sports Illustrated   21   14
>
> RES <- EXP$residuals # => (nij – Eij)/sqrt(Eij)
> RES
                    Geschlecht
Journal                   Frau      Mann
  Cosmopolitan        2.738613 -3.354102
  Economist          -1.333333  1.632993
  Sports Illustrated -2.400397  2.939874
>
> QK <- sum(RES^2) # ergibt QK
```

```
> QK
[1] 37.59921
>
> # Alternative 2:
> # Mehr zum Unabhängigkeitstest in Kap. 12
> chisq.test(table(Journal, Geschlecht))

        Pearson's Chi-squared test

data:  table(Journal, Geschlecht)
X-squared = 37.599, df = 2, p-value = 6.846e-09

> # Berechnung Kontingenzkoeffizient KK*
> m <- 2            # => Zeilenzahl (Gender) < Spaltenzahl (Journal) => m = 2
> n <- 100      # Anzahl der Beobachtungen
> QK <- 37.60# Vektor mit QK
>
> KK_star <- sqrt((QK / (QK + n)) * (m / (m - 1)))
> KK_star
[1] 0.7392642
```

## Lösung für Aufgabe 2.1)

| # | Getränk | Frau | # | Getränk | Frau |
|---|---|---|---|---|---|
| 1 | Bier | 0 | 14 | Mix | 1 |
| 2 | Mix | 0 | 15 | Mix | 1 |
| 3 | Wein | 0 | 16 | Bier | 1 |
| 4 | Bier | 0 | 17 | Mix | 1 |
| 5 | Bier | 0 | 18 | Mix | 1 |
| 6 | alkfrei | 0 | 19 | alkfrei | 1 |
| 7 | Bier | 0 | 20 | alkfrei | 1 |
| 8 | Bier | 0 | 21 | Mix | 1 |
| 9 | Bier | 0 | 22 | alkfrei | 1 |
| 10 | Wein | 0 | 23 | Wein | 1 |
| 11 | Wein | 1 | 24 | Wein | 1 |
| 12 | Wein | 1 | 25 | alkfrei | 1 |
| 13 | Wein | 1 | | | |

**Tabelle aus 2.1**

## Lösung a)

Zunächst erstellen wir die Kontingenztabelle mit den absoluten Häufigkeiten:

| $n_{ij}$ | Frau | | |
|---|---|---|---|
| Getränk | 0(Mann) | 1(Frau) | ∑ |
| Alkfrei | 1 | 4 | 5 |
| Bier | 6 | 1 | 7 |
| Mix | 1 | 5 | 6 |
| Wein | 2 | 5 | 7 |
| ∑ | 10 | 15 | 25 |

Nun berechnen wir wie im Bsp. 2.2 die Erwartungswerte bei statistischer Unabhängigkeit mithilfe der Randverteilungen:

$$alkfrei \cap Mann = 10 * \frac{5}{25} = 2.0$$

$$alkfrei \cap Frau = 15 * \frac{5}{25} = 3.0$$

$$Bier \cap Mann = 10 * \frac{7}{25} = 2.8$$

$$Bier \cap Frau = 15 * \frac{7}{25} = 4.2$$

$$Mix \cap Mann = 10 * \frac{6}{25} = 2.4$$

$Mix \cap Frau = 15 * \frac{6}{25} = 3.6$

$Wein \cap Mann = 10 * \frac{7}{25} = 2.8$

$Wein \cap Frau = 15 * \frac{7}{25} = 4.2$

Es entsteht folgende Tabelle:

| $E_{ij}$ | Frau | | |
|---|---|---|---|
| Getränk | 0(Mann) | 1(Frau) | $\sum$ |
| Alkfrei | 2.0 | 3.0 | 5.0 |
| Bier | 2.8 | 4.2 | 7.0 |
| Mix | 2.4 | 3.6 | 6.0 |
| Wein | 2.8 | 4.2 | 7.0 |
| $\sum$ | 10.0 | 15.0 | 25.0 |

Nun können wir die Quadratische Kontingenz ermitteln mithilfe folgender Formel:
$QK = \sum (n_{ij} - E_{ij})^2 / E_{ij}$

| Zeile $i$,Spalte $j$ | $n_{ij}$ | $E_{ij}$ | $(n_{ij} - E_{ij})^2 / E_{ij}$ |
|---|---|---|---|
| 1,1 | 1 | 2.0 | 0.5000 |
| 1,2 | 4 | 3.0 | 0.3333 |
| 2,1 | 6 | 2.8 | 3.6571 |
| 2,2 | 1 | 4.2 | 2.4381 |
| 3,1 | 1 | 2.4 | 0.8167 |
| 3,2 | 5 | 3.6 | 0.5444 |
| 4,1 | 2 | 2.8 | 0.2286 |
| 4,2 | 5 | 4.2 | 0.1524 |
| | | $QK =$ | 8.6706 |

Wie in der vorherigen Aufgabe benötigen wir zur Berechnung von $KK^*$ noch $m$, die kleinere Zahl der Zeilen- und Spaltenzahl in der Kontingenztabelle und $n$, die Anzahl der Beobachtungen.

$KK^* = \sqrt{\frac{QK \cdot m}{(QK+n)(m-1)}} = \sqrt{\frac{8.6706 \cdot 2}{(8.6706+25)(2-1)}} = 0.7177$

Somit beträgt die quadratische Kontingenz 8.6706 und der korrigierte Kontingenzkoeffizient 0.7177.

**Lösung b)**

Im vorliegenden Fall liegt ein relativ starker Zusammenhang zwischen den nominal skalierten Merkmalen Getränk und Frau vor. Das heißt, dass Frauen und Männer deutlich unterschiedliche Präferenzen für Getränke besitzen. So ist bei Frauen Wein relativ beliebter, Männer mögen hingegen Bier mehr als Frauen.

**Lösung mit R**

```
> # a) Kontingenztabelle erstellen:
> Getränk <- c("Bier", "Mix", "Wein", "Bier", "Bier", "alkfrei", "Bier",
+              "Bier", "Bier", "Wein", "Wein", "Wein", "Wein", "Mix", "Mix",
+              "Bier",   "Mix", "Mix", "alkfrei", "alkfrei", "Mix", "alkfrei",
+              "Wein", "Wein", "alkfrei")
> Frau    <- c(0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1)
```

```
> addmargins(table(Getränk, Frau))
         Frau
Getränk    0  1 Sum
  alkfrei  1  4   5
  Bier     6  1   7
  Mix      1  5   6
  Wein     2  5   7
  Sum     10 15  25
>
> # Quadratische Kontingenz QK
> # Alternative 1:
> # Berechnung der erwarteten Werte bei statistischer Unabhängigkeit
> EXP <- chisq.test(table(Getränk, Frau))
Warning message:
In chisq.test(table(Getränk, Frau)) :
  Chi-squared approximation may be incorrect

> EXP$expected
         Frau
Getränk     0   1
  alkfrei 2.0 3.0
  Bier    2.8 4.2
  Mix     2.4 3.6
  Wein    2.8 4.2
>
> RES <- EXP$residuals # => (nij – Eij)/sqrt(Eij)
> RES
         Frau
Getränk            0          1
  alkfrei -0.7071068  0.5773503
  Bier     1.9123658 -1.5614401
  Mix     -0.9036961  0.7378648
  Wein    -0.4780914  0.3903600
>
> QK <- sum(RES^2) # ergibt QK
> QK
[1] 8.670635
>
> # Alternative 2:
> # Mehr zum Test auf Unabhängigkeit in Kap. 12
> chisq.test(table(Getränk, Frau))

        Pearson's Chi-squared test

data:  table(Getränk, Frau)
X-squared = 8.6706, df = 3, p-value = 0.03401

Warning message:
In chisq.test(table(Getränk, Frau)) :
  Chi-squared approximation may be incorrect
>
> # Kontingenzkoeffizient KK_star
> m  <-  2   #  => Spaltenzahl (Frau) < Zeilenzahl (Getränk)
> n  <- 25   # Anzahl der Beobachtungen
> QK <- 8.67 # Vektor mit QK
> KK_star <- sqrt((QK / (QK + n)) * (m / (m - 1)))
> KK_star
[1] 0.717634
```

**Aufgabe 4.2: Kovarianz**

*Gegeben seien drei Variablen: X = {1,2,4,5}, $Y_1$ = {5,1,1,5} und $Y_2$ = {5,4,1,2}.*

a) *Berechnen Sie die Varianzen der drei Variablen und die Kovarianzen $c_{XY_1}$ und $c_{XY_2}$.*
b) *Stellen Sie den Zusammenhang zwischen X und $Y_1$ sowie zwischen X und $Y_2$ in einem Streudiagramm graphisch dar. Zeichnen Sie jeweils das arithmetische Mittel beider Variablen in das Streudiagramm ein. Welche Aussage hinsichtlich des Zusammenhangs zwischen X und $Y_1$ sowie zwischen X und $Y_2$ können Sie treffen?*
c) *Überprüfen Sie (4.15) anhand der Daten. Interpretation.*

**Lösung a)**

Die Varianz berechnet man mithilfe folgender Formel: $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Für die Kovarianz benötigen wir folgende Gleichung: $c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Zur vereinfachten Berechnung ermitteln wir folgende Werte: $(\bar{x} = 3, \bar{y}_1 = \bar{y}_2 = 3, n = 4)$

| $i$ | $X_i$ | $Y_{1_i}$ | $Y_{2_i}$ | $X_i - \bar{X}$ | $Y_{1_i} - \bar{Y}_1$ | $Y_{2_i} - \bar{Y}_2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 5 | 5 | -2 | 2 | 2 |
| 2 | 2 | 1 | 4 | -1 | -2 | 1 |
| 3 | 4 | 1 | 2 | 1 | -2 | -1 |
| 4 | 5 | 5 | 1 | 2 | 2 | -2 |

| $i$ | $(X_i - \bar{X})^2$ | $(Y_{1_i} - \bar{Y}_1)^2$ | $(Y_{2_i} - \bar{Y}_2)^2$ | $(X_i - \bar{X})(Y_{1_i} - \bar{Y}_1)^2$ | $(X_i - \bar{X})(Y_{2_i} - \bar{Y}_2)^2$ |
|---|---|---|---|---|---|
| 1 | 4 | 4 | 4 | -4 | -4 |
| 2 | 1 | 4 | 1 | 2 | -1 |
| 3 | 1 | 4 | 1 | -2 | -1 |
| 4 | 4 | 4 | 4 | 4 | -4 |
| $\sum$ | 10 | 16 | 10 | 0 | -10 |

Mithilfe der Ergebnisse in der Tabelle können wir nun die Varianzen der drei Variablen und die Kovarianzen $c_{XY_1}$ und $c_{XY_2}$ berechnen:

$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4}(10) = 2.5$

$s_{Y_1}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{1_i} - \overline{y_1})^2 = \frac{1}{4}(16) = 4$

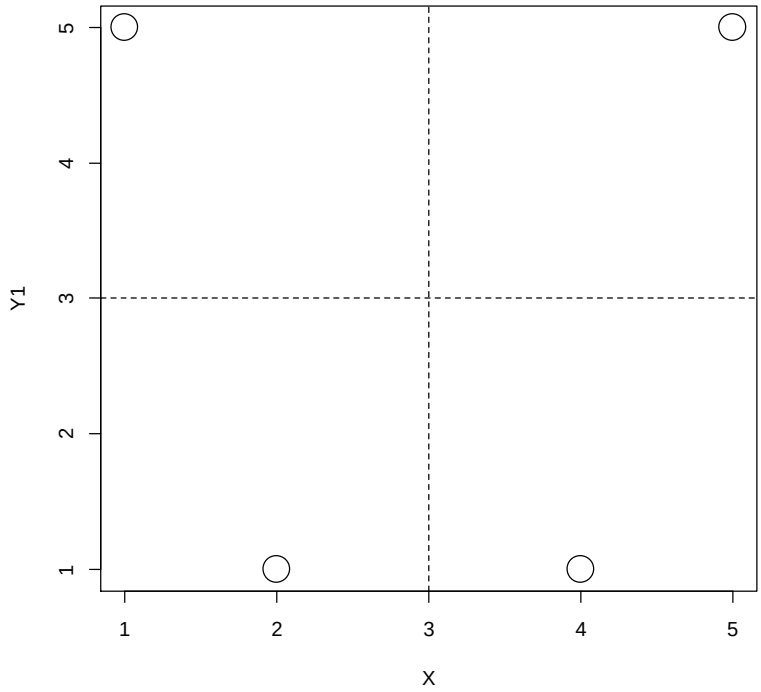
$s_{Y_2}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{2_i} - \overline{y_2})^2 = \frac{1}{4}(10) = 2.5$

$c_{X,Y_1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_{1i} - \bar{y}) = \frac{1}{4}(0) = 0$

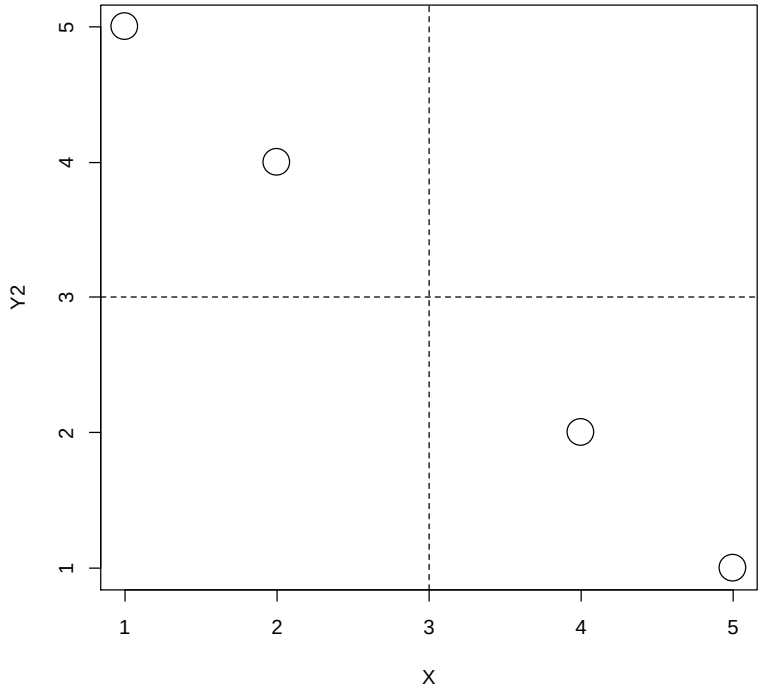
$c_{X,Y_2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_{2i} - \bar{y}) = \frac{1}{4}(-10) = -2.5$

**Lösung b)**

Zusammenhang zwischen $X$ und $Y_1$:




Es gibt einen näherungsweise quadratischen Zusammenhang zwischen beiden Variablen. Die Kovarianz ist 0, weil alle vier Quadranten gleichmäßig besetzt sind. In jedem Quadranten befindet sich genau ein Datenpunkt.
Zusammenhang zwischen X und $Y_2$:




Es gibt einen starken negativen linearen Zusammenhang zwischen beiden Variablen. Das bestätigt auch die Kovarianz, die hier negativ ausfällt (-2.5).

**Lösung c)**

| $X$ | $Y_1$ | $Y_2$ | $X+Y_1$ | $X+Y_2$ |
|---|---|---|---|---|
| 1 | 5 | 5 | 6 | 6 |
| 2 | 1 | 4 | 3 | 6 |
| 4 | 1 | 2 | 5 | 6 |
| 5 | 5 | 1 | 10 | 6 |

$\overline{X+Y_1} = 6, \quad \overline{X+Y_2} = 6, \quad n = 4$
Für die Varianz der Summe (gem. 4.15 im Buch) gilt:
$s^2_{X+Y} = s^2_X + s^2_Y + 2c_{XY}$
Somit gilt zu prüfen für $X$ und $Y_1$:
$s^2_{X+Y_1} = \frac{1}{n}\sum_{i=1}^{n}((x_i + y_{1i}) - 6)^2 = \frac{1}{4}(0 + 9 + 1 + 16) = \frac{26}{4} = \frac{13}{2} = 6.5$
$s^2_X + s^2_{Y_2} = 6.5 \Longrightarrow c_{XY_1} = 0$
Bedingung (4.15) ist erfüllt.
Die Kovarianz ist hier Null, es liegt keine lineare Abhängigkeit vor. Dennoch können wir nicht schließen, dass beide Variablen unabhängig voneinander sind (vgl. Streudiagramm)!
Für $X$ und $Y_2$:
$s^2_{X+Y_2} = \frac{1}{n}\sum_{i=1}^{n}((x_i + y_{2i}) - 6)^2 = \frac{1}{4}(0 + 0 + 0 + 0) = 0$
$s^2_{X+Y_2} = s^2_X + s^2_{Y_2} + 2c_{XY_2}$
$0 = (2.5 + 2.5) + \big(2 * (-2.5)\big)$
Bedingung (4.15) ist erfüllt.
Die Kovarianz ist negativ, es liegt eine negative lineare Abhängigkeit vor.

**Lösung mit R**

```
> # a)
> X  <- c(1,2,4,5)
> Y1 <- c(5,1,1,5)
> Y2 <- c(5,4,2,1)
> MX  <- mean(X)
> MY1 <- mean(Y1)
> MY2 <- mean(Y2)
> n   <- length(X)# Anzahl der Elemente in den Vektoren X, Y1, Y2
>
> # Varianzen
> var_X   <- 1/n * sum((X - MX)^2); var_X
[1] 2.5
> var_Y1  <- 1/n * sum((Y1 - MY1)^2); var_Y1
[1] 4
> var_Y2  <- 1/n * sum((Y2 - MY2)^2); var_Y2
[1] 2.5
>
> # Man beachte, dass var() die geschätzte Varianz der Grundgesamtheit
> # berechnet (mit Nenner 1/(n-1)). Wir rechnen hier mit der Varianz
> # der Stichprobe. Vgl. Buch, S. 51.

> # Kovarianzen
> cov_XY1 <- 1/n * sum((X - MX)*(Y1 - MY1))
> cov_XY1 # cov_XY1 = 0
[1] 0
> cov_XY2 <- 1/n * sum((X - MX)*(Y2 - MY2))
> cov_XY2 # cov_XY2 = -2.5
[1] -2.5
>
> # Man beachte, dass cov() die geschätzte Kovarianz der Grundgesamtheit
> # berechnet (mit Nenner 1/(n-1)). Wir rechnen hier mit der Kovarianz
> # der Stichprobe. Vgl. Buch, S. 70.
>
```

```
> # vereinfacht:
> cov_XY2 <- (1/n * sum(X*Y2)) - MX*MY2; cov_XY2
[1] -2.5

> # b)
> # Streudiagramm X und Y1
> plot(X, Y1, cex = 3)
> abline(v = mean(X), lty = 2)
> abline(h = mean(Y1), lty = 2)
>
> # Streudiagramm X und Y2
> plot(X, Y2, cex = 3)
> abline(v = mean(X), lty = 2)
> abline(h = mean(Y2), lty = 2)
```

**Aufgabe 4.3: Bildung**

*Eine Stichprobe von Arbeitnehmern wird nach „Bruttoeinkommen“ (Y, in 1000€) und nach „Bildungsgrad“ (Educ, in Jahre) befragt.*

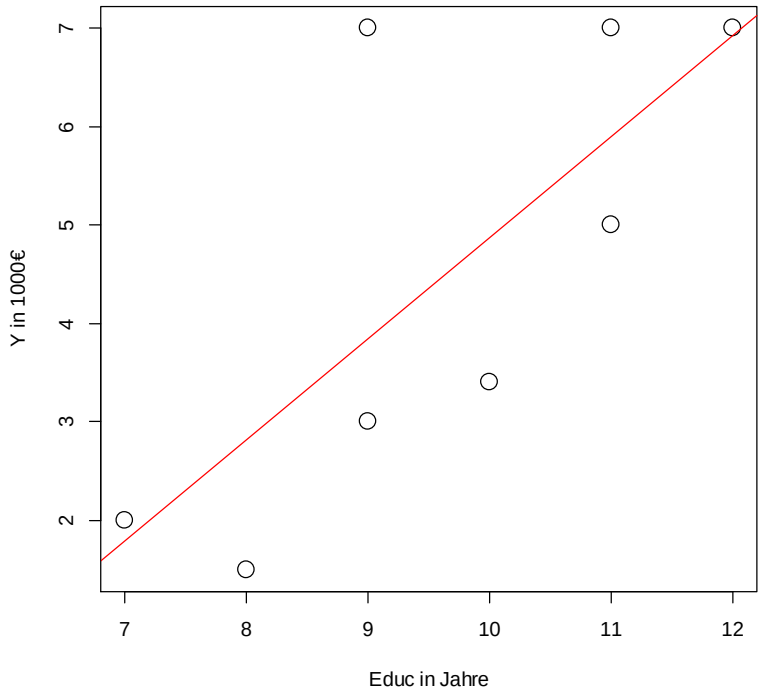
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Y*: | 3 | 5 | 7 | 1.5 | 7 | 2 | 3.4 | 7 |
| *Educ*: | 9 | 11 | 12 | 8 | 9 | 7 | 10 | 11 |

*Untersuchen Sie den Zusammenhang zwischen Y und Educ,*

*a) indem Sie den möglichen statistischen Zusammenhang geeignet graphisch darstellen und erläutern, und*

*b) den Zusammenhang rechnerisch ermitteln und interpretieren.*

**Lösung a)**




Zwischen den Variablen *Educ* und *Y* existiert ein positiver linearer Zusammenhang. Ein höherer Bildungsgrad geht einher mit höherem Einkommen und umgekehrt. Man beachte: Bei der Interpretation der Korrelation sollte nicht von einer Kausalität gesprochen werden.

**Lösung b)**

Da beide Variablen quantitativ sind, eignet sich der Bravais-Pearson-Korrelationskoeffizient als statistisches Maß für die Stärke des linearen Zusammenhangs.

Für die Berechnung des Pearson-Korrelationskoeffizienten verwenden wir folgende Formel:

$r_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_{s,i} y_{s,i} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right)$

Dabei bezeichnet $n$ die Anzahl der Beobachtungen und beträgt hier 8.

Die Standardabweichungen betragen:

$s_{Educ} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Educ_i^2 - \overline{Educ}^2} = 1.685$

$s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2} = 2.321$

Mithilfe der Standardabweichungen ergibt sich folgende Rechnung:

$r_{EducY} = \frac{1}{8}\sum_{i=1}^{8} \left(\frac{Educ_i - \overline{Educ}}{1.685}\right)\left(\frac{y_i - \bar{y}}{2.321}\right) = 0.7475$

Alternativ lässt sich auch folgende Formel verwenden:

$r_{EducY} = \frac{c_{Educ,Y}}{s_{Educ} \times s_Y} = \frac{2.923}{1.685 \times 2.321} = 0.7475$

Der Bravais-Pearson-Korrelationskoeffizient ist ein normiertes Maß für die Stärke des linearen Zusammenhangs zweier quantitativer Variablen. Anhand des berechneten Wertes kann man von einem linearen Zusammenhang mittlerer Stärke ausgehen. Dies macht auch das Streudiagramm oben deutlich.

**Lösung mit R**

```
# a) Eingabe der Daten und Plot
Y    <- c(3 ,  5 ,  7 , 1.5 , 7 , 2 , 3.4 ,  7)
Educ <- c(9 , 11 , 12 ,   8 , 9 , 7 ,  10 , 11)
plot(Educ , Y, cex = 2, xlab = "Educ in Jahre", ylab = "Y in 1000€")
abline(lm(Y ~ Educ), col = "red")

> # b) Drei Möglichkeiten zur Berechnung des Korrelationskoeffizienten:
> # Variante 1:
> # Berechnung der standardisierten Werte von Educ
> sEduc <- sqrt((1/n)*sum((Educ - mean(Educ))^2)); sEduc
[1] 1.57619
> z_Educ <- (Educ - mean(Educ))/sEduc
> # Berechnung der standardisierten Werte von Y
> sY <- sqrt((1/n)*sum((Y - mean(Y))^2)); sY
[1] 2.171081
> z_Y <- (Y - mean(Y)) / sY
> n   <- 8# Anzahl Merkmalsträger (Paare)
> # Pearson-Korrelationskoeffizient als arithmetisches Mittel der Produkte
> # der standardisierten Variablenpaare (Buch S. 73)
> (1/n)*sum(z_Educ * z_Y)
[1] 0.7474533
>
> # Variante 2:
> # mit geschätzter Standardabweichung und Kovarianz
> # der Grundgesamtheit (mit n im Nenner)
> # gleiches Resultat wie oben
> cov(Y, Educ)/(sd(Y)*sd(Educ))
[1] 0.7474533
>
> # Variante 3:
> # Syntax cor() in R
> cor(Y, Educ, method = "pearson")
[1] 0.7474533
```

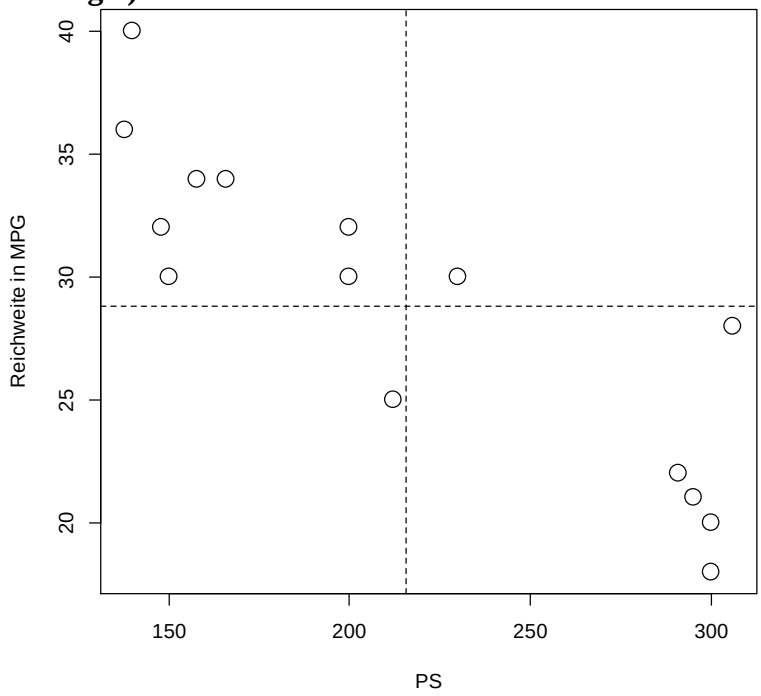
**Aufgabe 4.4: PS**

*Eine Studie zum Pkw-Markt in den USA liefert folgende Daten zu den Merkmalen „PS“ (Pferdestärken) und „mpg“ (Miles per Gallon, Reichweite).*

| Pkw | PS | mpg |
|---|---|---|
| Audi A4 | 200 | 32 |
| BMW 328 | 230 | 30 |
| Buick LaCrosse | 200 | 30 |
| Chevy Cobalt | 148 | 32 |
| Chevy TrailBlazer | 291 | 22 |
| Ford Expedition | 300 | 20 |
| GMC Yukon | 295 | 21 |
| Honda Civic | 140 | 40 |
| Honda Accord | 166 | 34 |
| Hyundai Elantra | 138 | 36 |
| Lexus IS 350 | 306 | 28 |
| Lincoln Navigator | 300 | 18 |
| Mazda Tribute | 212 | 25 |
| Toyota Camry | 158 | 34 |
| Volkswagen Beetle | 150 | 30 |

*a) Erstellen Sie ein Streudiagramm und berechnen Sie die Kovarianz. Interpretation.*
*b) Berechnen und interpretieren Sie den Pearson-Korrelationskoeffzienten.*

**Lösung a)**




Zur Berechnung der Kovarianz benötigen wir folgende Formel:
$c_{MPG,PS} = \frac{1}{n}\sum_{i=1}^{n}(MPG_i - \overline{MPG})(PS_i - \overline{PS})$
Hier ist: $n = 15$, $\overline{MPG} = 28.8$ , $\overline{PS} = 215.6$
$\Longrightarrow c_{MPG,PS} = -343.61$
Eine Kovarianz von -343.61 gibt an, dass sich relativ viele Datenpunkte im Quadranten links oben und im Quadranten rechts unten befinden. Es gibt also einen negativen linearen Zusammenhang, allerdings ist unklar, wie stark dieser ist.

**Lösung b)**
Da die Kovarianz bereits bekannt ist, bietet es sich hier an, die folgende Formel zur Berechnung des Korrelationskoeffizienten zu nutzen:

$r_{XY} = \frac{c_{XY}}{s_X s_Y}$

Dafür müssen zunächst noch die Standardabweichungen der beiden Variablen ermittelt werden.

$s_{MPG} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} MPG_i^2 - 28.8^2} = 6.18$

$s_{PS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} PS_i^2 - 215.6^2} = 64.03$

Eingesetzt in die Formel ergibt sich folgendes Ergebnis:

$r_{MPG,PS} = \frac{c_{MPG,PS}}{s_{MPG} s_{PS}} = \frac{-343.61}{6.18 * 64.03} = -0.868$

Der Korrelationskoeffizient von -0.868 weist auf einen starken negativen, linearen Zusammenhang zwischen MPG und PS hin. Eine stärkere PS-Zahl geht einher mit einer geringeren Reichweite und umgekehrt.

**Lösung mit R**

```
# a) Eingabe der Daten und Plot
PS  <- c(200, 230, 200, 148, 291, 300, 295,
         140, 166, 138, 306, 300, 212, 158, 150)

MPG <- c(32, 30, 30, 32, 22, 20, 21,
         40, 34, 36, 28, 18, 25, 34, 30)
plot(PS, MPG, cex = 2,
ylab = "Reichweite in MPG")            # Streudiagramm
abline(h = mean(MPG), lty = 2); abline(v = mean(PS), lty = 2)

> # Berechnung der deskriptiven Kovarianz:
> n <- 15
> kov <- 1/n*sum((PS - mean(PS)) * (MPG - mean(MPG)))
> kov
[1] -343.6133
>
> # b)
> # Standardabweichung der Variable PS:
> sdPS <- sqrt(1/n * sum((PS - mean(PS))^2))
> sdPS
[1] 64.03312
>
> # Standardabweichung der Variable MPG:
> sdMPG <- sqrt(1/n * sum((MPG - mean(MPG))^2))
> sdMPG
[1] 6.177378
>
> # Berechnung des Korrelationskoeffizienten
> kov/(sdPS*sdMPG)
[1] -0.8686827
> # oder
> cor(PS, MPG, method = "pearson")
[1] -0.8686827
```

**Aufgabe 4.5: Lebensmittel**

*Verschiedene Produkte eines Lebensmittels wurden auf Haltbarkeit und Geschmack getestet. Um die Produkte in ihrer Bewertung zu ordnen, wurde folgendes Schema verwendet {ausgezeichnet, sehr gut, gut, durchschnittlich, unterdurchschnittlich, schlecht}.*

| Produkt | Geschmack | Haltbarkeit |
|---|---|---|
| A | ausgezeichnet | schlecht |
| B | ausgezeichnet | durchschnittlich |
| C | durchschnittlich | ausgezeichnet |
| D | unterdurchschnittlich | ausgezeichnet |
| E | sehr gut | durchschnittlich |
| F | durchschnittlich | sehr gut |
| G | ausgezeichnet | unterdurchschnittlich |
| H | gut | gut |

*Gibt es einen Zusammenhang zwischen Haltbarkeit und Geschmack? Berechnen Sie den passenden Korrelationskoeffizienten und interpretieren Sie das Ergebnis.*

**Lösung**

Zunächst werden für die Bewertungen des Geschmacks und der Haltbarkeit numerische Werte auf einer Skala von 1 bis 6 zugeordnet, welche als Note bezeichnet werden können:

| Bewertung | Note (1-6) |
|---|---|
| Ausgezeichnet | 1 |
| Sehr gut | 2 |
| Gut | 3 |
| Durchschnittlich | 4 |
| Unterdurchschnittlich | 5 |
| Schlecht | 6 |

Anschließend werden die Produkte entsprechend ihrer Noten klassifiziert und Rängen zugeordnet. Haben zwei Werte denselben Rang, so erhalten beide den Mittelwert der aufeinanderfolgenden Ränge. Es entsteht folgende Tabelle:

| Produkt | Note Geschmack (X) | Note Haltbarkeit (Y) | Rang X | Rang Y |
|---|---|---|---|---|
| A | 1 | 6 | 2 | 8 |
| B | 1 | 4 | 2 | 5.5 |
| C | 4 | 1 | 6.5 | 1.5 |
| D | 5 | 1 | 8 | 1.5 |
| E | 2 | 4 | 4 | 5.5 |
| F | 4 | 2 | 6.5 | 3 |
| G | 1 | 5 | 2 | 7 |
| H | 3 | 3 | 5 | 4 |

Mithilfe der Ränge lässt sich nun der Spearman-Korrelationskoeffizient bestimmen:

$$r_{XY}^{Sp} = r_{rg(X),rg(Y)} = \frac{c_{rg(X),rg(Y)}}{s_{rg(X)} s_{rg(Y)}} = \frac{-5.3929}{2.3755 \times 2.4202} = -0.9380$$

Es gibt also einen sehr starken negativen linearen Zusammenhang zwischen den Rängen für Geschmack und Haltbarkeit. Das heißt, dass Produkte, die sehr gut schmecken, sehr schnell verderben und umgekehrt.

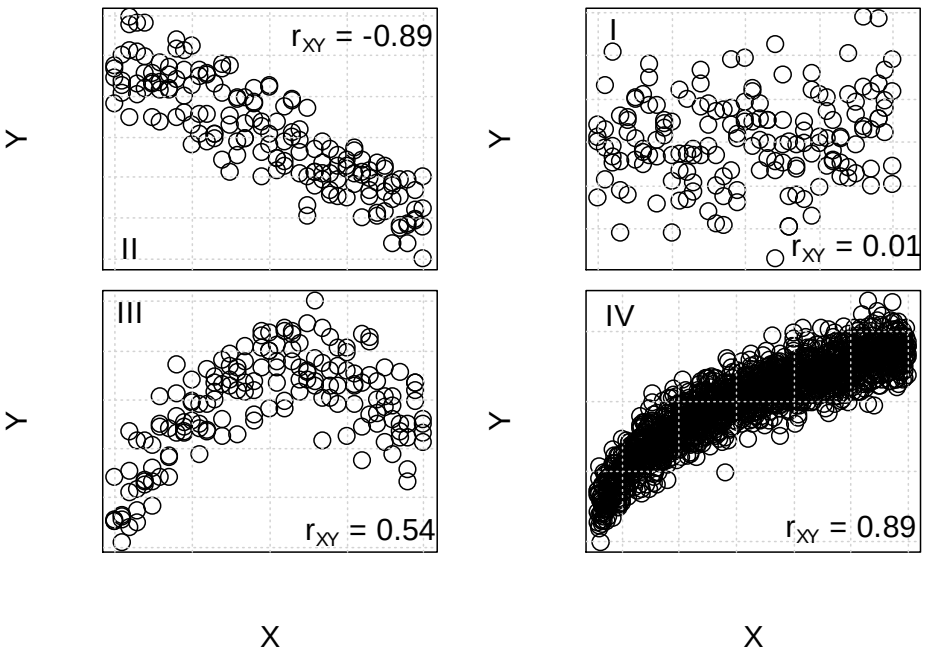
**Lösung mit R**

```
> # Zwei Alternativen:
> # 1) Wir bestimmen die Ränge und berechnen dann
> # den Pearson-Korrelationskoeffizienten der Ränge
> # "Note Geschmack" = G und "Note Haltbarkeit" = H
> G   <- c(1 , 1 , 4 , 5 , 2 , 4 , 1 , 3)
> H   <- c(6 , 4 , 1 , 1 , 4 , 2 , 5 , 3)
> # Wir bestimmen die Ränge
```

```
> rank(G)
[1] 2.0 2.0 6.5 8.0 4.0 6.5 2.0 5.0
> rank(H)
[1] 8.0 5.5 1.5 1.5 5.5 3.0 7.0 4.0
>
> # Die Korrelation der Ränge lässt sich nun mithilfe des Pearson-
> # Korrelationskoeffizienten berechnen:
> cor(rank(G), rank(H), method = "pearson")
[1] -0.9380511
>
> # 2) mit R-Syntax
> cor(G , H, method = "spearman")
[1] -0.9380511
>
> # Das hier wäre falsch:
> cor(G , H, method = "pearson")
[1] -0.9392339
```

**Aufgabe 4.6: Streudiagramme**

*Welche Aussagen lassen sich für die vier unten abgebildeten Streudiagramme und den dazugehörigen Korrelationskoeffizienten treffen?*




**Lösung**

Streudiagramm I: Die Datenpunkte verteilen sich gleichmäßig im gesamten Streudiagramm. Es ist kein linearer Zusammenhang der Datenpunkte erkennbar, dies zeigt auch der Korrelationskoeffizient, welcher praktisch Null ist.
Streudiagramm II: Die Datenpunkte weisen einen starken, negativen linearen Zusammenhang auf. Dafür spricht auch der Korrelationskoeffizient von -0.89.
Streudiagramm III: Die Datenpunkte verlaufen in Form einer nach unten offenen Parabel (inverser quadratischer Zusammenhang). Der Korrelationskoeffizient von 0.54 weist auf eine mäßig starke Korrelation hin. Hier liegt jedoch ein Trugbild vor, da der Korrelationskoeffizient lediglich den linearen Zusammenhang aufzeigt. Viele Datenpunkte liegen "links unten" und "rechts oben" im Streudiagramm, was somit den hohen Korrelationskoeffizienten erklärt. Der zugrundeliegende Zusammenhang ist aber nicht linear.

Streudiagramm IV: Der Korrelationskoeffizient von 0.89 weist auf eine positive, lineare Korrelation hin. Jedoch verlaufen die Datenpunkte in einer leicht degressiv steigenden Kurve und es ist klar erkennbar, dass keine linearer Zusammenhang besteht. Auffallend ist hier, dass die Datenpunkte sehr dicht aneinander liegen.
Fazit: Datenpunkte deshalb immer im Streudiagramm anzeigen lassen, um Fehlinterpretationen zu vermeiden. Einer betragsmäßig großen Korrelation muss nicht zwangsläufig auch ein linearer Zusammenhang zugrunde liegen.

# Kapitel 5: Lineare Regression

**Aufgabe 5.1: Einkommen und Ausgaben**

*Ein Bekleidungsunternehmen mit Internet-Vertrieb untersucht seine Datenbank, um herauszufinden, ob es einen Zusammenhang zwischen den jährlichen Ausgaben (in €) eines Kunden und dem Einkommen des Kunden (in €) gibt. Es sei angenommen, dass die Annahmen des linearen Regressionsmodells erfüllt sind.*

*Die Regressionsgerade ist* $\widehat{Ausgaben} = -31.6 + 0.012 * Einkommen$.

*a) Interpretieren Sie Anstiegsparameter und Achsenabschnitt.*
*b) Wenn ein Kunde ein Einkommen von 20000 € hat, wie hoch sind die prognostizierten jährlichen Ausgaben?*
*c) Wie groß ist der Fehlerterm, wenn die tatsächlichen Ausgaben 100 € betragen?*

**Lösung a)**
Der Anstiegsparameter ($\beta_1$) beträgt 0.012, das heißt, mit jedem zusätzlichen Euro Einkommen des Kunden werden 1.2 Cent für Bekleidung ausgegeben. Oder: Steigt das Einkommen um 100€, werden 1,20€ mehr für Bekleidung ausgegeben.
Der Achsenabschnitt von -31.6 (€) kann – rein technisch – interpretiert werden als die Ausgaben, die ein Kunde im Falle eines Einkommens von 0 € haben würde. Dies kann hier jedoch nicht sinnvoll interpretiert werden.

**Lösung b)**
Berechnung der geschätzten jährlichen Ausgaben $\hat{A}$:
$\hat{A}$ = geschätzte Ausgaben, $E$ = Einkommen
$\hat{A} = -31.6 + 0.012 \cdot E$
$\hat{A}(20000\ €) = -31.6 + 0.012 \cdot 20000 = 208.40€$
Bei einem Einkommen von 20000€ betragen die prognostizierten jährlichen Ausgaben des Kunden 208.40 €.

**Lösung c)**
Die Aufgabe bezieht sich auf das Einkommen b). Der Fehlerterm $u$ berechnet sich mithilfe folgender Formel:
$u = y - \hat{y} = Beobachtung - Vorhersage$
$u = 100\ € - 208.4\ € = -108.4€$
Der Fehlerterm beträgt -108.4€, wenn die tatsächlichen Ausgaben 100€ betragen.

**Aufgabe 5.2: Fahrleistung und Preis**

*Ein lineares Modell zur Vorhersage des Preises eines 2010er Dacia Sandero (in €) in Abhängigkeit von seinem Kilometerstand (in km) wurde für 38 Autos dieser Marke wie folgt geschätzt:* $\widehat{Preis} = 24356 - 0.055 * km$

*a) Was ist die unabhängige Variable? Was ist die abhängige Variable?*
*b) Was bedeutet der Anstiegsparameter in diesem Zusammenhang?*

c) *Was bedeutet der Achsenabschnitt in diesem Zusammenhang? Ist er von Bedeutung?*
d) *Welcher Preis ergibt sich für ein Auto mit einem Kilometerstand von 100000 km?*
e) *Wenn der tatsächliche Preis für ein Auto mit einem Kilometerstand von 100000 km bei 21000 € lag, wie groß ist dann der Fehlerterm?*
f) *Der durchschnittliche Preis der Autos lag bei 20847 €, die Standardabweichung betrug 923 € und die Korrelation zwischen Preis und Kilometerstand lag bei -0.269. Wenn bei einem 2010er Dacia Sandero der Kilometerstand eine Standard-abweichung unter dem durchschnittlichen Kilometerstand lag, wie hoch wäre dann der Preis für dieses Auto?*

**Lösung a)**
Die unabhängige Variable ist die Anzahl der gefahrenen Kilometer, die abhängige Variable ist der Preis. Dies ist dadurch zu erklären, dass der Preis vom Kilometerstand des Dacia Sandero abhängig ist.

**Lösung b)**
Der Anstiegsparameter von 0.055 bedeutet, dass sich der Preis (oder auch der Wert) eines Dacia Sandero um 5.5 Cent pro mehr gefahrenem Kilometer verringert.

**Lösung c)**
Der Achsenabschnitt bedeutet, dass ein Dacia Sandero mit einem Kilometerstand von 0 gefahrenen Kilometern einen Preis von 24356€ hätte. Damit gibt der Achsenabschnitt den geschätzten Neuwert eines Dacia Sandero an.

**Lösung d)**
Der Preis eines Autos mit 100000km ergibt sich wie folgt:
$\hat{P} = 24356 - 0.055 * 100000$
$\hat{P}(100000\ \text{km}) = 18856\ €$
Der prognostizierte Preis eines Dacia Sanderos (mit 100.000km Laufleistung) würde 18856 € betragen.

**Lösung e)**
$u = y - \hat{y} = Beobachtung - Vorhersage$
$u = 21000\ € - 18856\ € = 2144€$

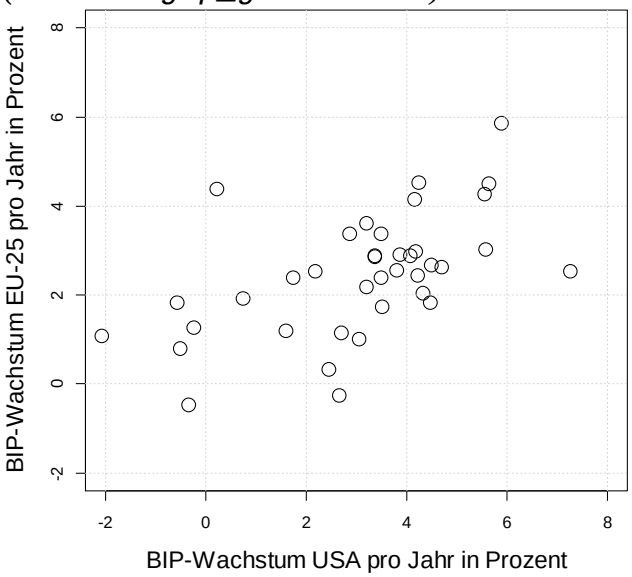
**Lösung f)**
Die Regressionsgleichung der standardisierten Variablen (S. 93 im Buch) ist:

$$\hat{y}_s = r_{XY} x_s$$

Nach dieser Gleichung führt eine Veränderung der standardisierten unabhängigen Variable $x_s$ um eine Standardabweichung zu einer Veränderung der standardisierten prognostizierten Werte in Höhe von $r_{XY}$. Dies gilt für alle Werte von $x$ und $y$, also auch für die Mittelwerte von $x$, $y$ und der prognostizierten $y$-Werte.
Wenn wir nun 1 Standardabweichung bei der unabhängigen Variablen vom Mittelwert nach unten abweichen (diese Information war gegeben), verändert sich die abhängige Variable genau um $r_{XY} \cdot (-1)$ Standardabweichungen von ihrem Mittelwert. Hier also um $+0.269 \cdot 923$. Der Mittelwert der abhängigen Variablen war gegeben.
Es ergibt sich der Wert: $20847 + 0.269 \cdot 923 = 21095.29$.
Der Preis für diesen Dacia Sandero würde also bei 21095.3€ liegen.

**Aufgabe 5.3: Wirtschaftswachstum EU und USA**

*Steht das Wirtschaftswachstum in Europa in einem Zusammenhang zu dem Wachstum in den USA? In der Abbildung unten sehen Sie ein Streudiagramm des durchschnittlichen Wachstums in 25 europäischen Ländern (in % des BIP) gegenüber dem Wachstum in den USA (in % des BIP). Jeder Punkt repräsentiert ein Jahr von 1970 bis 2007.*
*(Datensatz: `gdp_growth.csv`).*




*Der Output einer Regressionsanalyse sieht wie folgt aus:*

```
Coefficients:       Intercept              US
                    1.3297                 0.3616
                    R-squared:             0.2965
```

a) *Überprüfen Sie die Annahmen für das lineare Modell.*
b) *Erklären Sie die Bedeutung des Bestimmtheitsmaßes.*
c) *Bestimmen Sie die Gleichung der Regressionsgeraden. Welche Bedeutung hat der Achsenabschnitt? Macht das Sinn? Interpretieren Sie den Anstiegsparameter.*
d) *Im Jahr 2007 lag das Wachstum in den USA bei 3.2%, währenddessen die Wirtschaft in Europa nur um 2.16% gewachsen ist. Ist dieser Wert kleiner oder größer als der Wert, den man anhand der Regressionsgleichung berechnet hätte? Welchen Wert hat der Fehlerterm in diesem Jahr?*

**Lösung a)**
Annahme 1: Quantitative Daten. Diese Annahme ist erfüllt, da das BIP-Wachstum ein kardinal skaliertes Merkmal ist, welches abzählbar ist und eine Einheit (%) besitzt.
Annahme 2: Linearer Zusammenhang. Diese Annahme ist einigermaßen erfüllt (positiver linearer Zusammenhang). Vgl. Streudiagramm oben.
Annahme 3: Ausreißer: Es sind keine Ausreißer erkennbar.
Annahme 4: Homoskedastizität. Die Standardabweichung der Fehlerterme muss entlang der Regressionsgeraden konstant bleiben. Dieses Kriterium ist ausreichend erfüllt.

**Lösung b)**
Das Bestimmtheitsmaß quantifiziert im Intervall [0,1], wie gut sich die Regressionsgerade an die Punktwolke anpasst. Es wird berechnet, wie groß der Anteil der von der Regressionsgeraden erklärten Varianz an der gesamten Varianz des BIP-Wachstums der EU-Staaten (abhängige Variable) ist. Hier kann das Regressionsmodell ca. 30% der Varianz der

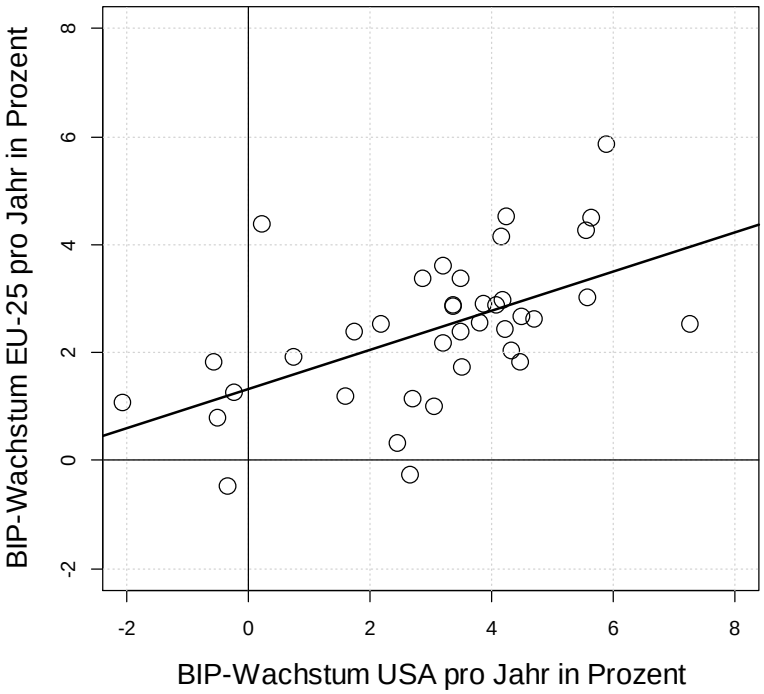
abhängigen Variablen erklären. Der Rest der Streuung bleibt unerklärt bzw. entfällt auf die Residuen.

**Lösung c)**
Gleichung der Regressionsgeraden:
$\widehat{BIP_{Wachstum_{EU}}} = 1.3297 + 0.3616 * BIP_{Wachstum_{USA}}$
Achtung: Die Wachstumsraten sind hier in Prozent (%) im Vergleich zum Vorjahr angegeben. Dies ist bei der Interpretation des Anstiegsparameters zu berücksichtigen.
Der Achsenabschnitt gibt an, wie groß das BIP-Wachstum der EU in % im Vergleich zum Vorjahr wäre, wenn das Wirtschaftswachstum der USA 0% wäre (wenn also das US-BIP im Vergleich zum Vorjahr konstant wäre). Das macht also durchaus Sinn.
Der Anstiegsparameter sagt aus, um wie viel %-Punkte die EU-Wirtschaft im Vergleich zum Vorjahr mehr wächst, wenn die US-Wirtschaft um einen Prozentpunkt im Vergleich zum Vorjahr mehr wächst. Hier: Ein zusätzlicher Prozentpunkt Wachstum im Vergleich zum Vorjahr in den USA induziert ein zusätzliches Wachstum im Vergleich zum Vorjahr in der EU um 0.3616 Prozentpunkte.

**Lösung d)**
Wirtschaftswachstum nach der Regressionsgeraden:
$\widehat{BIP_{Wachstum_{EU}}} = 1.3297 + 0.3616 * 3.2\ \%$
$\widehat{BIP_{Wachstum_{EU}}} = 2.486\%$
Das tatsächliche BIP-Wachstum der EU-Staaten 2007 war mit 2.16% kleiner.
Wert des Fehlerterms:
$u = y - \hat{y} = Beobachtung - Vorhersage$
$u = 2.16\% - 2.4868\%$
$u = -0.3268\ \%$
Der Fehlerterm für das Jahr 2007 ist negativ. Er beträgt -0.3268 %.

**Lösung mit R**
```
> data <- read.csv("gdp_growth.csv")
> data[1:3, ]
  Year     US Euro25
1 1970 0.2167 4.3748
2 1971 2.8760 3.3583
3 1972 5.5540 4.2545
> attach(data)

# Streudiagramm
plot(US, Euro25, cex = 2, xlim = c(-2,8), ylim = c(-2,8),
xlab = "BIP-Wachstum USA pro Jahr in Prozent",
cex.lab = 1.4,
ylab = "BIP-Wachstum EU-25 pro Jahr in Prozent")
grid()

> # Berechnung der Regressionsgerade
> ols <- lm(Euro25 ~ US); ols

Call:
lm(formula = Euro25 ~ US)

Coefficients:
(Intercept)           US
     1.3297       0.3616

>
> # Berechnung R2
> # Methode 1: Quadrat des Korrelationskoeffizienten
```

```
> cor(Euro25, US)^2
[1] 0.2965059
>

> # Methode 2: Anteil der Varianz der Werte auf der
> # Regressionsgeraden an der Varianz der abh. Variablen
> var(ols$fitted.values)/var(Euro25)
[1] 0.2965059

# Streudiagramm mit Regressionsgerade
plot(US, Euro25, cex = 2, xlim = c(-2,8), ylim = c(-2,8),
xlab = "BIP-Wachstum USA pro Jahr in Prozent",
cex.lab = 1.4,
ylab = "BIP-Wachstum EU-25 pro Jahr in Prozent")
grid()
abline(lm(Euro25 ~ US), lwd = 2)
abline(h = 0, v = 0)
```



**Aufgabe 5.4: Umsatz und Arbeitslosenquote**

*Die Daten eines großen Einzelhandelsunternehmens wurden dafür verwendet, um ein lineares Regressionsmodell zu berechnen, durch welches der Quartalsumsatz ($rev$, in Mrd. \$) in den USA auf der Grundlage der US-Arbeitslosenquote ($u$, in %) vorhergesagt wird.*
*Dieses Regressionsmodell ergab ein Bestimmtheitsmaß von 88.3% und einen Anstiegsparameter von $-2.99$.*

a) *Interpretieren Sie das Bestimmtheitsmaß.*
b) *Wie hoch ist die Korrelation zwischen dem Quartalsumsatz und der Arbeitslosenquote?*
c) *Wenn in einem Quartal die Arbeitslosenquote 1.0% höher ist als in einem anderen Quartal, wie hoch ist die prognostizierte Auswirkung auf den Umsatz in diesem Quartal?*
d) *Die Regressionsanalyse liefert folgendes lineares Model:*

$$\widehat{rev} = 20.91 - 2.99u$$

*Wenn die Arbeitslosenquote bei 6.0% liegt, welcher Quartalsumsatz ergibt sich anhand der Regressionsgleichung?*

**Lösung a)**
Der Anteil der Varianz der Regressionswerte (vorhergesagte Werte bzw. Werte auf der Regressionsgeraden) hat einen Anteil von 88% an der Varianz der abhängigen Variablen. Die Güte der Anpassung (man sagt auch „die Erklärungskraft des Modells“) ist damit als sehr gut zu beschreiben.

**Lösung b)**
Es gilt $\sqrt{R^2} = r_{XY} \Longrightarrow \sqrt{0.883} = 0.94$.
Der Pearson-Korrelationskoeffizient zwischen Quartalsumsatz und Arbeitslosenquote kann als die Quadratwurzel des Bestimmtheitsmaßes berechnet werden und beträgt 0.94, was auf einen starken positiven linearen Zusammenhang schließen lässt.

**Lösung c)**
Bei einem Anstieg der Arbeitslosenquote um 1%-Punkt, sinkt der Quartalsumsatz um 2.99 Milliarden Dollar. Dies ergibt sich aus dem Anstiegsparameter.

**Lösung d)**
Berechnung des Quartalsumsatzes bei einer Arbeitslosenquote von 6.0%:
$\widehat{rev} = 20.91 - 2.99 * 0.06$
$\widehat{rev} = 20.73 \text{ Mrd. \$}$
Der prognostizierte Umsatz für eine Arbeitslosenquote von 6% liegt bei 20.73 Mrd. Dollar.

**Aufgabe 5.5: Produktion und Überstunden**

*Die Statistik eines Ventilatoren-Herstellers sieht für das Jahr 2005 wie folgt aus:*

| Monat | Produktion in Stück | Überstunden |
|---|---|---|
| Januar | 2600 | 70 |
| Februar | 3100 | 90 |
| März | 3350 | 130 |
| April | 3500 | 150 |
| Mai | 3600 | 170 |
| Juni | 3850 | 200 |
| Juli | 3600 | 160 |
| August | 3500 | 140 |
| September | 3200 | 130 |
| Oktober | 3100 | 80 |
| November | 3000 | 60 |
| Dezember | 2800 | 50 |

*a) Berechnen Sie den Korrelationskoeffizienten zwischen Produktion und geleisteten Überstunden. Besteht ein linearer Zusammenhang?*

*b) Angenommen, Sie sollen die Überstunden als lineare Funktion der Produktion darstellen, wie gehen Sie vor? Berechnen Sie die Regressionsfunktion und interpretieren Sie den Achsenabschnitt und den Anstiegsparameter.*

*c) Wie groß ist das arithmetische Mittel der Fehlerterme?*

*d) Fertigen Sie eine graphische Darstellung der Beobachtungswerte und der Regressionsgeraden an.*

*e) Welche Schätzung für die Zahl der Überstunden würden Sie bei einer Monatsproduktion von 3950 Stück abgeben?*

**Lösung a)**

Zunächst berechnen wir die Kovarianz zwischen Produktion ($X$) und Überstunden ($Y$), sowie deren Standardabweichungen.

Es gilt $c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})$ mit $n = 12$

$\bar{x} = 3266.67$ und $\bar{y} = 119.17$

$c_{XY} = 15180.56$

$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = 350.20$

$s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2} = 46.09$

Mithilfe dieser Werte lässt sich der Korrelationskoeffizient bestimmen:

$$r_{XY} = \frac{c_{XY}}{s_X s_Y} = \frac{15180.56}{350.20 * 46.09} = 0.941$$

Zwischen den Variablen Produktion und Überstunden besteht ein starker positiver linearer Zusammenhang.

**Lösung b)**

Die Regressionsgleichung $\hat{y} = \beta_0 + \beta_1 x$ lässt sich in folgenden Schritten ermitteln und interpretieren:

1) Berechnen des Anstiegsparameters $\beta_1$ und des Achsenabschnitts $\beta_0$ :

$\beta_1 = \frac{c_{XY}}{s_X^2} = \frac{15180.56}{350.20^2} = 0.124$

$\beta_0 = \bar{y} - \beta_1 \bar{x} = 119.17 - 0.124 * 3266.67 = -285.9$

2) Somit ergibt sich folgende Regressionsgleichung:

$\hat{y} = -285.9 + 0.124x$

2) Interpretation des Anstiegsparameters und des Achsenabschnittes:

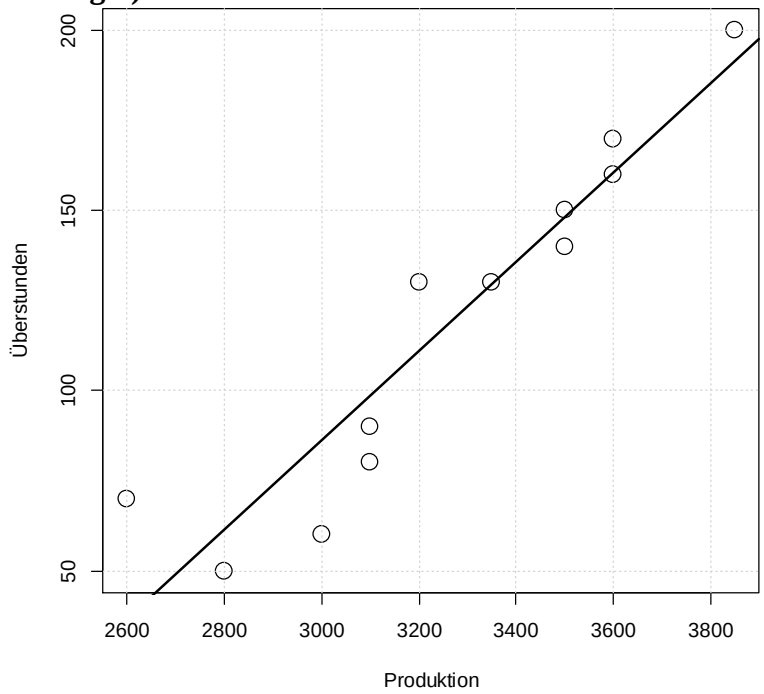
Der Anstiegsparameter $\beta_1$ sagt aus, dass wenn eine Einheit Produktion mehr erzeugt wird, 0.124 Überstunden (ca. 7.5 Minuten) mehr gearbeitet werden müssen. Der Achsenabschnitt $\beta_0$ $(-285.9)$ gibt an, wie viele Überstunden geleistet werden müssten, um 0 Einheiten der Produktion herzustellen. Dieser Wert ergibt jedoch keinen praktischen Sinn.

**Lösung c)**
Zur Ermittlung des arithmetischen Mittels der Fehlerterme müssen diese zunächst ermittelt werden: $u = y - \hat{y} = Beobachtung - Vorhersage$

| *x* | *y* | $\hat{y}$ | *u* |
|---|---|---|---|
| 2600 | 70 | 36.64 | 33.36 |
| 3100 | 90 | 98.54 | -8.54 |
| 3350 | 130 | 129.48 | 0.52 |
| 3500 | 150 | 148.05 | 1.95 |
| 3600 | 170 | 160.43 | 9.57 |
| 3850 | 200 | 191.37 | 8.63 |
| 3600 | 160 | 160.43 | -0.43 |
| 3500 | 140 | 148.05 | -8.05 |
| 3200 | 130 | 110.91 | 19.09 |
| 3100 | 80 | 98.54 | -18.54 |
| 3000 | 60 | 86.16 | -26.16 |
| 2800 | 50 | 61.40 | -11.40 |

Mithilfe dieser Tabelle lässt sich nun der Mittelwert der Fehlerterme $\bar{u} \approx 0$ bestimmen. Dies war zu erwarten (vgl. Buch S. 90).

**Lösung d)**




**Lösung e)**
Für die Prognose kann die Regressionsgleichung $\hat{y} = \beta_0 + \beta_1 x$ verwendet werden.
Schätzen der Überstundenanzahl für eine Produktion von $x = 3950$:
$\hat{y} = -285.9 + 0.124 * 3950$

$\hat{y} = 203.9$
Bei einer Produktion von 3950 Stück sind ca. 204 Überstunden zu erwarten.

**Lösung mit R**

```
> # a)
> # Eingabe der Daten von Produktion (P) und Überstunden (Ü)
> P <- c(2600, 3100, 3350, 3500, 3600, 3850, 3600, 3500, 3200, 3100, 3000, 2800)
> Ü <- c(70, 90, 130, 150, 170, 200, 160, 140, 130, 80, 60, 50)
> # Berechnung des Korrelationskoeffizienten
> cor(P, Ü)
[1] 0.9405
>
> # oder mit der Formel
> n <- length(P); n
[1] 12
> COV <- (1/n)*(sum((P-mean(P))*(Ü-mean(Ü)))); COV
[1] 15181
> SD_P <- sqrt((1/n)*(sum((P-mean(P))*(P-mean(P))))); SD_P
[1] 350.2
> SD_Ü <- sqrt((1/n)*(sum((Ü-mean(Ü))*(Ü-mean(Ü))))); SD_Ü
[1] 46.09
> COV/(SD_P*SD_Ü)
[1] 0.9405
>
> # Achtung: Hier wird mit den Werten der Kovarianz bzw. Varianz/SD
> # der Stichprobe gerechnet (Faktor (1/n)).
> # Das gleiche Ergebnis ergibt sich, wenn mit den Werten der Kovarianz
> # bzw. Varianz/SD der Grundgesamtheit gerechnet wird (Faktor (1/(n-1)))
>
> cov(P, Ü); sd(P); sd(Ü)
[1] 16561
[1] 365.8
[1] 48.14
> (1/(n-1))*(sum((P-mean(P))*(Ü-mean(Ü))))
[1] 16561
> sqrt((1/(n-1))*(sum((P-mean(P))*(P-mean(P)))))
[1] 365.8
> sqrt((1/(n-1))*(sum((Ü-mean(Ü))*(Ü-mean(Ü)))))
[1] 48.14
>
> cov(P, Ü)/(sd(P)*sd(Ü))
[1] 0.9405
>
> # b)
> # Berechnung der Regressionsgeraden
> b1 <- COV/SD_P^2; b1
[1] 0.1238
> b0 <- mean(Ü) - b1*mean(P); b0
[1] -285.2
>
> ols <- lm(Ü~P); ols

Call:
lm(formula = Ü ~ P)

Coefficients:
(Intercept)            P
   -285.190        0.124

> # Streudiagramm mit Regressionsgerade
> plot(P, Ü, cex = 2, xlab = "Produktion", ylab = "Überstunden")
> grid()
> abline(lm(Ü~P), lwd = 2)
>
> # c)
> y_dach <- ols$fitted.values # Werte auf der Regressionsgeraden
> y <- Ü # abhängige Variable
> err <- y - y_dach           # Fehlerterme
```

```
> options(digits = 4)         # Begrenzung der Ausgabe
> data.frame(y, y_dach, err)
     y y_dach      err
1   70  36.64  33.3550
2   90  98.54  -8.5362
3  130 129.48   0.5181
4  150 148.05   1.9507
5  170 160.43   9.5725
6  200 191.37   8.6268
7  160 160.43  -0.4275
8  140 148.05  -8.0493
9  130 110.91  19.0855
10  80  98.54 -18.5362
11  60  86.16 -26.1580
12  50  61.40 -11.4015
> mean(err)
[1] 2.961e-15
```

**Aufgabe 5.6: Einkommen und Kosmetik**

*Bei sechs deutschen Haushalten wurden jeweils das Haushaltsnettoeinkommen (X, in 1000 € pro Jahr) und die monatlichen Ausgaben für Kosmetikprodukte (Y, in €) erhoben:*

| | | | | | | |
|---|---|---|---|---|---|---|
| *X* | 50 | 40 | 60 | 70 | 30 | 65 |
| *Y* | 28 | 25 | 30 | 43 | 20 | 32 |

*a) Berechnen Sie zuerst $s_X^2$, $s_Y^2$ und $c_{XY}$.*
*b) Stellen Sie fest, ob ein linearer Zusammenhang zwischen dem Haushaltsnettoeinkommen und den monatlichen Ausgaben für Kosmetikprodukte besteht. Geben Sie gegebenenfalls die Richtung und die Stärke an.*
*c) Beschreiben Sie diesen Zusammenhang durch eine lineare Regressionsfunktion. Interpretieren Sie den Anstiegsparameter. Berechnen und interpretieren Sie das Bestimmtheitsmaß.*
*d) Berechnen Sie das arithmetische Mittel der Fehlerterme.*
*e) Stellen Sie die beobachteten Werte und die Regressionsgerade graphisch dar.*

**Lösung a)**

Es gilt $c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})$ mit $n = 6$.
Es ist $\bar{x} = 52.5$ und $\bar{y} = 29.67$
$c_{XY} = 90.833$
$s_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = 197.917$
$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = 50.222$

**Lösung b)**

$r_{XY} = \frac{c_{XY}}{s_X * s_Y} = \frac{90.833}{\sqrt{197.917} * \sqrt{50.222}} = 0.911$

Es liegt ein sehr starker positiver linearer Zusammenhang zwischen den Variablen Einkommen und Ausgaben für Kosmetik vor. Das heißt, ein Anstieg im Einkommen bedeutet einen konstanten Anstieg bei den Ausgaben für Kosmetik.

**Lösung c)**

1) Berechnen des Anstiegsparameters $\beta_1$ und des Achsenabschnitts $\beta_0$:

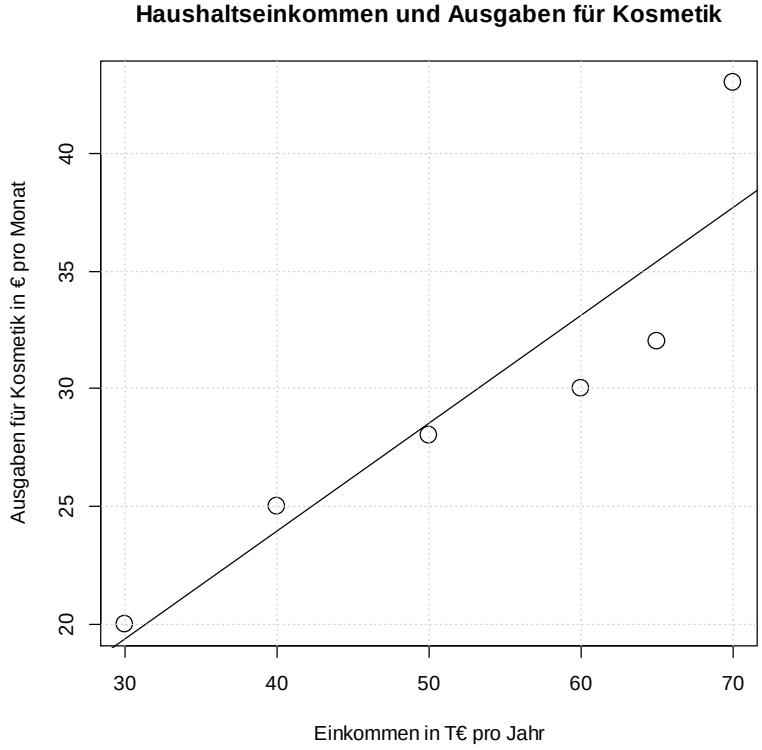
$\beta_1 = \frac{c_{XY}}{s_X^2} = \frac{90.833}{197.917} = 0.459$
$\beta_0 = \bar{y} - \beta_1 \bar{x} = 29.667 - 0.459 * 52.5 = 5.57$
2) Somit ergibt sich folgende Regressionsgleichung:
$\hat{y} = 5.57 + 0.459x$
Der Anstiegsparameter $\beta_1$ bedeutet, dass von 1000€ zusätzlichen Einkommen im Jahr ca. 0.46€ pro Monat (also ca. 5.50€ pro Jahr) für Kosmetikartikel ausgegeben werden.
3) Berechnung des Bestimmtheitsmaßes:
$R^2 = {r_{XY}}^2 = 0.911^2 = 0.83$
Das Bestimmtheitsmaß $R^2$ sagt aus, dass das lineare Regressionsmodell ca. 83% der Varianz der abhängigen Variablen (Ausgaben) erklären kann. Die restlichen ca. 17 % werden nicht erklärt und verbleiben in den Residuen.

**Lösung d)**
Zur Ermittlung des arithmetischen Mittels der Fehlerterme müssen diese zunächst ermittelt werden: $u = y - \hat{y} = Beobachtung - Vorhersage$

| *x* | *y* | $\hat{y}$ | *u* |
|---|---|---|---|
| 50 | 28 | 28.5 | -0.52 |
| 40 | 25 | 23.9 | 1.07 |
| 60 | 30 | 33.1 | -3.11 |
| 70 | 43 | 37.7 | 5.30 |
| 30 | 20 | 19.3 | 0.66 |
| 65 | 32 | 35.4 | -3.40 |

Mithilfe dieser Tabelle lässt sich nun der Mittelwert der Fehlerterme $\bar{u}$ bestimmen. Er ist – wie zu erwarten – nahe 0.

**Lösung e)**



**Lösung mit R**

```
> # Eingabe der Daten für Einkommen (X) und Kosmetikprodukte (Y)
> X   <- c(50 , 40 , 60 , 70 , 30 , 65)
> Y   <- c(28 , 25 , 30 , 43 , 20 , 32)
> MX  <- mean(X); MX
[1] 52.5
> MY  <- mean(Y); MY
[1] 29.66667
> n   <- length(X) # entspricht auch length(Y)
> # Eingabe der Daten für Einkommen (X) und Kosmetikprodukte (Y)
> X   <- c(50 , 40 , 60 , 70 , 30 , 65)
> Y   <- c(28 , 25 , 30 , 43 , 20 , 32)
> MX  <- mean(X); MX
[1] 52.5
> MY  <- mean(Y); MY
[1] 29.66667
> n   <- length(X) # entspricht auch length(Y)
>
> # Berechnung der Varianzen und Kovarianz
> # Varianz X
> var_X  <- (sum((X - MX)^2)) / n
> var_X
[1] 197.9167
> # Varianz Y
> var_Y  <- (sum((Y - MY)^2)) / n
> var_Y
[1] 50.22222
> # Kovarianz
> cov_XY <- 1/6 * (sum((X - MX) * (Y - MY)))
> cov_XY
[1] 90.83333
> # Alternativ dazu:
> cov_XY <- 1/6 * sum(X * Y) - MX * MY
> cov_XY
[1] 90.83333
> # b) Korrelation
> cor_XY <- cov_XY / (sqrt(var_X) * sqrt(var_Y))
> cor_XY
[1] 0.9110791
>
> # c) Berechnung von Achsenabschnitt (beta0) und Anstiegsparameter (beta1)
> beta1  <- cov_XY / var_X; beta1
[1] 0.4589474
> beta0  <- MY - beta1 * MX; beta0
[1] 5.57193
> # oder
> ols <- lm(Y ~ X); ols

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     5.5719       0.4589

>
> # Berechnung des Bestimmtheitsmaßes R2
> R2 <- (cor_XY)^2; R2
[1] 0.8300652
>
> # d) Berechnung der Fehlerterme u
> y_dach <- ols$fitted.values # prognostizierte Werte des Regressionsmodells
> err <- Y - y_dach  # Fehlerterme
> options(digits = 5)

> data.frame(X, Y, y_dach, err)
   X  Y y_dach      err
1 50 28 28.519 -0.51930
2 40 25 23.930  1.07018
```

```
3 60 30 33.109 -3.10877
4 70 43 37.698  5.30175
5 30 20 19.340  0.65965
6 65 32 35.404 -3.40351
>
> # e) Streudigramm
> plot(X, Y, cex = 2,
+ ylab = "Ausgaben für Kosmetik in € pro Monat",
+ xlab = "Einkommen in T€ pro Jahr",
+ main = "Haushaltseinkommen und Ausgaben für Kosmetik")
> abline(lm(Y~X))
> grid()
```

# Kapitel 6: Zufall und Wahrscheinlichkeit

### Aufgabe 6.1: Würfel

*Ein Würfel wird zweimal hintereinander geworfen.*

*a) Geben Sie jedes Element des Ereignisraums explizit an.*

*b) Wie groß ist die Wahrscheinlichkeit, dass die Augensumme 8 oder größer ist, wenn beim ersten Wurf eine 5 erscheint?*

c) *… wenn bei mindestens einem Wurf eine 4 erscheint?*

**Lösung a)**

Die Elemente des Ereignisraums $S$ sind:

$S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),$
$(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),$
$(3,1),(3,2), (3,3), (3,4), (3,5), (3,6),$
$(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),$
$(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),$
$(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$

**Lösung b)**

Gesucht ist die bedingte Wahrscheinlichkeit für:

$\mathrm{P}(\text{Augensumme } \geq 8 \mid 1.\,\text{Wurf "5"}) \implies \mathrm{P}(\mathrm{A}|\mathrm{B})$

Wir definieren: Ereignis $A$ = Augensumme $\geq 8$

Die bedingte Wahrscheinlichkeit berechnet sich mit: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Ereignis $B$ = $1.\,\text{Wurf eine "5"}$

Es gibt 6 Elemente des Ereignisraums, die $B$ erfüllen:

$\mathrm{B} = \{(5,1), (5,2), (5,3), (5,4), (5,5), (5,6)\}$

Wahrscheinlichkeit für das Ereignis $B$: $P(B) = \frac{6}{36} = \frac{1}{6}$

Ereignis $A \cap B$ = Augensumme 8 oder größer und 1. Wurf eine "5"

Es gibt 4 Elemente des Ereignisraums, die A & B erfüllen:

$\mathrm{A} \cap \mathrm{B} = \{(5,3), (5,4), (5,5), (5,6)\}$

Wahrscheinlichkeit für das Ereignis A&B: $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \frac{4}{36} = \frac{1}{9}$

Gesuchte bedingte Wahrscheinlichkeit:

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/9}{1/6} = \frac{6}{9} = \frac{2}{3}$

Die unbedingte Wahrscheinlichkeit für A ist dagegen kleiner (da die Bedingung den Ereignisraum einschränkt): $P(A|B) > P(A)$.

Es gibt 15 Elemente des Ereignisraums, die $A$ erfüllen:
$A = \{(2,6),(3,5),(3,6),(4,4),(4,5),(4,6),(5,3),(5,4),(5,5),(5,6),(6,2),(6,3),(6,4),(6,5),(6,6)\}$
Wahrscheinlichkeit für das Ereignis $A$:
$P(A) = \frac{Anzahl\ der\ günstigen\ Ausgänge}{Anzahl\ der\ möglichen\ Ausgänge} = \frac{15}{36} = \frac{5}{12}$

**Lösung c)**
Ereignis $B$: Bei mindestens einem Wurf eine 4.
Es gibt 11 Elemente des Ereignisraums, die $B$ erfüllen:
$B = \{(4,1),(4,2),(4,3),(4,5),(4,6),(1,4),(2,4),(3,4),(4,4),(5,4),(6,4)\}$
Wahrscheinlichkeit für das Ereignis $B$: $P(B) = \frac{11}{36} = 0.306$
Ereignis $A \cap B$: Augensumme ist mindestens 8 und bei mindestens einem Wurf eine 4
Elemente des Ereignisraums, die A&B erfüllen:
$A \cap B = \{(4,4),(4,5),(5,4),(4,6),(6,4)\}$
Wahrscheinlichkeit für das Ereignis A&B:
P(A&B) = 5/36 = 0.139
Gesuchte bedingte Wahrscheinlichkeit: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{5/36}{11/36} = \frac{5}{11} = 0.454$

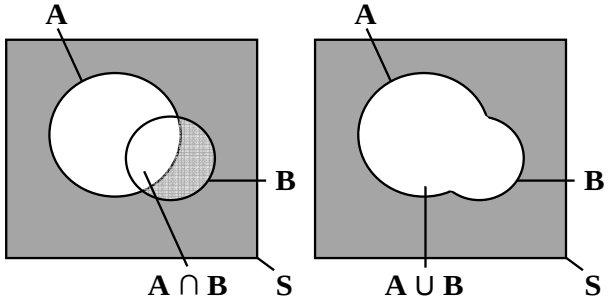
Die unbedingte Wahrscheinlichkeit für A ist dagegen kleiner (da die Bedingung den Ereignisraum einschränkt): $P(A|B) > P(A)$.
Es gibt 15 Elemente des Ereignisraums, die $A$ erfüllen:
$A = \{(2,6),(3,5),(3,6),(4,4),(4,5),(4,6),(5,3),(5,4),(5,5),(5,6),(6,2),(6,3),(6,4),(6,5),(6,6)\}$
Die unbedingte Wahrscheinlichkeit für das Ereignis $A$ ist daher $P(A) = \frac{15}{36}$

**Aufgabe 6.2: Wahrscheinlichkeiten**

*Für die Ereignisse A und B seien die Wahrscheinlichkeiten gegeben:*
$P(A) = 0.6, \quad P(B) = 0.2, \quad P(A \cap B) = 0.1.$
*Geben Sie – wenn möglich – die folgenden Wahrscheinlichkeiten zahlenmäßig an.*
*a)* $P(A \cup B)$ *b)* $P(A|B)$ *c)* $P(\overline{A \cap B})$ *d)* $P(\overline{A \cup B})$
*Hinweis: Nutzen Sie Venn-Diagramme zur Darstellung der Wahrscheinlichkeiten.*

Folgende Werte sind gegeben:
$P(A) = 0.6, P(B) = 0.2, P(A \cap B) = 0.1$

**Lösung a)**
$P\ (A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.2 - 0.1 = 0.7$
Die Wahrscheinlichkeit für $A \cup B$ beträgt 70%.
Die folgenden beiden Abbildungen zeigen das Problem graphisch (allerdings nicht maßstabsgetreu).



**Lösung b)**

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.2} = 0.5$

Die Wahrscheinlichkeit für $A|B$ beträgt 50%.

**Lösung c)**

$P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - 0.1 = 0.9$

Die Wahrscheinlichkeit für $\overline{A \cap B}$ beträgt 90%.

**Lösung d)**

$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.7 = 0.3$

Die Wahrscheinlichkeit für $\overline{A \cup B}$ beträgt 30%.

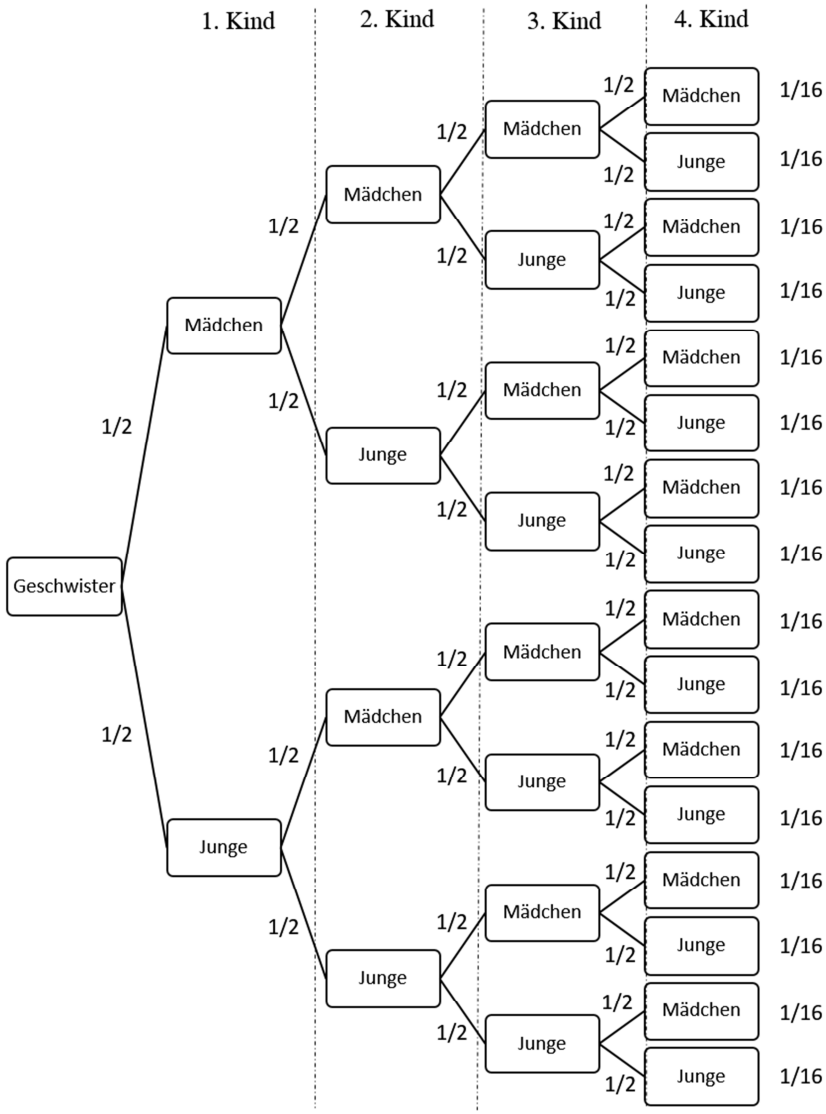
**Aufgabe 6.3: Geschwister**

*Wie groß ist die Wahrscheinlichkeit, dass von vier Geschwistern*

*a) alle vier Jungen sind,*
*b) alle vier Mädchen sind,*
*c) das älteste ein Junge, die folgenden Mädchen sind,*
*d) die drei älteren Jungen sind und das jüngste ein Mädchen ist,*
*e) zwei Jungen und zwei Mädchen sind?*

*Hinweis: Lösen Sie die Aufgabe mittels eines Baumdiagramms.*

**Baumdiagramm**
Die Abbildung zeigt das Baumdiagramm für dieses 4-stufige Zufallsexperiment.




**Lösung a)**
$P(Alle\ vier\ sind\ Jungen) = (\frac{1}{2})^4 = \frac{1}{16} = 0.0625$
Die Wahrscheinlichkeit, dass alle vier Jungen sind, ist 6.25%. Dies ist die Eintrittswahrscheinlichkeit für den untersten Pfad in der Abbildung oben.

**Lösung b)**
$P(Alle\ vier\ sind\ Mädchen) = (\frac{1}{2})^4 = \frac{1}{16} = 0.0625$. Dies ist die Eintrittswahrscheinlichkeit für den obersten Pfad in der Abbildung oben.

**Lösung c)**
Ereignis $C$: Das älteste ist ein Junge, die folgenden sind Mädchen:
$P(C) = (\frac{1}{2})^4 = \frac{1}{16} = 0.0625$
Die Wahrscheinlichkeit, dass das älteste ein Junge und die folgenden Mädchen sind, ist ebenfalls 6.25%.

**Lösung d)**
Ereignis $D$: Die drei ältesten sind Jungen, das jüngste ist ein Mädchen:
$P(D) = (\frac{1}{2})^4 = \frac{1}{16} = 0.0625$
Die Wahrscheinlichkeit, dass die drei ältesten Jungen sind und das Jüngste ein Mädchen ist, ist auch 6.25%.

**Lösung e)**
Ereignis $E$: Die Wahrscheinlichkeit für zwei Jungen und zwei Mädchen:
$E = \{(JJMM), (MMJJ), (JMMJ), (JMJM), (MJMJ), (MJJM)\}$
$P(E) = 6 \times \frac{1}{16} = \frac{6}{16} = 0.375$
Die Wahrscheinlichkeit, dass es zwei Jungen und zwei Mädchen sind, ist 37.5%. Diese Wahrscheinlichkeit ergibt sich über die Summe der Wahrscheinlichkeiten für die einzelnen Pfade.

**Aufgabe 6.4: Gerät**

*Ein technisches Gerät G wird aus drei Einzelteilen A, B, C zusammengefügt. Das Gerät funktioniert nur dann, wenn alle drei Einzelteile funktionieren und außerdem bei der Montage M keine Fehler unterlaufen. Die Wahrscheinlichkeiten, dass die Teile A, B, C defekt sind, betragen $P(\bar{A}) = P(\bar{B}) = 0.02$, $P(\bar{C}) = 0.04$. Die Wahrscheinlichkeit, dass bei der Montage ein Fehler gemacht werde, betrage $P(\bar{M}) = 0.03$. Die Fehler treten unabhängig voneinander auf.*

*a) Wie groß ist die Wahrscheinlichkeit $P(G) = (A \cap B \cap C \cap M)$, dass das Gerät funktioniert?*
*b) Wie groß ist die Wahrscheinlichkeit $P(\bar{G})$, dass das Gerät nicht funktioniert?*

**Lösung a)**
Ereignis $G$: Das Gerät funktioniert nach der Montage:
$P(G) = (A \cap B \cap C \cap M) = P(A) * P(B) * (C) * P(M)$
$P(G) = (1 - 0.02) * (1 - 0.02) * (1 - 0.04) * (1 - 0.03) = 0.894$
Die Wahrscheinlichkeit für ein funktionierendes Gerät beträgt 89.4%.

**Lösung b)**
Gegenereignis $\bar{G}$: Das Gerät funktioniert nicht.
$P(\overline{G}) = 1 - 0.894 = 0.106$
Die Wahrscheinlichkeit für ein fehlerhaftes Gerät beträgt 10.6%.

**Aufgabe 6.5: Urne**

*In einer Urne befindet sich eine große Zahl von Kugeln. 70% der Kugeln sind weiß, 30% sind schwarz.*

*a) Wie groß ist die Wahrscheinlichkeit, eine schwarze Kugel bei blindem Hineingreifen zu ziehen?*
*b) Wie groß ist die Wahrscheinlichkeit, bei drei zufälligen Stichproben mit Zurücklegen genau drei weiße Kugeln zu finden?*
*c) Wie groß ist die Wahrscheinlichkeit, bei drei zufälligen Stichproben mit Zurücklegen genau drei schwarze Kugeln zu finden?*
*d) Wie groß ist die Wahrscheinlichkeit, bei drei zufälligen Stichproben mit Zurücklegen eine schwarze und zwei weiße Kugeln zu finden?*

*Hinweis: Vgl. Lösungsansatz bei Aufgabe 6.3.*

**Lösung a)**
Wahrscheinlichkeit für das Ziehen einer schwarzen Kugel bei einmaligem Ziehen mit Zurücklegen: $P(A) = 0.3$
Die Wahrscheinlichkeit beträgt also 30%.

**Lösung b)**
Wahrscheinlichkeit für das Ziehen von drei weißen Kugeln bei dreimaligem Ziehen mit Zurücklegen: $P(B) = 0.7 * 0.7 * 0.7 = 0.343$
Die Wahrscheinlichkeit beträgt 34.3%.

**Lösung c)**
Wahrscheinlichkeit für das Ziehen von drei schwarzen Kugeln bei dreimaligem Ziehen mit Zurücklegen: $P(C) = 0.3 * 0.3 * 0.3 = 0.027$
Die Wahrscheinlichkeit beträgt 2.7%.

**Lösung d)**
Wahrscheinlichkeit für das Ziehen einer schwarzen und dreier weißer Kugeln bei dreimaligem Ziehen mit Zurücklegen:
$P(D) = \{(\text{SWW}), (\text{WWS}), (\text{WSW})\}$
$P(D) = (0.3 * 0.7 * 0.7) + (0.7 * 0.7 * 0.3) + (0.7 * 0.3 * 0.7)$
$P(D) = 0.441$
Die Wahrscheinlichkeit beträgt 44.1%.

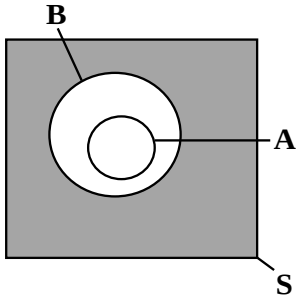
**Aufgabe 6.6: Bedingte Wahrscheinlichkeiten**

***Bedingte Wahrscheinlichkeiten.*** *Treffen Sie eine Aussage über die bedingte Wahrscheinlichkeit $P(A|B)$, wenn*

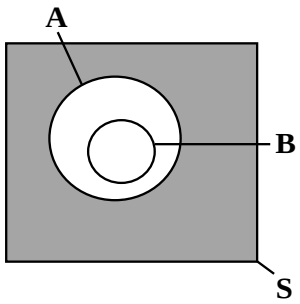
*a) A eine Teilmenge von B ist,*
*b) B eine Teilmenge von A ist,*
*c) A und B sich gegenseitig ausschließen.*

*Hinweis: Nutzen Sie Venn-Diagramme zur Darstellung der Wahrscheinlichkeiten.*

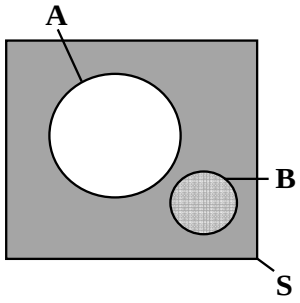
**Lösung a)**
Da $A$ eine Teilmenge von $B$ ist, gilt für die bedingte Wahrscheinlichkeit:
$P(A|B) = \frac{P(A \cap B)}{P(B)} < 1$
Zur Veranschaulichung dient das folgende Venn-Diagramm.




**Lösung b)**
$B$ ist Teilmenge von $A$:
$P(A|B) = \frac{P(A \cap B)}{P(B)} = 1$
Zur Veranschaulichung dient das folgende Venn-Diagramm.




**Lösung c)**
Wenn $A$ und $B$ sich gegenseitig ausschließen, ist $P(A \cap B) = 0$. Somit ergibt sich für die bedingte Wahrscheinlichkeit: $P(A|B) = 0$. Vgl. das folgende Venn-Diagramm.




**Aufgabe 6.7: Glücksspiel**

*Man bietet Ihnen das folgende Glücksspiel an: Sie bezahlen einen Einsatz von X Cent, dann dürfen Sie dreimal einen Würfel werfen. Werfen Sie mindestens einmal eine 6, erhalten Sie 1€.*

*a) Wie groß ist die Wahrscheinlichkeit, dass Sie bei diesem Spiel gewinnen?*
*b) Sie dürfen das Spiel hinreichend oft wiederholen. Welchen Einsatz X wären Sie höchstens bereit zu zahlen, um langfristig zu gewinnen?*

**Lösung a)**
Die Wahrscheinlichkeit zu gewinnen ($G$) ist gleich der Wahrscheinlichkeit, mindestens eine 6 zu würfeln. Diese Wahrscheinlichkeit ergibt sich als Gegenwahrscheinlichkeit zum Ereignis, bei drei Versuchen 0 x eine 6 zu würfeln.

$$P(keine\ 6\ bei\ 3\ Versuchen) = \left(\frac{5}{6}\right)^3 = 0.5787$$

Dann ist

$$P(G) = 1 - \left(\frac{5}{6}\right)^3 = \frac{91}{216} = 0.421$$

Die Wahrscheinlichkeit zu gewinnen beträgt 42.1%.

**Lösung b)**
Der zu erwartende Gewinn $G$ müsste den Einsatz $X$ übersteigen. Langfristig gewinnt man bei diesem Spiel mit einer Wahrscheinlichkeit von 42.1% den Betrag 1€. Deshalb gewinnt man im Durchschnitt mindestens 0.421 € im Spiel. Daraus ergibt sich, dass höchstens ein Einsatz von 0.42 € pro Spiel anzusetzen ist, um langfristig zu gewinnen.

**Lösung mit R**

```
> # ?pbinom Hilfe zur Binomialverteilung
> # vgl. Kap. 7 im Buch
> # in R
> (5/6)^3
```

```
[1] 0.5787037
> # Massefunktion der Binomialverteilung
> # bei n = 3 und p = 1/6
> dbinom(0:3, 3, 1/6)
[1] 0.57870370 0.34722222 0.06944444 0.00462963
> # W'keit für X = 1, 2 oder 3 Erfolge für
> # Massefunktion der Binomialverteilung
> # bei n = 3 und p = 1/6
> sum(dbinom(1:3, 3, 1/6))
[1] 0.4212963
```

**Aufgabe 6.8: Promotion**

*Bei einer Verkaufsaktion platziert der Hersteller Gewinnsymbole unter die Flaschendeckel von 10% aller Flaschen.*

*a) Angenommen, Sie kaufen drei Flaschen. Wie groß ist die Wahrscheinlichkeit etwas zu gewinnen?*

*b) Wenn Sie ein Six-Pack kaufen, mit welcher Wahrscheinlichkeit gewinnen Sie etwas?*

*Hinweis: Es genügt, die Wahrscheinlichkeit zu ermitteln, mit der Sie nicht gewinnen.*

**Lösung a)**

Vorüberlegungen: Die Gewinnwahrscheinlichkeit für eine Flasche ist 0.1, dementsprechend ist die Wahrscheinlichkeit bei einer Flasche nichts zu gewinnen gleich 0.9.

Ereignis $\bar{A}$: Man hat bei drei Versuchen keinen Gewinn:

$P(\bar{A}) = 0.9^3 = 0.729$

Ereignis $A$: Man hat bei drei Versuchen mindestens einen Gewinn:

$P(A) = 1 - P(\bar{A}) = 1 - 0.9^3 = 0.271$

Die Wahrscheinlichkeit für mindestens einen Gewinn beträgt 27.1%.

**Lösung b)**

Ereignis $B$: Mindestens einen Gewinn beim Kauf eines Six-Packs (6 Versuche):

$P(B) = 1 - 0.9^6 = 0.47$

Die Wahrscheinlichkeit für mind. einen Gewinn beim Kauf eines Sixpacks ist 47%.

**Lösung mit R**

```
> # ?pbinom Hilfe zur Binomialverteilung
> # vgl. Kap. 7 im Buch
> # in R
> 0.9^3
[1] 0.729
> 1 - 0.9^3
[1] 0.271
>
> # Massenfunktion der Binomialverteilung
> # bei n = 3 und p = 0.1
> dbinom(0:3, 3, 0.1)
[1] 0.729 0.243 0.027 0.001
>
> # W'keit für X = 1, 2 oder 3 Erfolge für
> # Massenfunktion der Binomialverteilung
> # bei n = 3 und p = 0.1
> sum(dbinom(1:3, 3, 0.1))
[1] 0.271
```

**Aufgabe 6.9: Beratungsfirma I**

*Angenommen, Sie arbeiten für eine große Beratungsfirma. Von allen Mitarbeitern im Beratungsgeschäft der Firma haben 55% keine Erfahrung in der Telekommunikationsindustrie, 32% haben etwas Erfahrung (< 5 Jahre) und der Rest hat große Erfahrung (≥ 5 Jahre). Sie und zwei andere Analysten werden zufällig ausgewählt in ein Team, welches ein Projekt in der Telekommunikation bearbeiten soll.*
*Mit welcher Wahrscheinlichkeit hat der erste Kollege im Team, dem Sie begegnen,*

*a) große Erfahrung im Bereich Telekommunikation?*
*b) zumindest etwas Erfahrung im Bereich Telekommunikation?*
*c) nicht mehr als etwas Erfahrung im Bereich Telekommunikation?*

**Lösung a)**
Die Wahrscheinlichkeit für einen Mitarbeiter mit großer Erfahrung ($E$):
$P(große\ E) = 1 - P(keine\ E) - P(etwas\ E) = 1 - 0.55 - 0.32 = 0.13$
Die Wahrscheinlichkeit, dass der erste Kollege, dem Sie im Team begegnen, große Erfahrung im Bereich Telekommunikation besitzt, beträgt 13%.

**Lösung b)**
Die Wahrscheinlichkeit für einen Mitarbeiter mit zumindest etwas Erfahrung:
$P(zumindest\ etwas\ E) = P(etwas\ E) + P(große\ E) = 0.32 + 0.13 = 0.45$
Die Wahrscheinlichkeit, dass der erste Kollege, dem Sie im Team begegnen, zumindest etwas Erfahrung im Bereich Telekommunikation besitzt, beträgt 45%.

**Lösung c)**
Die Wahrscheinlichkeit für einen Mitarbeiter mit höchstens etwas Erfahrung:
$P(höchstens\ etwas\ E) = P(keine\ E) + P(etwas\ E) = 0.55 + 0.32 = 0.87$
Die Wahrscheinlichkeit, dass der erste Kollege im Team dem Sie begegnen, höchstens etwas Erfahrung im Bereich Telekommunikation besitzt, beträgt 87%.

**Aufgabe 6.10: Beratungsfirma II**

*Gleiches Szenario wie in der Aufgabe zuvor. Mit welcher Wahrscheinlichkeit hat/haben von den zwei Ihrer Kollegen im Team*

*a) keiner Erfahrung im Bereich Telekommunikation?*
*b) beide etwas Erfahrung im Bereich Telekommunikation?*
*c) zumindest einer große Erfahrung im Bereich Telekommunikation?*
*d) Sie nutzen eine bestimmte Rechenregel zur Bestimmung der Wahrscheinlichkeiten. Was muss zur Anwendung dieser Regel gelten? Erklärung.*

**Lösung a)**
Die Wahrscheinlichkeit für zwei Mitarbeiter ohne Erfahrung (d.h. beide keine Erfahrung):
$P(A) = 0.55 * 0.55 = 0.3025$
Die Wahrscheinlichkeit, dass zwei Kollegen im Team keine Erfahrung im Bereich Telekommunikation besitzen, beträgt 30.25%.

**Lösung b)**
Die Wahrscheinlichkeit für zwei Mitarbeiter mit etwas Erfahrung:
$P(B) = 0.32 * 0.32 = 0.1024$
Die Wahrscheinlichkeit, dass zwei Kollegen im Team etwas Erfahrung im Bereich Telekommunikation besitzen, beträgt 10.24%.

**Lösung c)**
Die Wahrscheinlichkeit für einen Mitarbeiter mit großer Erfahrung ist:
$P(große\ E) = 0.13$
Die Wahrscheinlichkeit, dass ein Mitarbeiter keine große Erfahrung hat, ist:
$P(keine\ große\ E) = P(\overline{große\ E}) = 1 - 0.13 = 0.87$
Die Wahrscheinlichkeit, dass zwei Mitarbeiter keine große Erfahrung haben, ist:
$P(beide\ keine\ große\ E) = 0.87 * 0.87 = 0.7569$
Die gesuchte Wahrscheinlichkeit ist die Gegenwahrscheinlichkeit zu diesem Ereignis, also
$P(C) = 1 - 0.7569 = 0.2431$
Es gibt natürlich auch die Möglichkeit, die gesuchte Wahrscheinlichkeit über die Wahrscheinlichkeiten für die möglichen Kombinationen von Ereignissen auszurechen.
Die Wahrscheinlichkeit, dass zumindest ein Mitarbeiter große Erfahrung besitzt, ist:
$P(C) = (0.55 * 0.13 + 0.32 * 0.13) * 2 + 0.13 * 0.13 = 0.2431$
Die Wahrscheinlichkeit, dass zumindest ein Kollege im Team etwas Erfahrung im Bereich Telekommunikation besitzt, beträgt 24.31%.

**Lösung d)**
Der Multiplikationssatz gilt nur für Ereignisse, die stochastisch unabhängig voneinander sind, d.h., wenn z.B. ein erfahrener Kollege im Team ist, hat das keine Auswirkung auf die Wahrscheinlichkeit, dass der andere Kollege ebenfalls erfahren ist. In einem großen Unternehmen mit vielen Mitarbeitern kann diese Bedingung als erfüllt angesehen werden.

**Aufgabe 6.11: Immobilien**

*Nach Informationen einer Immobilienfirma haben 70% aller zum Verkauf stehender Häuser eine Garage, 15% haben einen Pool und 10% haben beides. Wie groß ist die Wahrscheinlichkeit, dass ein zum Verkauf stehendes Haus:*

*a) einen Pool oder eine Garage hat?*
*b) weder einen Pool noch eine Garage hat?*
*c) einen Pool aber keine Garage hat?*
*d) Wenn ein Haus zum Verkauf eine Garage hat, wie groß ist die Wahrscheinlichkeit, dass es auch einen Pool hat?*
*e) Sind die Eigenschaften „Garage“ und „Pool“ unabhängige Ereignisse? Erklärung.*
*f) Sind die Eigenschaften „Garage“ und „Pool“ sich gegenseitig ausschließende Ereignisse? Erklärung.*

*Hinweis: Erstellen Sie eine Kontingenztabelle oder ein Venn-Diagramm.*

Aus den Angaben lässt sich die folgende Kontingenztabelle konstruieren:

| $h_{ij}$ | Pool | | |
|---|---|---|---|
| Garage | Ja | Nein | $\sum$ |
| Ja | 0.10 | 0.60 | 0.70 |
| Nein | 0.05 | 0.25 | 0.30 |
| $\sum$ | 0.15 | 0.85 | 1.00 |

**Lösung a)**
Gesucht ist die Vereinigung der Ereignisse $P$ und $G$ (Definition des logischen „oder“, also $\cup$).
Es gilt: Das Ereignis $P$ oder $G$ tritt genau dann ein, wenn Ereignis $P$ oder Ereignis $G$ oder beide zugleich eintreten.
Die Wahrscheinlichkeit für einen Pool oder eine Garage ist:
$P(P \cup G) = 0.05 + 0.60 + 0.10 = 0.75$

Wir haben es hier mit disjunkten Ereignissen zu tun, also können wir die Wahrscheinlichkeiten für die drei Ereignisse (entweder Pool, entweder Garage, Pool und Garage) addieren.
Man könnte die Aufgabe auch so verstehen: Wie groß ist die Wahrscheinlichkeit für das alleinige Auftreten von $P$ und $G$? Da es sich um disjunkte Ereignisse handelt, können wir diese Wahrscheinlichkeiten direkt aus der Kontingenztabelle entnehmen mit:
$P(P_{ja}, G_{nein}) + P(P_{nein}, G_{ja}) = 0.05 + 0.60 = 0.65$
Hier ist aber nur die Lösung mit dem logischen „oder“ richtig.

**Lösung b)**
Die Wahrscheinlichkeit für keinen Pool und keine Garage:
$P(\overline{P \cup G}) = 1 - 0.75 = 0.25$
Das ist also die Gegenwahrscheinlichkeit zum Ereignis in a). Alternativ lässt sich die gesuchte Wahrscheinlichkeit auch aus der Kontingenztabelle unter $P(Garage\ nein \cap Pool\ nein) = 0.25$ ablesen.

**Lösung c)**
Die Wahrscheinlichkeit für einen Pool aber keine Garage:
$P(Pool \cap keine\ Garage) = 0.05$

**Lösung d)**
Die Wahrscheinlichkeit für Pool unter der Bedingung, dass das Haus eine Garage hat, ist:
$P(P|G) = \frac{P(P \cap G)}{P(G)} = \frac{0.1}{0.7} = 0.1429$

**Lösung e)**
Variante 1: Vergleichen der Randverteilung mit der bedingten Verteilung eines Merkmals. Hier sind die bedingten Verteilungen der Merkmals Pool dargestellt. Die Bedingungen sind Garage = Ja und Garage = Nein.

| $h_{j\|i}$ | Pool | | |
|---|---|---|---|
| Garage | Ja | Nein | $\sum$ |
| Ja | 0.1429 | 0.8571 | 1 |
| Nein | 0.1667 | 0.8333 | 1 |

Die bedingten Verteilungen des Merkmals Pool stimmen mit der Randverteilung {0.15, 0.85} fast überein.
Variante 2: Vergleichen der Randverteilung mit der bedingten Verteilung eines Merkmals. Hier sind die bedingten Verteilungen der Merkmals Garage dargestellt. Die Bedingungen sind Pool = Ja und Pool = Nein.

| $h_{i\|j}$ | Pool | |
|---|---|---|
| Garage | Ja | Nein |
| Ja | 0.6667 | 0.7059 |
| Nein | 0.3333 | 0.2941 |
| | 1 | 1 |

Die bedingten Verteilungen des Merkmals Garage stimmen mit der Randverteilung {0.70, 0.30} fast überein.
Ergebnis: Die beiden Merkmale Garage und Pool sind stochastisch unabhängig.

**Lösung f)**
Ereignisse, die sich gegenseitig ausschließen (also nicht gemeinsam auftreten), werden als disjunkte Ereignisse bezeichnet. Es muss gelten $P(A \cap B) = 0$
Hier ist die Wahrscheinlichkeit $P(Pool \cap Garage) = 0.1 > 0$. Pool und Garage sind somit nicht disjunkt.

**Erstellung der Kontingenztabelle mit R**

```
> # Erstellen einer Kontingenztabelle
> Garage <- c(rep("Ja", 70), rep("Nein", 30))
> Pool <- c(rep("Ja", 10), rep("Nein", 60), rep("Ja", 5), rep("Nein", 25))
> addmargins(prop.table(table(Garage,Pool)))
      Pool
Garage   Ja Nein  Sum
  Ja   0.10 0.60 0.70
  Nein 0.05 0.25 0.30
  Sum  0.15 0.85 1.00
>
> # Tabelle der bedingten Verteilungen
> options(digits=4)
> # Bedingung = Zeile = 1
> addmargins(prop.table(table(Garage,Pool),1))
      Pool
Garage     Ja   Nein    Sum
  Ja   0.1429 0.8571 1.0000
  Nein 0.1667 0.8333 1.0000
  Sum  0.3095 1.6905 2.0000
> # Bedingung = Spalte = 2
> addmargins(prop.table(table(Garage,Pool),2))
      Pool
Garage     Ja   Nein    Sum
  Ja   0.6667 0.7059 1.3725
  Nein 0.3333 0.2941 0.6275
  Sum  1.0000 1.0000 2.0000
```

**Aufgabe 6.12: Marktforschung**

*Marktforschungsinstitute kontaktieren bei Befragungen oft zufällig ausgewählte Teilnehmer. Nach Angaben eines Marktforschungsinstitutes lag die Kontaktrate (= Wahrscheinlichkeit einen ausgewählten Teilnehmer zu erreichen) im Jahr 2007 bei 67% und im Jahr 2012 bei 76%. Von den Befragten, die erfolgreich kontaktiert wurden, waren im Jahr 2007 58% bereit an der Befragung teilzunehmen und im Jahr 2012 nur 38%.*

*a) Wie groß ist die Wahrscheinlichkeit im Jahr 2012, dass ein erfolgreiches Interview mit einer zufällig ausgewählten Person zustande kommt? Dabei muss die Person kontaktiert werden und der Befragung zustimmen.*
*b) Was ist wahrscheinlicher: Ein erfolgreiches Interview mit einer zufällig ausgewählten Person im Jahr 2007 oder im Jahr 2012?*
*c) Mit welcher Wahrscheinlichkeit wird im Jahr 2012 eine zufällig ausgewählte Person kontaktiert aber lehnt es ab, an der Befragung teilzunehmen?*
*d) Mit welcher Wahrscheinlichkeit wird im Jahr 2012 eine zufällig ausgewählte Person nicht kontaktiert oder wird kontaktiert, lehnt es aber ab, an der Befragung teilzunehmen? Zeigen Sie zwei Wege zur Berechnung dieser Wahrscheinlichkeit.*

**Lösung a)**
Die Wahrscheinlichkeit für eine erfolgreiche Befragung im Jahr 2012:
$P(A) = 0.76 * 0.38 = 0.2888$

**Lösung b)**
Die Wahrscheinlichkeit für eine erfolgreiche Befragung im Jahr 2007:
$P(B) = 0.67 * 0.58 = 0.3886 > P(A)$
Die Wahrscheinlichkeit eine erfolgreiche Befragung im Jahr 2007 zu führen war mit 38.86% größer als im Jahr 2012 (28.88%).

**Lösung c)**
Die Wahrscheinlichkeit für eine abgelehnte Befragung im Jahr 2012:
$P(C) = P(Kontakt \cap keine\ Teilnahme) = 0.76 * 0.62 = 0.4712$
Mit einer Wahrscheinlichkeit von 47.12% wird im Jahr 2012 Kontakt hergestellt, aber keine Befragung geführt werden.

**Lösung d)**
Wahrscheinlichkeit für keinen Kontakt oder Kontakt, aber Ablehnung der Befragung (2012):
Möglichkeit 1: Über die Gegenwahrscheinlichkeit
$P(D) = 1 - P(Kontakt \cap Teilnahme) = 1 - 0.76 * 0.38 = \ 0.7112$
Möglichkeit 2: Über den Additionssatz
$P(D) = \big(1 - P(Kontakt)\big) + P(Kontakt \cap keine\ Teilnahme)$
$P(D) = (1 - 0.76) + 0.76 * (1 - 0.38) = 0.7112$
Die Wahrscheinlichkeit für keinen Kontakt oder Kontakt, aber Ablehnung der Befragung beträgt 71.12%.

**Aufgabe 6.13: GfK I**

*Die GfK befragt Konsumenten aus fünf Ländern danach, ob sie der Aussage "Ich betreibe täglich Gesichtspflege" zustimmen. Die Angaben sind in der Kontingenztabelle unten dargestellt.*

| $n_{ij}$ | Zustimmung | | | |
|---|---|---|---|---|
| Land | ja | nein | weiß nicht | $\sum$ |
| China | 361 | 988 | 153 | 1502 |
| Frankreich | 695 | 763 | 81 | 1539 |
| Indien | 828 | 689 | 18 | 1535 |
| UK | 597 | 898 | 62 | 1557 |
| USA | 668 | 841 | 48 | 1557 |
| $\sum$ | 3149 | 4179 | 362 | 7690 |

*Angenommen, wir wählen zufällig eine Person aus dieser Stichprobe aus.*

*a) Mit welcher Wahrscheinlichkeit wird diese Person der Aussage zustimmen?*
*b) Mit welcher Wahrscheinlichkeit stammt diese Person aus China?*
*c) Mit welcher Wahrscheinlichkeit stammt diese Person aus China und stimmt der Aussage zu?*
*d) Mit welcher Wahrscheinlichkeit stammt diese Person aus China oder stimmt der Aussage zu?*

**Lösung a)**
Die Wahrscheinlichkeit, dass die Person zustimmt:
$P(A) = \frac{3149}{7690} = 0.4095$

**Lösung b)**
Die Wahrscheinlichkeit, dass die Person aus China stammt:

$P(B) = \frac{1502}{7690} = 0.1953$

**Lösung c)**
Die Wahrscheinlichkeit, dass die Person aus China stammt und der Aussage zustimmt:
$P(C) = \frac{361}{7690} = 0.0469$

**Lösung d)**
Die Wahrscheinlichkeit, dass die Person aus China stammt oder der Aussage zustimmt:
$P(D) = \frac{1502}{7690} + \frac{3149}{7690} - \frac{361}{7690} = 0.1953 + 0.4095 - 0.0469 = 0.5579$

**Erstellung der Kontingenztabelle mit R**

```
> # Kontingenztabelle erstellen:
> Land <-   c(rep("China", 1502),  rep("Frankreich", 1539),
+             rep("Indien", 1535), rep("UK", 1557), rep("USA", 1557))
>
> Zustimmung <- c(rep("ja", 361), rep("nein", 988), rep("unschlüssig",153),
+         rep("ja", 695), rep("nein", 763), rep("unschlüssig", 81),
+         rep("ja", 828), rep("nein", 689), rep("unschlüssig", 18),
+         rep("ja", 597), rep("nein", 898), rep("unschlüssig", 62),
+         rep("ja", 668), rep("nein", 841), rep("unschlüssig", 48))
>
> options(digits=2)
> # Kontingenztabelle mit relativen Häufigkeiten


> addmargins(prop.table(table(Land, Zustimmung)))
           Zustimmung
Land             ja   nein unschlüssig    Sum
  China      0.0469 0.1285      0.0199 0.1953
  Frankreich 0.0904 0.0992      0.0105 0.2001
  Indien     0.1077 0.0896      0.0023 0.1996
  UK         0.0776 0.1168      0.0081 0.2025
  USA        0.0869 0.1094      0.0062 0.2025
  Sum        0.4095 0.5434      0.0471 1.0000
>
> # Kontingenztabelle mit absoluten Häufigkeiten
> addmargins(table(Land, Zustimmung))
           Zustimmung
Land           ja nein unschlüssig  Sum
  China       361  988         153 1502
  Frankreich  695  763          81 1539
  Indien      828  689          18 1535
  UK          597  898          62 1557
  USA         668  841          48 1557
  Sum        3149 4179         362 7690
```

**Aufgabe 6.14: GfK II**

*Betrachten Sie nochmals die Angaben zur GfK-Befragung aus der Aufgabe zuvor.*

a) *Wenn eine Person zufällig ausgewählt wird, mit welcher Wahrscheinlichkeit wird eine Person aus den USA gewählt, die zugleich der Aussage zustimmt?*
b) *Unter den Befragten aus den USA, mit welcher Wahrscheinlichkeit ist mit Zustimmung zur Aussage zu rechnen?*
c) *Mit welcher Wahrscheinlichkeit stammt ein Teilnehmer, der der Aussage zustimmt, aus den USA?*
d) *Sind Zustimmung und Herkunftsland unabhängige Merkmale?*

Die Kontingenztabelle mit relativen Häufigkeiten ist:

| $h_{ij}$ | Zustimmung | | | |
|---|---|---|---|---|
| Land | Ja | Nein | unschlüssig | $\sum$ |
| China | 0.0469 | 0.1285 | 0.0199 | 0.1953 |
| Frankreich | 0.0904 | 0.0992 | 0.0105 | 0.2001 |
| Indien | 0.1077 | 0.0896 | 0.0023 | 0.1996 |
| UK | 0.0776 | 0.1168 | 0.0081 | 0.2025 |
| USA | 0.0869 | 0.1094 | 0.0062 | 0.2025 |
| $\sum$ | 0.4095 | 0.5435 | 0.0470 | 1.0000 |

**Lösung a)**
Die Wahrscheinlichkeit, dass die Person aus den USA stammt und der Aussage zustimmt:
$P(A) = \frac{668}{7690} = 0.0869$

**Lösung b)**
Die Wahrscheinlichkeit, dass eine aus den USA stammende Person zustimmt:
$P(B) = P(Zustimmung|USA) = \frac{Zustimmung \cap USA}{USA} = \frac{668}{1557} = 0.4290$

**Lösung c)**
Die Wahrscheinlichkeit, dass eine der Aussage zustimmende Person aus den USA stammt:
$P(C) = P(USA|Zustimmung) = \frac{USA \cap Zustimmung}{Zustimmung} = \frac{668}{3149} = 0.2121$

**Lösung d)**
Gleiche Vorgehensweise wie bei Aufgabe 6.11 e). Vergleichen der Randverteilungen (Tabelle siehe oben) mit den bedingten Verteilungen beider Merkmale.

Tabelle der bedingten Verteilungen für das Spaltenmerkmal Zustimmung:

| $h_{j\|i}$ | Zustimmung | | | |
|---|---|---|---|---|
| Land | Ja | Nein | unschlüssig | $\sum$ |
| China | 0.24 | 0.66 | 0.10 | 1.00 |
| Frankreich | 0.45 | 0.50 | 0.05 | 1.00 |
| Indien | 0.54 | 0.45 | 0.01 | 1.00 |
| UK | 0.38 | 0.58 | 0.04 | 1.00 |
| USA | 0.43 | 0.54 | 0.03 | 1.00 |

Die beiden Merkmale sind nicht unabhängig, weil die bedingten Verteilungen nicht gleich der Randverteilung des Merkmals Zustimmung {0.41, 0.54, 0.05} sind, was sich z.B. deutlich im Falle von China vs. Frankreich zeigt.

**Erstellung der Kontingenztabelle für d) mit R und Berechnung von KK***

```
# Kontingenztabelle der bedingten Verteilungen
Land <-        c(rep("China", 1502), rep("Frankreich", 1539),
              rep("Indien", 1535), rep("UK", 1557), rep("USA", 1557))

Zustimmung <- c(rep("ja", 361), rep("nein", 988), rep("unschlüssig",153),
                rep("ja", 695), rep("nein", 763), rep("unschlüssig", 81),
                rep("ja", 828), rep("nein", 689), rep("unschlüssig", 18),
                rep("ja", 597), rep("nein", 898), rep("unschlüssig", 62),
                rep("ja", 668), rep("nein", 841), rep("unschlüssig", 48))
```

```
options(digits=1)
addmargins(prop.table(table(Land, Zustimmung),1))

> tab <- table(Land, Zustimmung)
# vgl. zum Test auf Unabhängigkeit Kap. 11
> chisq.test(table(Land, Zustimmung))

        Pearson's Chi-squared test

data:  table(Land, Zustimmung)
X-squared = 400, df = 8, p-value <2e-16

>
> # Kontingenzkoeffizient KK_star
> # ist mit 0.3 deutlich größer als Null => keine Unabhängigkeit
> m  <-  3     #  => kleine Zahl von Spaltenzahl und Zeilenzahl
> n  <- 7690   # Anzahl der Beobachtungen
> QK <- 400    # QK
> KK_star <- sqrt((QK / (QK + n)) * (m / (m - 1)))
> KK_star
[1] 0.3
```

# Kapitel 7: Zufallsvariablen und ausgewählte Verteilungen

**Aufgabe 7.1: Zwei Würfel**

*Es werden zwei Würfel geworfen. Die Zufallsvariable X sei definiert als die absolute Differenz der Augenzahlen beider Würfel.*

*a) Bestimmen Sie die Wahrscheinlichkeiten $P(X \leq 1)$ und $P(X > 4)$.*

*b) Auf welchen Voraussetzungen beruhen die von Ihnen verwendeten Wahrscheinlichkeiten?*

c) *Bestimmen Sie für diese diskrete Zufallsvariable die Massen- und Verteilungsfunktion. Zeichnen Sie beide Funktionen.*

**Lösung a)**

Die Elemente des Ereignisraumes *S* können folgendermaßen dargestellt werden:

$S$ = {(1,1), (2,1), (3,1), (4,1), (5,1), (6,1),
(1,2), (2,2), (3,2), (4,2), (5,2), (6,2),
(1,3), (2,3), (3,3), (4,3), (5,3), (6,3),
(1,4), (2,4), (3,4), (4,4), (5,4), (6,4),
(1,5), (2,5), (3,5), (4,5), (5,5), (6,5),
(1,6), (2,6), (3,6), (4,6), (5,6), (6,6)}

Die Zufallsvariable $X$ ist dabei definiert als $X$ = |Augenzahl 1. Wurf − Augenzahl 2. Wurf |
Die Werte lassen sich wie folgt berechnen:
Wahrscheinlichkeit $P(X \leq 1) = \frac{16}{36} = \frac{4}{9}$, da das Ereignis 16 Elementarereignisse besitzt:
$P(X \leq 1)$ = {(1,1), (2,1), (1,2), (2,2), (3,2), (2,3), (3,3), (4,3), (3,4), (4,4), (5,4), (4,5), (5,5), (6,5), (5,6), (6,6)}
Wahrscheinlichkeit $P(X > 4) = \frac{2}{36} = \frac{1}{18}$, da das Ereignis 2 Elementarereignisse besitzt:
$P(X > 4)$ = {(6,1), (1,6)}

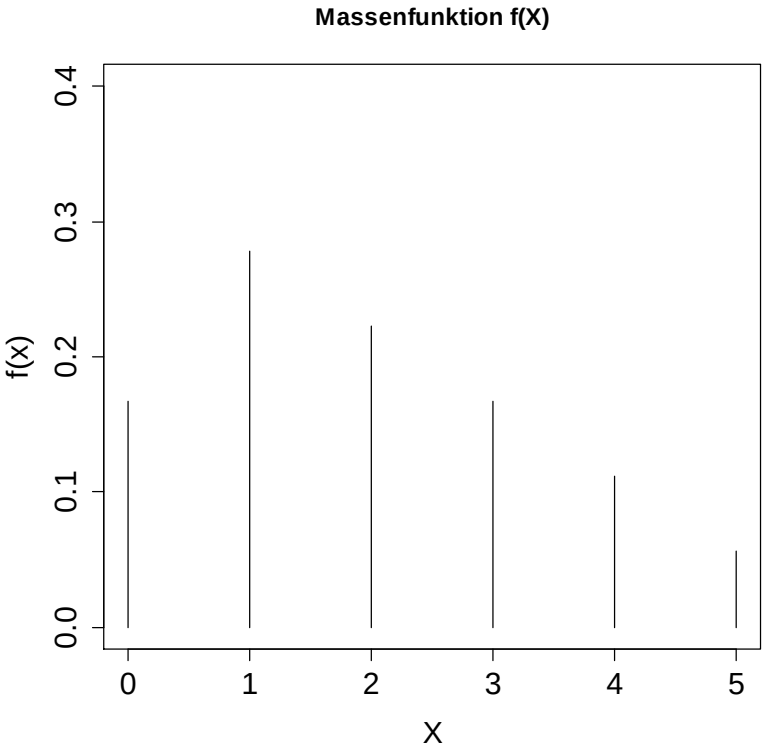
**Lösung b)**

Die Ereignisse (Ergebnis 1. Wurf, Ergebnis 2. Wurf) sind unabhängig. Mit anderen Worten, die Wahrscheinlichkeit für eine bestimmte Augenzahl im 2. Wurf wird nicht von der

Augenzahl im 1. Wurf beeinflusst. Daraus folgt z.B., dass $P(1.Wurf = 2, 2.Wurf = 3) = P(1.Wurf = 3, 2.Wurf = 2) = \frac{1}{36}$.

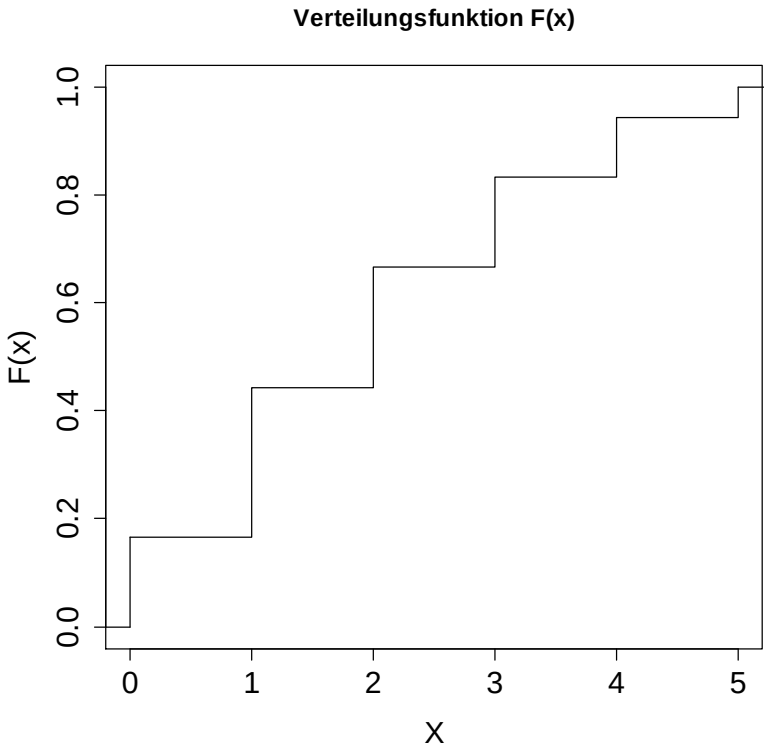
**Lösung c)**
Für die Massenfunktion gilt:

$$f(x) = \begin{cases} \frac{6}{36} \text{ für } x = 0 \\ \frac{10}{36} \text{ für } x = 1 \\ \frac{8}{36} \text{ für } x = 2 \\ \frac{6}{36} \text{ für } x = 3 \\ \frac{4}{36} \text{ für } x = 4 \\ \frac{2}{36} \text{ für } x = 5 \\ 0 \quad sonst \end{cases}$$




Für die Verteilungsfunktion $F$ gilt:

$$F(x) = \begin{cases} \frac{6}{36} f\text{ür } x = 0 \\ \frac{16}{36} f\text{ür } x = 1 \\ \frac{24}{36} f\text{ür } x = 2 \\ \frac{30}{36} f\text{ür } x = 3 \\ \frac{34}{36} f\text{ür } x = 4 \\ \frac{36}{36} f\text{ür } x = 5 \end{cases}$$




**Lösung mit R**

```
> # Zeichnen der Massenfunktion in R:
> X <- c(0:5) # Variable X definieren
> f <- c(6/36, 10/36, 8/36, 6/36, 4/36, 2/36) # Variable f definieren
> ecdf_f <- cumsum(f) # CDF berechnen
>
> plot(X, f, main = "Massenfunktion f(X)", cex.lab = 1.5, ylim = c(0, 0.4),
+ cex.axis = 1.5, type = "h", xlab = "X", ylab = "f(x)")
>
> # Zeichnen der Verteilungsfunktion in R:
> plot(X, ecdf_f, main = "Verteilungsfunktion F(x)",
+ ylim = c(0, 1), cex.lab = 1.5, cex.axis = 1.5,
+ type = "s",xlab = "X", ylab = "F(x)")
> segments(0,0,0,0.1667); segments(-1,0, 0, 0); segments(5,1,6,1)
>
> # Die Werte in a) lassen sich auch aus der Verteilungsfunktion ablesen:
> # Wahrscheinlichkeit P(X <= 1) = 16/36
> # Wahrscheinlichkeit P(X >  4) = 1 - 34/36 = 2/36
```

**Aufgabe 7.2: Mensch-ärgere-Dich-nicht**

*Um beim Mensch-ärgere-Dich-nicht herauskommen zu können („Erfolg“), muss man bei drei Würfen mindestens eine Sechs würfeln. Definieren Sie eine Zufallsvariable X mit „Augenzahl = 6“ bei $n = 3$ Würfen.*

*a) Ermitteln Sie die Massenfunktion der Zufallsvariablen.*

*b) Wie groß ist die Wahrscheinlichkeit, bei diesem Spiel herauszukommen, d.h. $P(X \geq 1)$?*

**Lösung a)**

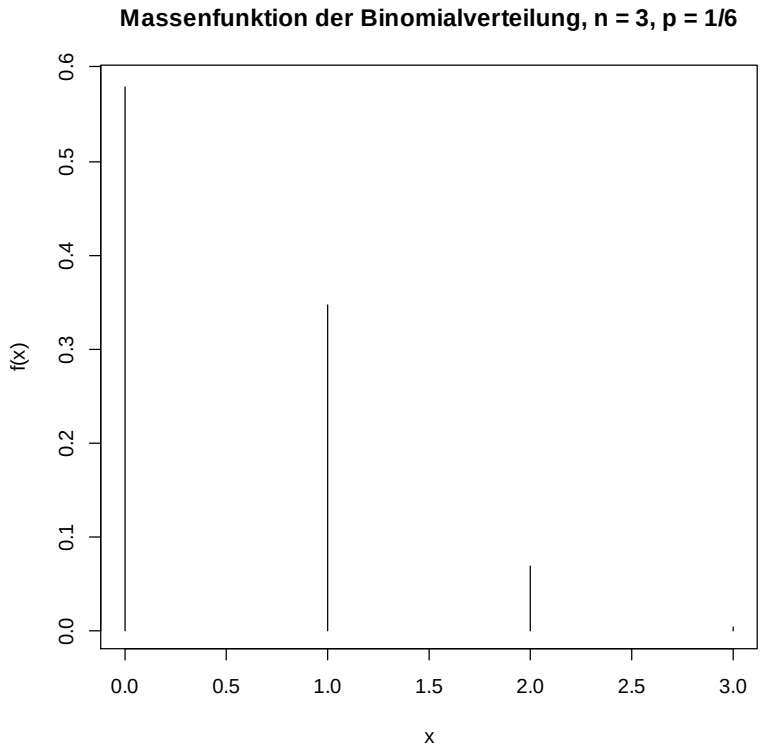
Hier liegt eine binomialverteilte Zufallsvariable vor, da mit $n$ Versuchen ein Bernoulli-Experiment mit konstanter Eintrittswahrscheinlichkeit $p$ durchgeführt wird.
Die Massenfunktion einer binomialverteilten Zufallsvariable ist

$$f_{Bi}(x,p,n) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{für } x = 0,1,2,\dots,n \\ 0 & sonst \end{cases}$$

wobei $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.
Hier ist $n = 3$ und $p = \frac{1}{6}$. Die Zufallsvariable $X$ ist die Zahl der Erfolge (6) beim Würfeln.
Für die Massenfunktion ergibt sich:

$$f(x) = \begin{cases} 0.5787 \text{ für } x = 0 \\ 0.3472 \text{ für } x = 1 \\ 0.0694 \text{ für } x = 2 \\ 0.0046 \text{ für } x = 3 \\ 0 \; sonst \end{cases}$$

Diese Werte müssen ausgerechnet werden, denn die Tabelle der Massenfunktion der Binomialverteilung im Buch (S. 292/293) weist nur Werte für $p = 0.15$ und $p = 0.2$ aus.
Die folgende Abbildung zeigt die Massenfunktion




**Lösung b)**
Die Wahrscheinlichkeit, mindestens eine Sechs zu würfeln, beträgt:
$P(X \geq 1) = 1 - 0.5787 = 0.4313$.

**Lösung mit R**

```
> # a) # Massenfunktion für Binomialverteilung
> # ?dbinom # Hilfe für die Binomialverteilung
> # dbinom(x, size, prob)
> # x = Zufallsvariable, size = n (Zahl der Versuche), prop = p
> dbinom(0:3, 3, 1/6)
[1] 0.57870 0.34722 0.06944 0.00463
>
> f <- dbinom(0:3, 3, 1/6) # Massenfunktion
> x <- c(0:3)              # x Variable definieren
>
> # gezeichnet:
> plot(x, f, main = "Massenfunktion der Binomialverteilung, n = 3, p = 1/6",
+ type = "h", xlab = "x", ylab = "f(x)")
```

**Aufgabe 7.3: Promotion**

*Ein Internet-Provider gibt bekannt, dass jeder 5. neue Nutzer zufällig ausgewählt wird und ein Geschenk erhält. Eine Gruppe von 10 Nachbarn beschließt, den neuen Internet-Service zu buchen.*

*a) Wie groß ist die Wahrscheinlichkeit, dass mindestens vier Teilnehmer der Gruppe das Geschenk erhalten?*

*b) Wie groß ist die Wahrscheinlichkeit, dass kein Teilnehmer der Gruppe das Geschenk bekommt?*

**Lösung a)**

Es liegt ein 10 x wiederholtes Bernoulli-Experiment mit der Erfolgswahrscheinlichkeit von $p = 0.2$ vor. Die Zufallsvariable „Zahl Erfolge“ $X$ ist somit binomialverteilt. Gesucht ist die Wahrscheinlichkeit, dass 4 oder mehr Teilnehmer gewinnen, also

$$P(X \geq 4) = P(x = 4) + P(x = 5) \ldots + P(x = 10)$$

Die Eintrittswahrscheinlichkeiten können als Werte der Massenfunktion aus der Tabelle der Binomialverteilung im Buch (S. 292/293) abgelesen werden. Natürlich könnte man auch die Formel für die Massenfunktion benutzen, nur wäre dies deutlich aufwändiger. Die Wahrscheinlichkeit für 4 Erfolge bei $n = 10$ und $p = 0.2$ ist z.B. $P(x = 4) = 0.0881$. Es ergibt sich:

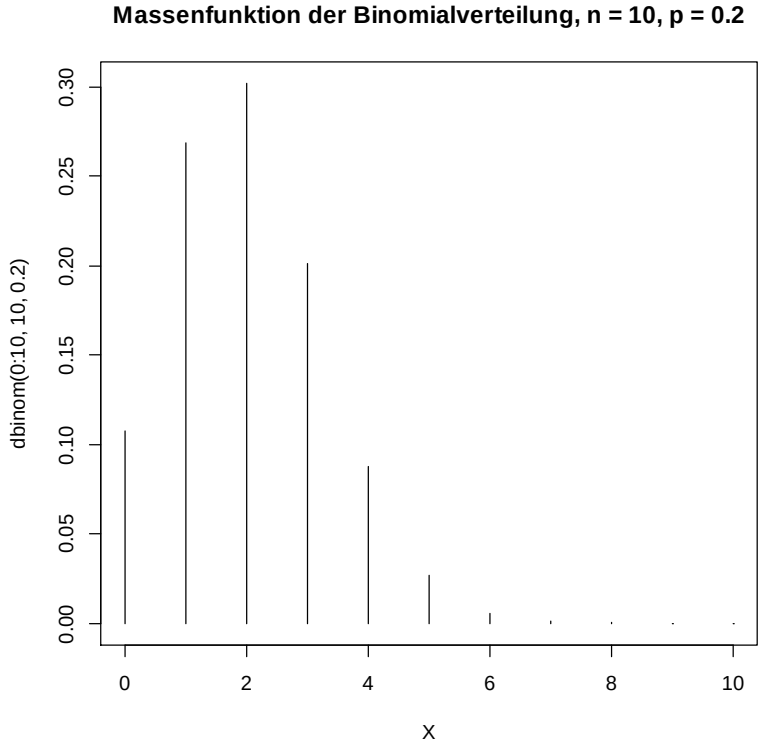
$P(X \geq 4) = 0.0881 + 0.0264 + 0.0055 + 0.0008 + 0.0001 + 0.0000 + 0.0000 = 0.1209$

Mit einer Wahrscheinlichkeit von ca. 12% bekommen mindestens vier der zehn Nachbarn ein Geschenk.

**Lösung b)**

Hier muss $P(x = 0)$ bestimmt werden. Die Wahrscheinlichkeit, dass kein Nachbar ein Geschenk bekommt, beträgt ca. 10.74%.

Die folgende Abbildung zeigt die Massenfunktion $f_{Bi}(x, n = 10, p = 0.2)$



**Lösung mit R**

```
> options(scipen="10") # Ausgabe in Dezimaldarstellung
>
> # a)
> sum(dbinom(4:10, 10, 0.2))
[1] 0.1209
>
> # b)
> dbinom(0:10, 10, 0.2)
 [1] 0.1073741824 0.2684354560 0.3019898880 0.2013265920 0.0880803840
 [6] 0.0264241152 0.0055050240 0.0007864320 0.0000737280 0.0000040960
[11] 0.0000001024
> dbinom(0, 10, 0.2)
[1] 0.1074
> # Abbildung der Massenfunktion
> plot(0:10, main = "Massenfunktion der Binomialverteilung, n = 10,
+ p = 0.2", dbinom(0:10, 10, 0.2), xlab = "X", type = "h")
```

**Aufgabe 7.4: Basketball**

*Ein Spieler wirft beim Training außerhalb der Drei-Punkte-Linie auf den Korb. Er trifft bei jedem Wurf mit einer Wahrscheinlichkeit von $P = \frac{1}{2}$. Die Zufallsvariable X sei definiert als die Anzahl der Treffer bei einer Serie von $n = 4$ Würfen.*

*a) Geben Sie die Massenfunktion der Zufallsvariablen X an.*
*b) Wie groß sind Modus, Median, Erwartungswert und Varianz von X?*
*c) Zeichnen Sie die Verteilungsfunktion von X.*

**Lösung a)**

Nach der Formel für die Massenfunktion der Binomialverteilung ergibt sich

$f_{Bi}\left(x, \frac{1}{2}, 4\right) = \binom{4}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x}$ für $x = 0,1,2,3,4$

Für $x = 0$ ergibt sich:

$f_{Bi}\left(0, \frac{1}{2}, 4\right) = \binom{4}{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{4-0} = \frac{4!}{0!(4-0)!}1\left(\frac{1}{2}\right)^{4} = \frac{4!}{4!}1 * 0.0625 = 0.0625$

Für $x = 1$ ergibt sich:

$f_{Bi}\left(1, \frac{1}{2}, 4\right) = \binom{4}{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{4-1} = \frac{4!}{1!(4-1)!}\frac{1}{2}\left(\frac{1}{2}\right)^{3} = \frac{4*3*2*1}{3*2*1} * 0.5 * 0.125 = 0.25$

Für $x = 2,3,4$ vgl. die Werte unten.

Natürlich ließen sich die Wahrscheinlichkeiten auch über einen Wahrscheinlichkeitsbaum ermitteln. Hierzu sind folgende Überlegungen nötig:

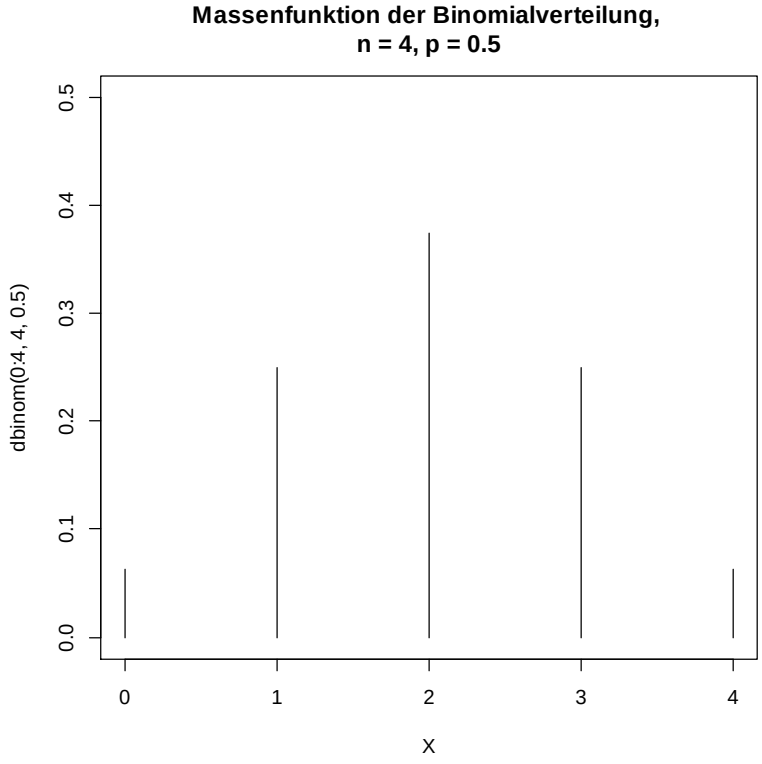
Die 16 Kombinationen sind (0 = kein Treffer, 1 = Treffer):

| | | | | |
|---|---|---|---|---|
| 1. Wurf (kein Treffer) | 0 | | | |
| 1.&2. Wurf | 00, 01 | | | |
| 1.&2.&3. Wurf | 000, 001 | 010, 011 | | |
| 1.&2.&3.&4. Wurf | 0000, 0001 | 0010, 0011 | 0100, 0101 | 0110, 0111 |
| 1. Wurf (Treffer) | 1 | | | |
| 1.&2. Wurf | 10, 11 | | | |
| 1.&2.&3. Wurf | 100, 101 | 110, 111 | | |
| 1.&2.&3.&4. Wurf | 1000, 1001 | 1010, 1011 | 1100, 1101 | 1110, 1111 |

Es gibt nur eine 4er-Serie, d.h., alle 4 Würfe sind ein Treffer. Analog gibt es nur eine 0er Serie. Das Auszählen der Treffer-Ereignisse ergibt die Massenfunktion dieser Zufallsvariable:

$$f(x) = \begin{cases} \frac{1}{16} \text{ für } x = 0 \\ \frac{4}{16} \text{ für } x = 1 \\ \frac{6}{16} \text{ für } x = 2 \\ \frac{4}{16} \text{ für } x = 3 \\ \frac{1}{16} \text{ für } x = 4 \\ 0 \quad sonst \end{cases}$$

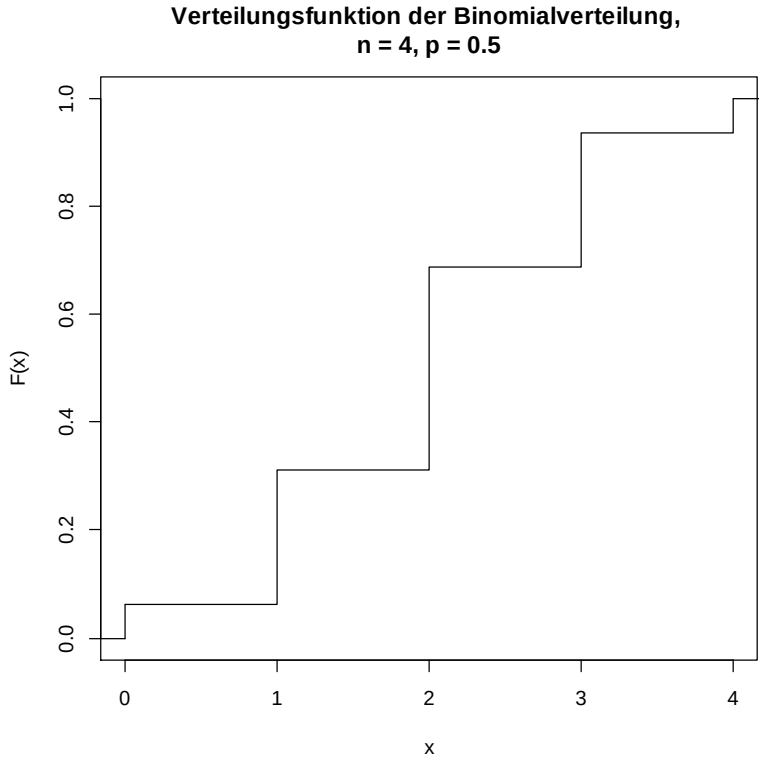
Die Werte der Massenfunktion könnten auch über die Tabelle der Binomialverteilung (vgl. Buch, S. 292/293) für $n = 4$ und $p = 0.5$ bestimmt werden. Es ergibt sich folgende Abbildung




**Lösung b)**
Modus: Der Wert mit der größten Wahrscheinlichkeitsmasse, hier $x_{Mod} = 2$.
Median: $x_{[0.5]} = 2$. Dieser Wert kann am besten grafisch ermittelt werden.
Wegen Symmetrie ist der Erwartungswert $E(X) = 2$.
Die Varianz ist nach der Formel $V(X) = np(1-p)$ zu berechnen.
Es ergibt sich $V(X) = 4 * 0.5(1-0.5) = 1$.
Alternativ ließe sich die Varianz auch berechnen mit
$V(X) = \sum_{i=1}^{n} \big(x_i - E(X)\big)^2 f(x_i)$
$= (0-2)^2 * \frac{1}{16} + (1-2)^2 * \frac{4}{16} + (2-2)^2 * \frac{6}{16} + (3-2)^2 * \frac{4}{16} + (4-2)^2 * \frac{1}{16} = 1$

**Lösung c)**
Die Verteilungsfunktion ergibt sich durch Kumulieren der Massenfunktion. Die Verteilungsfunktion dieser Zufallsvariable ist:

$$F(x) = \begin{cases} 0 \text{ für } x < 0 \\ \frac{1}{16} \text{ für } 0 \leq x < 1 \\ \frac{5}{16} \text{ für } 1 \leq x < 2 \\ \frac{11}{16} \text{ für } 2 \leq x < 3 \\ \frac{15}{16} \text{ für } 3 \leq x < 4 \\ 1 \text{ für } 4 \leq x \end{cases}$$




## Lösung mit R

```
> # a)
> dbinom(0:4, 4, 0.5) # Massenfunktion
[1] 0.0625 0.2500 0.3750 0.2500 0.0625

> # b)
> # So könnte man E(X) berechnen, aber das wäre
> # recht umständlich:
> sum(dbinom(0:4, 4, 0.5)*(0:4))
[1] 2
>
> # c) Abbildungen
> F <- dbinom(0:4, 4, 0.5) # Massenfunktion
> X <- c(0:4)              # X Variable definieren
> ecdf_F <- cumsum(F)      # CDF berechnen
>
> # Zeichnen der Massenfunktion
> plot(0:4, main = "Massenfunktion der Binomialverteilung,
+ n = 4, p = 0.5", dbinom(0:4, 4, 0.5),
+ ylim = c(0, 0.5), xlab = "X", type = "h")
>
> # Zeichnen der Verteilungsfunktion
> plot(X, ecdf_F, main = "Verteilungsfunktion der Binomialverteilung,
+ n = 4, p = 0.5", type = "s", xlab = "x", ylab = "F(x)", ylim = c(0, 1))
> segments(0,0,0,0.0625); segments(-1,0,0,0); segments(4,1,5,1)
```

**Aufgabe 7.5: Verteilung**

*Die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariable $X$, die nur die drei Werte 0, 1 und 2 annehmen kann, sei durch die Massenfunktion*

$$f(x) = P(X = x) = \begin{cases} \frac{20}{27}a - 3^{-(1+x)} & \text{für } x = 0,1,2 \\ 0 & sonst \end{cases}.$$

*definiert. Dabei ist $a$ eine geeignet zu wählende Konstante.*

*a) Wie groß muss $a$ sein? Begründung. Erstelle Sie eine Skizze der Massenfunktion.*
*b) Wie groß sind die Wahrscheinlichkeiten $P(X > 1)$ und $P(1 < X \leq 3)$?*
*c) Wie groß ist die bedingte Wahrscheinlichkeit $P(X = 2|X \geq 1)$?*
*d) Berechnen Sie Erwartungswert $E(X)$ und Varianz $V(X)$.*

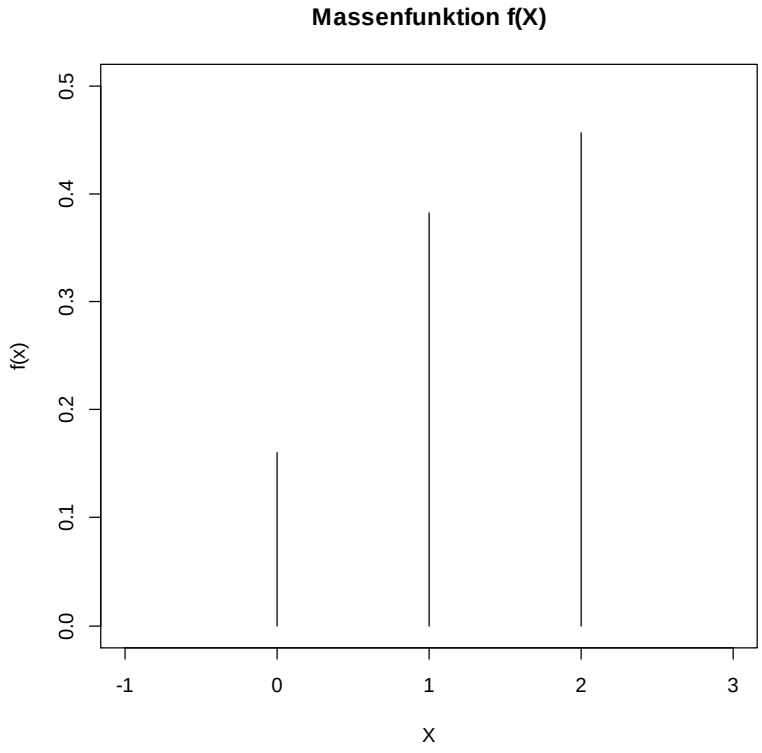
**Lösung a)**

Die Summe der Werte einer Wahrscheinlichkeitsfunktion muss Eins sein, daher lässt sich der Parameter $a$ folgendermaßen herleiten:

$$\sum f(x) = \left(\frac{20}{27}a - 3^{-1}\right) + \left(\frac{20}{27}a - 3^{-2}\right) + \left(\frac{20}{27}a - 3^{-3}\right)$$
$$= \frac{20}{9}a - \frac{1}{3} - \frac{1}{9} - \frac{1}{27} = \frac{20}{9}a - \frac{13}{27} = 1$$
$$\Leftrightarrow \frac{20}{9}a = 1 + \frac{13}{27} \Leftrightarrow a = \frac{2}{3}$$

Die Massenfunktion ist dann:

$$f(x) = \begin{cases} \frac{20}{27} * \frac{2}{3} - 3^{-(1+0)} = 0.160 \text{ für } x = 0 \\ \frac{20}{27} * \frac{2}{3} - 3^{-(1+1)} = 0.383 \text{ für } x = 1 \\ \frac{20}{27} * \frac{2}{3} - 3^{-(1+2)} = 0.457 \text{ für } x = 2 \\ 0 \ sonst \end{cases}$$



**Lösung b)**
$P(X > 1) = 0.457 = P(1 < X \leq 3)$

**Lösung c)**
$P\ (X = 2\ |X \geq 1) = \frac{P(X=2 \cap X \geq 1)}{P(X \geq 1)} = \frac{0.457}{(0.383+0.457)} = 0.544$

**Lösung d)**
Erwartungswert:
$E(X) = 0 * 0.160 + 1 * 0.383 + 2 * 0.457 = 1.297$
Varianz:
$E(X^2) = 0^2 * 0.160 + 1^2 * 0.383 + 2^2 * 0.457 = 2.211$
$V(X) = 2.211 - (1.297)^2 = 0.529$

**Lösung mit R**

```
> # b) Massenfunktion in R zeichnen:
> X <- c(0:2)# Variable X definieren
> f <- (20/27)*(2/3) - 3^(-(1+X))# Variable f definieren
> plot(X, f, main = "Massenfunktion f(X)", type = "h",
+ ylim = c(0, 0.5), xlim = c(-1, 3), xlab = "X", ylab = "f(x)")
> # c) Die Lösung für c) ist aus der Massenfunktion abzulesen.
```

**Aufgabe 7.6: Dreiecksverteilung**

*Die Verteilung einer stetigen Zufallsvariable X sei durch folgende Dichtefunktion definiert*

$$f(x) = \begin{cases} 4ax & 0 < x < 1 \\ 6a - 2ax \quad \textit{für} & 1 \leq x < 3. \\ 0 & sonst \end{cases}$$

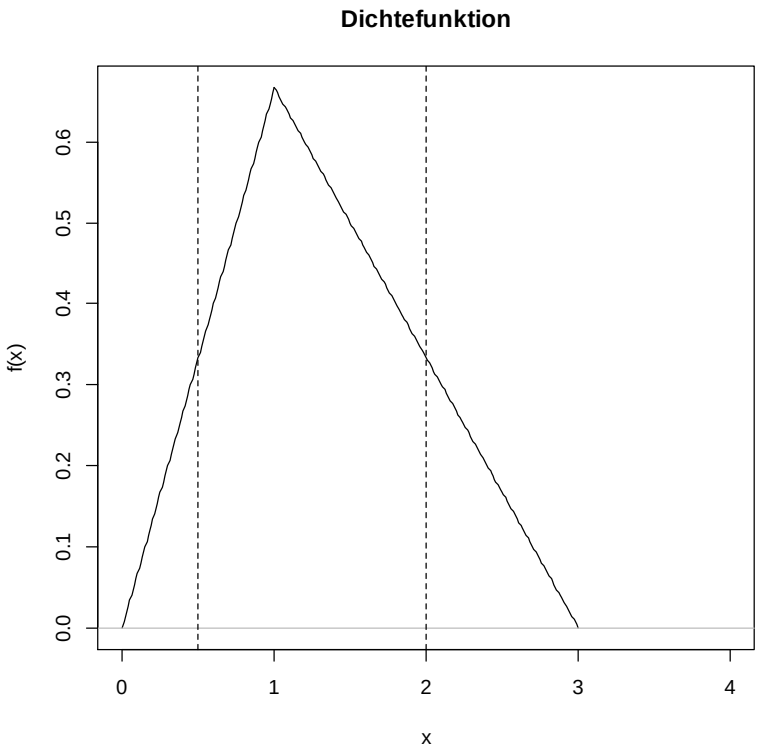
*Dabei ist a eine geeignet zu wählende Konstante.*

a) *Wie groß muss a sein? Begründung. Erstellen Sie eine Skizze der Dichtefunktion.*
b) *Wie groß sind die Wahrscheinlichkeiten $P(X = 1)$, $P(0.5 < X < 2)$ und $P(X < 2)$?*
c) *Wie groß ist die bedingte Wahrscheinlichkeit $P(X < 1|X < 0.5)$?*
d) *Berechnen Sie den Erwartungswert der Zufallsvariablen X.*

**Lösung a)**
Der Parameter $a$ muss so gewählt werden, dass das Integrale unter der Dichtefunktion exakt Eins ergibt, also:
$\int_{-\infty}^{+\infty} f(x)dx = 1 \Leftrightarrow 1 = \int_0^1 (4ax)dx + \int_1^3 (6a - 2ax)dx = 2ax^2 + (6ax - ax^2)$
$1 = 2a + ((18a - 9a) - (6a - a)) \Leftrightarrow 1 = 2a + (9a - 5a)$
$1 = 6a \Leftrightarrow a = \frac{1}{6}$
Damit ergibt sich folgende Dichtefunktion:

$$f(x) = \begin{cases} \frac{2}{3}x & 0 < x < 1 \\ 1 - \frac{1}{3}x \quad \textit{für} & 1 \leq x < 3 \\ 0 & sonst \end{cases}$$

Die Abbildung der Dichte (mit Hilfslinien für $x = 0.5$ und $x = 2$, vgl. b)):




**Lösung b)**
Die Punktwahrscheinlichkeit einer stetigen Zufallsvariable mit unendlichen vielen Ausprägungen ist Null. Es ist also: $P(x = 1) = 0$. Der Wert der Dichte an der Stelle 1 ist hingegen $f(x = 1) = \frac{2}{3}$.
Die beiden anderen Wahrscheinlichkeiten berechnen sich über Flächen unter der Dichtefunktion. Für $P(0.5 < X < 2)$ berechnet man am besten die (Dreiecks)-Flächen links von $x = 0.5$ und rechts von $x = 2$ und zieht diese Flächen von 1 ab.
$P(0.5 < X < 2) = 1 - \frac{1}{2}\frac{1}{3}\frac{1}{2} - 1\frac{1}{3}\frac{1}{2} = \frac{12-1-2}{12} = \frac{3}{4}$
Für $P(X < 2)$ berechnet man die (Dreiecks)-Fläche rechts von $x = 2$ und zieht diese Fläche von 1 ab.
$P(X < 2) = 1 - 1\frac{1}{3}\frac{1}{2} = \frac{5}{6}$

**Lösung c)**
$P(X < 1|X < 0.5) = 1$
Die Bedingung $(X < 0.5)$ ist eine Teilmenge von $(X < 1)$.
Folglich ist $P(X < 1|X < 0.5) = 1$

**Lösung d)**
$E(X) = \int_{-\infty}^{+\infty} xf(x) = \int_0^1 \left(\frac{2}{3}x^2\right) dx + \int_1^3 \left(x - \frac{1}{3}x^2\right) dx$
$= \frac{2}{3}\int_0^1 x^2\, dx + \int_1^3 x dx - \frac{1}{3}\int_1^3 x^2\, dx = \frac{2}{3}\frac{1}{3}[x^3]_0^1 + \frac{1}{2}[x^2]_1^3 - \frac{1}{3}\frac{1}{3}[x^3]_1^3$
$= \frac{2}{9} + \frac{1}{2}(9-1) - \frac{1}{9}(27-1) = 1.33$
Da die Dichtefunktion linkssteil bzw. rechtsschief ist, ist es plausibel, dass $E(X) > 1$ ist.

**Lösung mit R**
```
> # b)
> X1 <- seq(0, 1, 0.01)# für 0 <  X <  1
> f1 <- 2/3*X1
> X2 <- seq(1, 3, 0.01)# für 1 <= X <= 3
```

```
> f2 <- 1-1/3*X2
>
> # Zeichnen der Dichtekurve:
> plot(X1, f1, main = "Dichtefunktion", type = "l",
+ xlim = c(0, 4), xlab = "x", ylab = "f(x)")
> lines(X2, f2)
> abline(h = 0, col = "grey")
>
> abline(v = 0.5, lty = 2)
> abline(v = 2, lty = 2)
```

**Aufgabe 7.7: Rechteckverteilung**

*Eine stetige Zufallsvariable $X$ sei im Intervall $[2,7]$ rechteckverteilt.*

*a) Fertigen Sie eine Skizze für die Dichtefunktion und die Verteilungsfunktion an.*

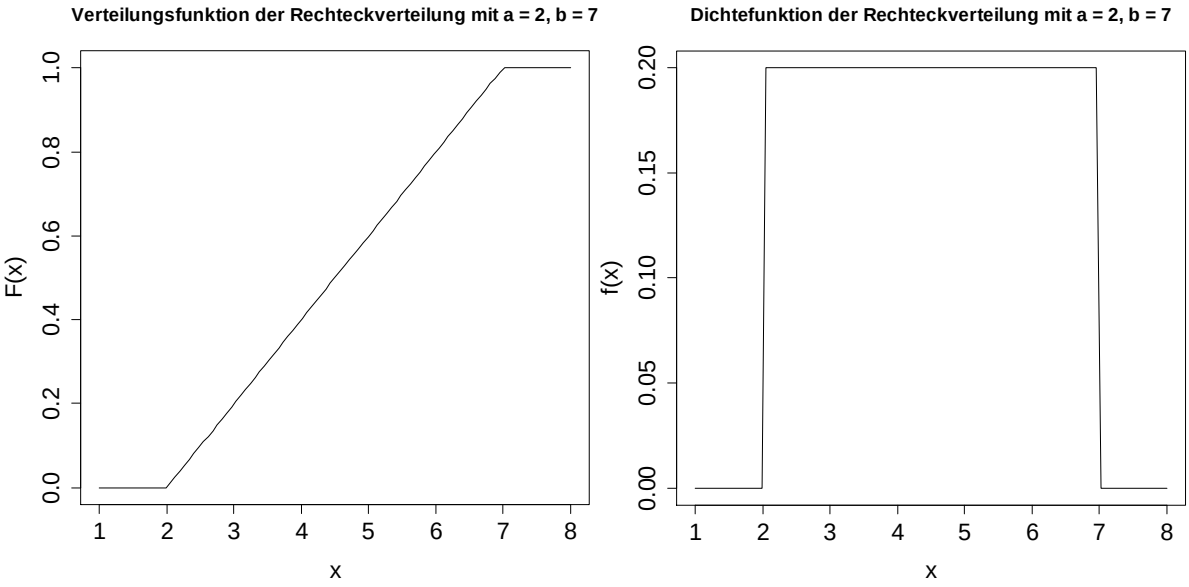
*b) Wie groß sind die Wahrscheinlichkeiten $P(X > 3)$, $P(1.5 < X < 3)$ und $P(X < 5)$?*

*c) Wie groß ist die bedingte Wahrscheinlichkeit $P(X < 3|X < 4)$?*

*d) Berechnen Sie den Erwartungswert und die Varianz der Zufallsvariablen $X$.*

**Lösung a)**

Mit den Intervallgrenzen und der Rechteckverteilung ergeben sich folgende Skizzen:




**Lösung b)**

Hinweis: Hier wie im Folgenden gelten die Beziehungen (7.24)-(7.26) des Lehrbuchs.

$P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - \frac{3-2}{7-2} = 1 - \frac{1}{5} = \frac{4}{5}$

Die Wahrscheinlichkeit für $P(X > 3)$ beträgt 80%.

$P(1.5 < X < 3) = F(3) - F(1.5) = \frac{1}{5} - 0 = \frac{1}{5}$

Die Wahrscheinlichkeit für $P(1.5 < X < 3)$ beträgt 20%.

$P(X < 5) = P(X \leq 5) = F(5) = \frac{5-2}{7-2} = \frac{3}{5}$

Die Wahrscheinlichkeit für $P(X < 5)$ beträgt 60%.

Man beachte, dass $P(X < 5) = P(X \leq 5)$ wegen $P(X = 5) = 0$.

**Lösung c)**

Die bedingte Wahrscheinlichkeit ist

$P(X < 3|X < 4) = \frac{0.2}{0.4} = 0.5$

**Lösung d)**

$E(X) = \frac{a+b}{2} = \frac{2+7}{2} = 4.5$

$V(X) = \frac{(b-a)^2}{12} = \frac{(7-2)^2}{12} = 2.083$

**Lösung mit R**

```
> # a)
> # Verteilungsfunktion der Rechteckverteilung
> a = 2; b = 7
> curve(punif(x, a, b), cex.lab = 1.5, cex.axis = 1.5,
+ main = "Verteilungsfunktion der Rechteckverteilung mit a = 2, b = 7",
+ xlab = "x", ylab = "F(x)", xlim = c(1, 8))
>

> # Dichtefunktion der Rechteckverteilung
> curve(dunif(x, a, b), cex.lab = 1.5, cex.axis = 1.5,
+ main = "Dichtefunktion der Rechteckverteilung mit a = 2, b = 7",
+ xlab = "x", ylab = "f(x)", xlim = c(1, 8))
>
> # b)
> # P(X > 3)
> 1 - punif(3, a, b)
[1] 0.8
>
> # P(1.5 < X < 3)
> punif(3, a, b) - punif(1.5, a, b)
[1] 0.2
>
> # P(X < 5)
> punif(5, a, b)
[1] 0.6
> punif(4.9999999, a, b)
[1] 0.6
>
> # c)
> # P(X < 3 | X < 4)
> punif(3, a, b) / punif(4, a, b)
[1] 0.5
```

**Aufgabe 7.8: 1-kg-Pakete**

*Bei einer Gewichtskontrolle von 1-kg-Paketen wurde festgestellt, dass das Gewicht näherungsweise normalverteilt ist mit Mittelwert 1.02 kg und Standardabweichung 0.03 kg.*

*a) Wie viel Prozent aller Pakete wiegen mindestens 1 kg? Skizzieren Sie die Dichtefunktion.*

*b) Jenseits welchen Betrags befinden sich die 5% schwersten Pakete?*

*c) Wie viel Prozent aller Pakete wiegen mindestens 1.02 kg?*

**Lösung a)**

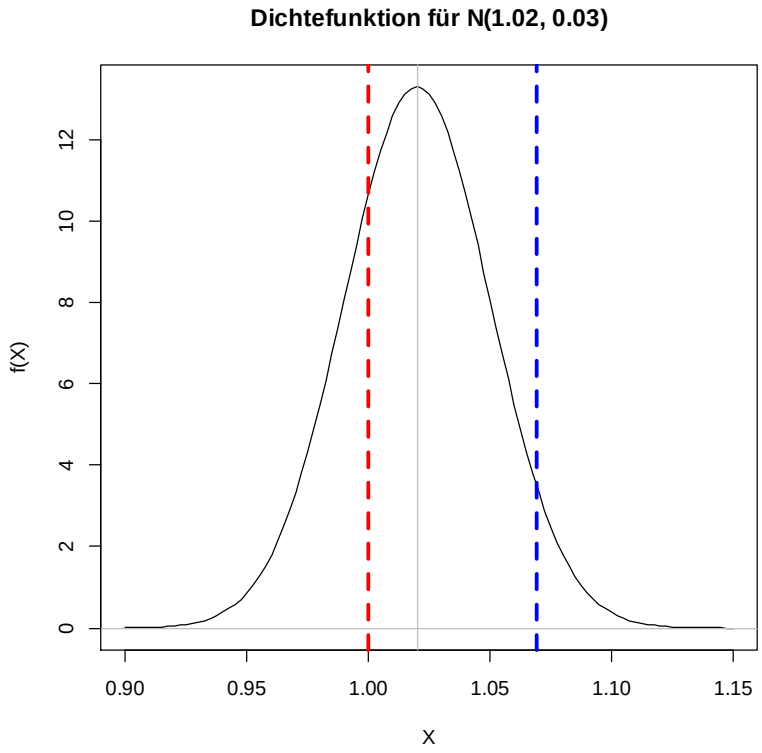
Vorüberlegungen: Das Gewicht folgt annahmegemäß einer $N(1.02,0.03)$-Verteilung. Wir können diese Verteilung durch Standardisieren in eine $N(0,1)$-Verteilung überführen und dann Wahrscheinlichkeitsaussagen treffen. Die Wert der Verteilungsfunktion der $N(0,1)$-Verteilung, $F_{St}(z)$, bzw. die Quantile der $N(0,1)$-Verteilung, $z[q]$, können aus der Tabelle im Buch (S. 294) abgelesen werden. Wir nehmen an, dass es beliebig viele Werte für das Gewicht gibt (Stetigkeit).

Standardisieren: $z = \frac{x-\mu_x}{\sigma_x} = \frac{1-1.02}{0.03} = -0.67$

$P(X \geq 1kg) = 1 - P(X \leq 1kg) = 1 - P(z \leq -0.67) = 1 - F_{St}(-0.67)$

$= 1 - \big(1 - F_{St}(0.67)\big) = F_{St}(0.67) = 0.7486$

Ca. 75% der Pakete wiegen mindestens 1kg.
In der Abbildung der Dichtefunktion der $N(1.02,0.03)$-Verteilung unten, ist diese Wahrscheinlichkeit die Fläche rechts von der gestrichelten roten Linie.




**Lösung b)**
Das 95%-Quantil der $N(0,1)$-Verteilung beträgt $z[0.95] = 1.64$.
Es gilt $z = \frac{x-\mu_x}{\sigma_x}$. Es ergibt sich $z = \frac{x - 1.02}{0.03} = 1.64$
Rückstandardisieren ergibt: $x = 1.02 + 1.64 \cdot 0.03 = 1.069$
Die kritische Grenze liegt bei 1.069 kg, d.h., jenseits dieses Betrags sind die 5% der schwersten
Pakete (vgl. Fläche rechts von der blau gestrichelten Linie in der Abbildung oben).

**Lösung c)**
Standardisieren: $z = \frac{x-\mu_x}{\sigma_x} = \frac{(1.02-1.02)}{0.03} = 0$
$P(X \geq 1.02) = P(z \geq 0) = 1 - P(z \leq 0) = F_{St}(0) = 0.5$
50% aller Pakete wiegen mindestens 1.02 kg.

**Lösung mit R**

```
> # a) Zeichnen der Dichtefunktion
> z <- qnorm(0.95); z
[1] 1.644854
> X95 <- 1.02 + z*0.03; X95
[1] 1.069346
> curve(dnorm(x, mean = 1.02, sd = 0.03), xlim = c(0.9, 1.15),
+ main = "Dichtefunktion für N(1.02, 0.03)", xlab = "X", ylab = "f(X)")
> abline(v = 1.02, col = "grey"); abline(h = 0, col = "grey")
> abline(v = 1, lty = 2, lwd = 3, col = "red")
> abline(v = X95, lty = 2, lwd = 3, col = "blue")
>
> # Berechnung der Wahrscheinlichkeit P(X >= 1.00)
> 1 - pnorm(1, 1.02, 0.03) # mind. 1kg
[1] 0.7475075
>
> # b) P(X > X_krit) = 0.95
> qnorm(0.95, 1.02, 0.03)
[1] 1.069346
>
> # c) P(X >= 1.02)
> 1 - pnorm(1.02, 1.02, 0.03)
[1] 0.5
```

**Aufgabe 7.9: Kaffee**

*Eine Kaffeerösterei weiß aus Erfahrung, dass das Füllgewicht der 500g-Packung Kaffee näherungsweise einer Normalverteilung mit $\mu = 500g$ und $\sigma = 5g$ unterliegt.*

*a) Wie groß ist die Wahrscheinlichkeit, dass eine Packung exakt 500g wiegt?*
*b) Wie groß ist die Wahrscheinlichkeit, dass eine Packung zwischen 495 g und 505 g wiegt?*
*c) Wie groß ist die Wahrscheinlichkeit, dass eine Packung weniger als 490 g wiegt?*
*d) Welches Gewicht unterschreitet eine Packung mit einer Wahrscheinlichkeit von 5%?*

**Lösung a)**

Das Merkmal Gewicht ist stetig, also beliebig genau messbar. Wir messen also nicht nur Gramm-genau, sondern im Bereich von 1/1000 Gramm oder noch genauer (z.B. 1/1000000 Gramm). Die Wahrscheinlichkeit, dass ein Paket exakt 500.000g (oder noch genauer) wiegt ist daher nahe Null. Es ist also $P(X = 500) \approx 0$.

**Lösung b)**

Das vorgegebene Intervall ist ausgehend von Mittelwert je eine Standardabweichung nach oben und nach unten. Wir können also die Empirical Rule (vgl. S. 53 im Buch) anwenden und als Ergebnis ist zu erwarten, dass ca. 2/3 aller Beobachtungen in diesem Intervall liegen.
Die Wahrscheinlichkeit für ein „Gewicht kleiner 495 g“ lässt sich folgendermaßen berechnen:

$$P(X < 495) = P\left(\frac{x-\mu}{\sigma} < \frac{495-500}{5}\right) = P(z < -1) = F_{St}(-1) = 1 - F_{St}(1)$$
$$= 1 - 0.8413 = 0.1587$$

Die Wahrscheinlichkeit für ein „Gewicht größer 505 g“ ist:

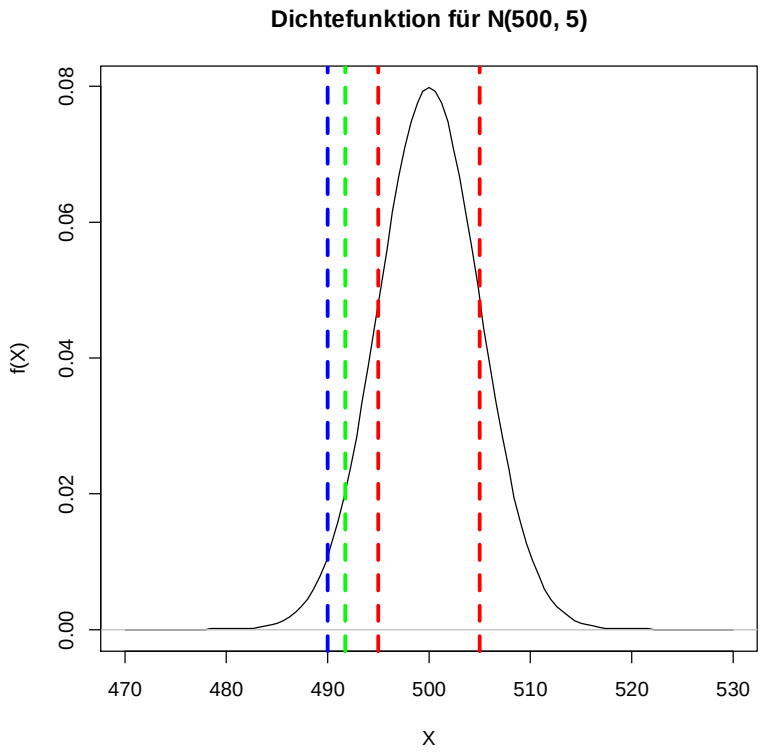
$$P(X > 505) = P\left(\frac{x-\mu}{\sigma} > \frac{505-500}{5}\right) = P(z > 1) = 1 - P(z < 1) = 1 - F_{St}(1)$$
$$= 1 - 0.8413 = 0.1587$$

Die Wahrscheinlichkeit für ein „Gewicht kleiner 495 g“ und ein „Gewicht größer 505 g“ ist:
$P(X < 495 \cap X > 505) = 0.1587 + 0.1587 = 0.3174$
Damit ist die gesuchte Gegenwahrscheinlichkeit:
$P(495 \leq X \leq 505) = 1 - 0.3174 = 0.6826$

Die Wahrscheinlichkeit, dass eine Packung zwischen 495g und 505g wiegt, beträgt 68.26% (ca. 2/3). In der Abbildung unten ist dies die Fläche, die durch die Dichtefunktion und die rot gestrichelten Linien begrenzt ist.

**Lösung c)**
Die Wahrscheinlichkeit für „Gewicht kleiner 490 g“ ist:
$P(X < 490) = P\left(\frac{x-\mu}{\sigma} < \frac{490-500}{5}\right) = P(z < -2) = F_{St}(-2) = 1 - F_{St}(2)$
$= 1 - 0.9772 = 0.0228$
Die Wahrscheinlichkeit, dass eine Packung weniger als 490g wiegt, beträgt 2.28%. In der Abbildung unten ist dies die Fläche unter der Dichte links der blau gestrichelten Linie.

**Lösung d)**
Gesucht ist das 5% Quantil der Standardnormalverteilung und damit $z_{[0.05]}$. Auf Grund der Symmetrie von $f(z)$ gilt:
$z_{[0.05]} = -z_{[0.95]} = -1.645$ (vgl. Tab. 14.1 im Lehrbuch)
Durch Rückstandardisieren kann der entsprechende Wert der Zufallsvariablen $X$ berechnet werden.
$-1.645 = \frac{x-500}{5} \Longrightarrow x = 491.775g$
Mit einer Wahrscheinlichkeit von 5% wird das Gewicht eines Pakets 491.775g unterschreiten. In der Abbildung unten ist dies die Fläche unter der Dichte links der grün gestrichelten Linie.




**Lösung mit R**

```
> # a)
> # Achtung, R liefert den Funktionswert für die Dichte:
> dnorm(500, 500, 5)
[1] 0.07978846
> # Dies ist NICHT die Punktwahrscheinlichkeit (diese ist Null),
> # da wir es mit einer stetigen Zufallsvariablen zu tun haben.
> # b)
> pnorm(505, 500, 5) - pnorm(495, 500, 5)
```

```
[1] 0.6826895
> # c)
> pnorm(490, 500, 5)
[1] 0.02275013
> # d)
> qnorm(0.05, 500, 5)
[1] 491.7757
>
> # Dichtefunktion für N(500,5)
> curve(dnorm(x, mean = 500, sd = 5), xlim = c(470, 530),
+ main = "Dichtefunktion für N(500, 5)", xlab = "X", ylab = "f(X)")
> abline(h = 0, col = "grey")
> abline(v = 495, lty = 2, lwd = 3, col = "red")
> abline(v = 505, lty = 2, lwd = 3, col = "red")
> abline(v = 490, lty = 2, lwd = 3, col = "blue")
> abline(v = 491.775, lty = 2, lwd = 3, col = "green")
```

**Aufgabe 7.10: Reifen**

*Ein Reifenproduzent ist sich sicher, dass die Lebensdauer seiner Winterreifen durch eine Normalverteilung mit einem Mittelwert von 32000 km und einer Standardabweichung von 2500 km charakterisiert ist.*

a) *Angenommen Sie kaufen einen dieser Winterreifen. Wie wahrscheinlich ist es, dass dieser Reifen 40000 km hält? Erklärung.*
b) *Ungefähr welcher Anteil der produzierten Reifen hält erwartungsgemäß weniger als 30000 km?*
c) *Ungefähr welcher Anteil der produzierten Reifen hält erwartungsgemäß zwischen 30000 km und 35000 km?*
d) *Bestimmen und interpretieren Sie den Interquartilsabstand der Daten.*
e) *Ein lokaler Händler plant eine Marketing-Strategie, bei der jeder Kunde eine Entschädigung erhält, wenn die Reifen nicht eine Mindestzahl an km halten. Der Händler möchte aber sein finanzielles Risiko begrenzen und plant, nur einem von 25 Kunden eine Entschädigung zu gewähren. Welche Mindestzahl an km muss für die Garantie gewählt werden?*

**Lösung a)**

Standardisieren: $z = \frac{x-\mu_x}{\sigma_x} = \frac{40000-32000}{2500} = 3.2$

$P(X \geq 40000) = 1 - P(X \leq 40000) = 1 - P(z < 3.2) = 1 - F_{St}(3.2)$

$= 1 - 0.9993 = 0.0007$

Mit anderen Worten: 40000 km sind 3.2 Standardabweichungen über dem arithmetischen Mittel. Das ist extrem viel und daher ist eine solche Mindestlaufleistung sehr unwahrscheinlich. In der Abb. unten ist diese Wahrscheinlichkeit die Fläche unter der Dichte rechts der gestrichelten roten Linie.

**Lösung b)**

$z = \frac{x-\mu_x}{\sigma_x} = \frac{30000-32000}{2500} = -0.8$

$P(X < 30000) = P(X \leq 30000) = P(z < -0.8) = F_{St}(-0.8) = 1 - F_{St}(0.8)$

$= 1 - 0.7991 = 0.2119$

Daraus folgt, dass 21.2% aller Reifen erwartungsgemäß weniger als 30000 km halten.
In der Abb. unten ist diese Wahrscheinlichkeit die Fläche unter der Dichte links der linken gestrichelten blauen Linie.

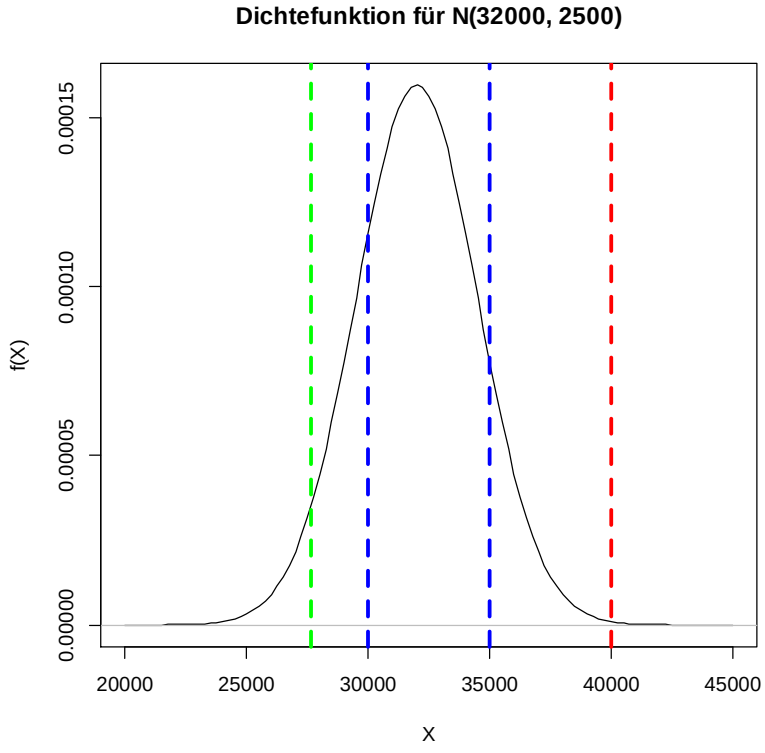
**Lösung c)**

$P(30000 \leq X \leq 35000) = P\left(\frac{30000-32000}{2500} \leq z \leq \frac{35000-32000}{2500}\right)$

$= P(-0.8 \leq z \leq 1.2) = F_{St}(1.2) - F_{St}(-0.8) = 0.8849 - 0.2119 = 0.673$
67.3% der Reifen halten zwischen 30000 km und 35000 km. In der Abb. unten ist diese Wahrscheinlichkeit die Fläche unter der Dichte zwischen den gestrichelten blauen Linien.

**Lösung d)**
$z_{[0.75]} = 0.67$
$0.67 = \frac{X-32000}{2500} \Longrightarrow X_{[0.75]} = 33675\ km$
$z_{[0.25]} = -0.67$
$-0.67 = \frac{X-32000}{2500} \Longrightarrow X_{[0.25]} = 30325\ km$
$IQA = 33675\ km - 30325\ km = 3350km$
Der Interquartilsabstand beträgt 3350 km. Das bedeutet, dass die mittleren 50% der Reifen zwischen 30325 km und 33675 km, in einem Intervall der Länge 3350km, halten.

**Lösung e)**
Die gewünschte Eintrittswahrscheinlichkeit ist $\frac{1}{25} = 0.04$. Wir suchen also das 4%-Quantil der Standardnormalverteilung. Dieser Wert kann dann über die Formel für die Standardisierung wieder in einen Wert für $X$ umgerechnet werden.
Es gilt: $z_{[a]} = -z_{[1-a]}$ mit α als Wahrscheinlichkeit im Intervall (0,1).
Hier:
$z_{[0.04]} = -z_{[0.96]} = -1.75$. Der korrespondierende $z$-Wert ist also –1.75.
$-1.75 = \frac{X-32000}{2500} \rightarrow X_{[0.04]} = 27625km$
Der Händler sollte seine Garantie bei der Unterschreitung von 27625 km einführen, um sicher zu gehen, dass nur ca. 4% der Kunden die Entschädigung erhalten. In der Abb. unten ist diese Wahrscheinlichkeit die Fläche unter der Dichte links der gestrichelten grünen Linie.




**Lösung mit R**
```
# a)
1 - pnorm(40000, 32000, 2500)
```

```
# b)
pnorm(30000, 32000, 2500)
# c)
pnorm(35000, 32000, 2500) - pnorm(30000, 32000, 2500)
# d)
qnorm(0.75, 32000, 2500) - qnorm(0.25, 32000, 2500)
# e)
qnorm(0.04, 32000, 2500)

# Dichtefunktion für N(32000,2500)
curve(dnorm(x, mean = 32000, sd = 2500), xlim = c(20000, 45000),
main = "Dichtefunktion für N(32000, 2500)", xlab = "X", ylab = "f(X)")
abline(h = 0, col = "grey")
abline(v = 40000, lty = 2, lwd = 3, col = "red")
abline(v = 30000, lty = 2, lwd = 3, col = "blue")
abline(v = 35000, lty = 2, lwd = 3, col = "blue")
abline(v = 27625, lty = 2, lwd = 3, col = "green")
```

**Aufgabe 7.11: IQ-Test**

*In einer Personalabteilung werden Intelligenztests durchgeführt. Die erzielten Punkte der Teilnehmer sind näherungsweise normalverteilt mit Mittelwert 100 und Standardabweichung 16. Mit welcher Wahrscheinlichkeit erzielt ein zufällig ausgewählter Teilnehmer eine Punktzahl von*

*a) 100 oder mehr*
*b) über 148*
*c) zwischen 84 und 116*
*d) über 132*

*Welche Punktwerte trennen*

*e) die kleinsten 16% aller Teilnehmer*
*f) die mittleren 95% aller Teilnehmer*
*g) die höchsten 2.5% aller Teilnehmer?*

**Lösung a)**

Wir unterstellen ein stetiges Merkmal.

$z = \frac{X-\mu_x}{\sigma_x} = \frac{100-100}{16} = 0$

$P(X \geq 100) = 1 - P(X \leq 100) = 1 - P(z \leq 0) = 1 - F_{St}(0) = 1 - 0.5 = 0.5$

Mit einer Wahrscheinlichkeit von 50% werden 100 Punkte oder mehr erreicht.

**Lösung b)**

$z = \frac{X-\mu_x}{\sigma_x} = \frac{148-100}{16} = 3$

$P(X \geq 148) = 1 - P(X \leq 148) = 1 - P(z \leq 3) = 1 - F_{St}(3) = 1 - 0.9987 = 0.0013$

Mit einer Wahrscheinlichkeit von 0.13% erreicht ein Teilnehmer über 148 Punkte. In der Abb. unten ist diese Wahrscheinlichkeit die Fläche rechts von der gestrichelten roten Linie.

**Lösung c)**

$P(84 \leq X \leq 116) = P\left(\frac{84-100}{16} \leq z \leq \frac{116-100}{16}\right) = P(-1 \leq z \leq 1)$

$= F_{St}(1) - F_{St}(-1) = 0.8413 - (1 - 0.8413) = 0.6826$

Mit einer Wahrscheinlichkeit von 68.26% liegt der Punktewert zwischen 84 und 116 Punkten. Das Ergebnis hätte man auch mit der Empirical Rule erhalten, da es hier um die Wahrscheinlichkeit für das Intervall $\mu \pm 1\sigma$ handelt und diese ist ca. 2/3.

**Lösung d)**

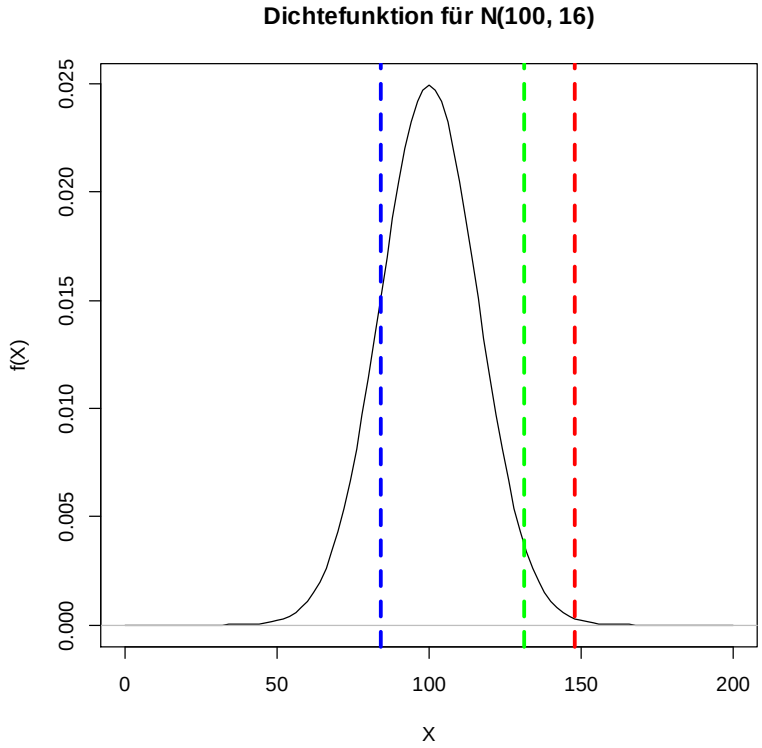
$z = \frac{X-\mu_x}{\sigma_x} = \frac{(132-100)}{16} = 2$

$P(X \geq 132) = 1 - P(X \leq 132) = 1 - P(z \leq 2) = 1 - F_{St}(2) = 1 - 0.9772 = 0.0228$
Mit einer Wahrscheinlichkeit von 2.28% erreicht ein Teilnehmer mehr als 132 Punkte.

**Lösung e)**
$z_{[0.16]} = -0.99$
$-0.99 = \frac{X-100}{16} \Longrightarrow X = 84.16$
16% aller Teilnehmer erreichen 84.16 Punkte oder weniger. In der Abb. unten ist diese Wahrscheinlichkeit die Fläche links von der gestrichelten blauen Linie.

**Lösung f)**
$z_{[0.975]} = 1.96$
$1.96 = \frac{X-100}{16} \Longrightarrow X = 131.36$
$z_{[0.025]} = -1.96$
$-1.96 = \frac{X-100}{16} \Longrightarrow X = 68.64$
Die mittleren 95% der Punkte der Teilnehmer liegen zwischen 68.64 und 131.36.

**Lösung g)**
$z_{[0.975]} = 1.96$
$1.96 = \frac{X-100}{16} \Longrightarrow X = 131.36$
2.5 % aller Teilnehmer haben 131.36 Punkte oder mehr. In der Abb. unten ist diese Wahrscheinlichkeit die Fläche rechts von der gestrichelten grünen Linie.




**Lösung mit R**

```
> # a)
> pnorm(100, 100, 16)
[1] 0.5
> # b)
> 1 - pnorm(148, 100, 16)
[1] 0.001349898
```

```
> # c)
> pnorm(116, 100, 16) - pnorm(84, 100, 16)
[1] 0.6826895
> # d)
> 1 - pnorm(132, 100, 16)
[1] 0.02275013
> # e)
> qnorm(0.16, 100, 16)
[1] 84.08867
> # f)
> qnorm(0.975, 100, 16)
[1] 131.3594
> qnorm(0.025, 100, 16)
[1] 68.64058
> # g)
> qnorm(0.975, 100, 16)
[1] 131.3594
>
> # Dichtefunktion für N(100,16)
> curve(dnorm(x, mean = 100, sd = 16), xlim = c(0, 200),
+ main = "Dichtefunktion für N(100, 16)", xlab = "X", ylab = "f(X)")
> abline(h = 0, col = "grey")
> abline(v = 148, lty = 2, lwd = 3, col = "red")
> abline(v = 84, lty = 2, lwd = 3, col = "blue")
> abline(v = 131.35, lty = 2, lwd = 3, col = "green")
```

# Kapitel 8: Grenzwertsätze

**Aufgabe 8.1: Stichprobenmittelwert**

*Die Verteilung eines metrischen Merkmals $X$ in einer sehr großen Grundgesamtheit sei unbekannt. Jedoch wissen wir, dass der Mittelwert 1000 sei und die Standardabweichung gleich 14 ist. Nun wird eine Stichprobe vom Umfang 500 gezogen.*

*a) Wie groß ist der Erwartungswert des Stichprobenmittelwertes?*

*b) Wie groß ist die Standardabweichung des Stichprobenmittelwertes?*

**Lösung a)**

Die wahren Parameter der Grundgesamtheit, Mittelwert $\mu = 1000$ und Standardabweichung $\sigma = 14$, sind gegeben. Der Erwartungswert des Stichprobenmittelwerts ist $E(\bar{X}_n) = \mu = 1000$. Der Erwartungswert des Stichprobenmittelwertes entspricht dem Mittelwert der Grundgesamtheit, d.h., im Mittel „liegt“ man mit einer großen Stichprobe also richtig. Dies folgt aus dem „Gesetz der großen Zahl“, nach dem der Mittelwert einer Stichprobe gegen den Erwartungswert (hier der Mittelwert der Grundgesamtheit) konvergiert.

**Lösung b)**

Die Varianz des Stichprobenmittelwerts ist $V(\bar{X}_n) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{14^2}{500} = 0.392$.

Die Standardabweichung des Stichprobenmittelwertes ist $\sigma_{\bar{X}} = \sqrt{0.392} = 0.626$.

Man beachte: die Streuung von Mittelwerten einer Zufallsstichprobe ist immer kleiner als die Streuung in der Grundgesamtheit.

**Aufgabe 8.2: Wähler der CDU**

*Der Anteil der potentiellen CDU-Wähler in Sachsen betrage ca. 40%. In einer Zufallsstichprobe vom Umfang 100 werden Wahlberechtigte aus Sachsen nach ihrer Wahlabsicht befragt.*

*a) Welcher Anteil der Befragten wird erwartungsgemäß angeben, CDU wählen zu wollen?*

b) *Wie groß ist die Varianz dieses Stichprobenanteilswertes?*

**Lösung a)**

Auch hier sind die wahren Parameter der Grundgesamtheit, Anteilswert $p = 0.4$ und Varianz $p(1-p) = 0.24$, gegeben. Der Erwartungswert des Stichprobenanteilswerts ist $E(H_n) = p = 0.4$. Im Mittel „liegt" man mit einer hinreichend großen Stichprobe also richtig (Gesetz der großen Zahl).

**Lösung b)**

Die Varianz des Stichprobenanteilswertes ist $V(H_n) = \frac{p(1-p)}{n} = \frac{0.4*0.6}{100} = 0.0024$. Die Standardabweichung des Stichprobenanteils ist damit $\sigma_{H_n} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4*0.6}{100}} = 0.0489$.

**Aufgabe 8.3: Haushaltseinkommen**

*Das mittlere Jahreseinkommen aller Haushalte in einer Großstadt beträgt 20000 €, bei einer Standardabweichung von 4500 €. Eine Zufallsstichprobe mit 900 Haushalten wird gezogen.*

*a) Wie groß ist die Wahrscheinlichkeit, in der Stichprobe ein mittleres Jahreseinkommen von über 19500 € vorzufinden?*

*b) Mit welcher Wahrscheinlichkeit liegt der Stichprobenmittelwert zwischen 19400 € und 20600 €?*

*c) Geben Sie, ohne zu rechnen oder statistische Tafeln zu benutzen, an, in welchem der folgenden drei Intervalle*
*(1) 18500 € - 19500 €,(2) 20100 € - 21100 €, (3) 19500 € - 20500 €*
*der Stichprobenmittelwert mit großer Wahrscheinlichkeit liegen wird.*

*d) Nun weist man Sie darauf hin, dass das Haushaltseinkommen sicher nicht normalverteilt ist, vielmehr eine deutliche Schiefe aufweist und stellt Ihre Ergebnisse zu a) bis c) in Frage. Wie begegnen Sie diesem Argument?*

**Lösung a)**

Bei *n* = 900 handelt es sich um eine große Stichprobe. Wir gehen nach dem Zentralen Grenzwertsatz davon aus, dass der Stichprobenmittelwert mit approximativ normalverteilt ist mit $E(\bar{X}) = \mu = 20000$ und $\sigma_{\bar{X}} = \frac{4500}{\sqrt{900}} = 150$. Der standardisierte Stichprobenmittelwert folgt also einer $N(0,1)$-Verteilung.
Damit ist $P(\bar{X} > 19500) = 1 - P(\bar{X} \leq 19500) = 1 - P(z \leq 3.33) = F_{st}(3.33) = 0.9996$.
Die Wahrscheinlichkeit für ein mittleres Einkommen in der Stichprobe von über 19500 € ist 99.96%.

**Lösung b)**

Gesucht ist die Wahrscheinlichkeit $P(19400 \leq \bar{X}_n \leq 20600)$.
Wir berechnen über die Standardisierung

$$P(19400 \leq \bar{X}_n \leq 20600) = P\left(\frac{19400-20000}{\frac{4500}{\sqrt{900}}} \leq \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}}} \leq \frac{20600-20000}{\frac{4500}{\sqrt{900}}}\right)$$

$= P(-4 \leq Z \leq 4) = F_{St}(4) - F_{St}(-4) = F_{St}(4) - \left(1 - F_{St}(4)\right)$
$= 0.9999 - 1 + 0.9999 = 0.9998$
Die gesuchte Wahrscheinlichkeit ist 0.9998 (geringe Abweichung zur Lösung mit R, vgl. unten).

**Lösung c)**
Mit großer Wahrscheinlichkeit wird der Stichprobenmittelwert im Intervall *(3) 19500 € - 20500 €* liegen. In diesem Intervall liegt der wahre Mittelwert der Grundgesamtheit und damit auch der Erwartungswert des Stichprobenmittels.

**Lösung d)**
Das Haushaltseinkommen ist nicht normalverteilt, sondern üblicherweise linkssteil bzw. rechtsschief, aber die Verteilungsfunktion des standardisierten Stichprobenmittelwerts strebt mit wachsender Zahl der Beobachtungen gegen die Verteilungsfunktion der Standardnormalverteilung (zentraler Grenzwertsatz). Dieser Satz gilt unabhängig von der Verteilung des Merkmals in der Grundgesamtheit. Daher sind die Wahrscheinlichkeitsaussagen approximativ in Ordnung.

**Lösung mit R**
```
> # a)
> 1 - pnorm(19500, 20000, 4500/sqrt(900))
[1] 0.9995709
> # b)
> pnorm(20600, 20000, 4500/sqrt(900)) - pnorm(19400, 20000, 4500/sqrt(900))
[1] 0.9999367
```

**Aufgabe 8.4: Mindestgewicht**

*Das mittlere Gewicht von Zementsäcken eines Zementwerks sei $\mu = 25.4kg$ bei einer Standardabweichung von $s = 3kg$. Die Zementsäcke werden auf Paletten mit je 40 Säcken verpackt und transportiert. Jede Palette wird vom Abnehmer geprüft, wobei ein mittleres Mindestgewicht eines Zementsacks von 25.0 kg gesetzt ist.*

*a) Wie groß ist die Wahrscheinlichkeit, dass eine Palette dieses Limit übersteigt? Prüfen Sie die Bedingungen für die Anwendung des zentralen Grenzwertsatzes.*
*b) Auf welchen Wert müsste die Standardabweichung der Zementsäcke reduziert werden, damit man zu 99% sicher sein kann, dass der Abnehmer die Lieferung nicht beanstandet?*
*c) Wenn man die Standardabweichung nicht reduzieren kann, auf welchen Wert müsste das mittlere Gewicht erhöht werden, um das Sicherheitsniveau in b) von 99% einzuhalten?*
*d) Angenommen, die Verteilung der Zementsäcke selbst folgt einer Normalverteilung. Wie groß ist die Wahrscheinlichkeit, dass ein zufällig gewählter Sack mindestens 25.0 kg wiegt? Erklärung und Bezug zu a). Argumentieren Sie mit der unterschiedlichen Streuung von Beobachtungswerten und Stichprobenmittelwerten.*

**Lösung a)**
Die wahren Parameter der Grundgesamtheit, Mittelwert $\mu$ und Standardabweichung $\sigma$ (hier mit $s$ bezeichnet) sind gegeben. Die Stichprobengröße ist $n = 40$. Gesucht ist die Wahrscheinlichkeit, dass das mittlere Gewicht eines Sacks auf der Palette 25kg übersteigt.

$$P(\bar{X} > 25) = 1 - P(\bar{X} \leq 25) = 1 - P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{25 - 25.4}{\frac{3}{\sqrt{40}}}\right) = 1 - P(z \leq -0.84)$$

$= P(z \leq 0.84) = F_{St}(0.84) = 0.7995$
Die Wahrscheinlichkeit, dass eine Palette das Limit von 25kg übersteigt, beträgt ca. 80%.

Die Voraussetzungen zur Anwendung des zentralen Grenzwertsatzes sind:

1) Die Stichprobe muss hinreichend groß sein: Die Bedingung $n > 30$ ist erfüllt.
2) Die Stichprobe darf nicht mehr als 5% der Grundgesamtheit umfassen, was auch als erfüllt angesehen werden kann, weil die Gesamtzahl der Zementsäcke deutlich größer ausfallen dürfte (keine Angabe in der Aufgabe).
3) Die Gewichte der einzelnen Zementsäcke untereinander sind unabhängig. Das Gewicht des einen Zementsackes beeinflusst nicht das Gewicht eines anderen Zementsackes.

**Lösung b)**

Das mittlere Mindestgewicht der Säcke soll mit 99% 25kg betragen. Also: $P(\bar{X} > 25) = 0.99 \Leftrightarrow P(\bar{X} < 25) = 0.01$ (die Punktwahrscheinlichkeit ist näherungsweise Null). Die Wahrscheinlichkeit von 99% entspricht einem $z$-Wert von 2.326 (vgl. Tabelle der Verteilungsfunktion der Standardnormalverteilung). Daraus ergibt sich der z-Wert für 1% mit $z = -2.326$.

Da $\bar{X}$ einer Normalverteilung folgt mit Mittelwert $25.4$ und Standardabweichung $\frac{\sigma}{\sqrt{n}}$, können wir über die Formel für die Standardisierung die gesuchte Standardabweichung bestimmen:

$z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \Leftrightarrow -2.326 = \frac{(25-25.4)*\sqrt{40}}{\sigma} \Longrightarrow \sigma = 1.088$

Die Standardabweichung bei der Abfüllung der Säcke muss 1.088 betragen, damit der Abnehmer die Lieferung mit einer Wahrscheinlichkeit von 1% nicht beanstandet.

**Lösung c)**

$P(\bar{X} > 25)$ müsste 99% betragen. Dies entspricht einem $z$-Wert von -2.326 (vgl. b)).

$z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \Leftrightarrow -2.326 = \frac{25-\mu}{\frac{3}{\sqrt{40}}} \Longrightarrow \mu = -\left(-2.326 * \left(\frac{3}{\sqrt{40}}\right) - 25\right) = 26.12\ kg$

Das mittlere Gewicht der abgefüllten Säcke müsste auf 26.12 kg erhöht werden, um das Sicherheitsniveau von 1% einzuhalten.

**Lösung d)**

Wir betrachten nun die Verteilung der Grundgesamtheit und nicht mehr die Verteilung des Mittelwerts der Stichprobe mit Größe $n$. Die Umrechnung der kritischen Größe 25 ergibt:

$z = \frac{X-\mu}{\sigma} = \frac{25-25.4}{3} = -0.13$

Daraus folgt:

$P(X > 25) = P(z > -0.13) = 1 - P(z \leq -0.13) = F_{st}(0.13) = 0.5517$

Die Wahrscheinlichkeit, dass ein Zementsack das Gewicht von 25kg übersteigt, liegt bei 55.2%.

Die Wahrscheinlichkeit ist deutlich kleiner als in a) mit 80%, weil hier wie in b) die Wahrscheinlichkeit des Auftretens einer Ausprägung aus der Grundgesamtheit berechnet wird (Verteilung der Zementsäcke selbst). In a) wurde die Wahrscheinlichkeit des Auftretens eines Stich-probenmittelwertes bestimmt. Mittelwerte haben kleinere Varianzen als Beobachtungen aus Grundgesamtheiten.

**Lösung mit R**

```
> # a)
> 1 - pnorm(25, 25.4, 3/sqrt(40))
[1] 0.8004624
> # d)
> 1 - pnorm(25, 25.4, 3)
[1] 0.553051
```

**Aufgabe 8.5: Ausschuss**

*Bei der Produktion eines Massenartikels fällt erfahrungsgemäß 10% Ausschuss an. Es wird eine Zufallsstichprobe vom Umfang n = 800 genommen. Wie groß ist die Wahrscheinlichkeit, dass sich in der Stichprobe a) mehr als 90 und b) weniger als 60 schlechte Stücke befinden?*

**Lösung a)**

Bei einer Stichprobe, die eine Zufallsauswahl zwischen „gut" und „schlecht" vornimmt, ist die Summe guter bzw. schlechter Stücke binomialverteilt. Wir wissen nach (8.16) im Buch, dass bei einer großen Stichprobe die Normalverteilung zur Approximation der Binomialverteilung verwendet werden kann. Nach (8.16) können an die Stelle der Summen auch Anteilswerte treten. Wir wählen im Folgenden diesen 2. Weg. Hierzu muss – wie im Folgenden ersichtlich – die Aufgabe ein wenig verändert werden. Man kann an diesem Beispiel auch leicht überprüfen, dass beide Wege zum gleichen Ziel führen.

Gesucht ist $P\left(H_n > \frac{90}{800}\right) = 1 - P\left(H_n \leq \frac{90}{800}\right)$. Der Erwartungswert ist $E(H_n) = 0.1$.

Hinweis: $H_n$ ist der Stichprobenanteilswert bzw. der Mittelwert der Stichprobe (das Gegenstück zu $\bar{X}_n$ bzw. $\bar{X}$).

Es gilt $\frac{n}{N} < 0.05$ mit $N$ = Größe der Grundgesamtheit, da hier die Grundgesamtheit sehr groß ist. Weiterhin ist die Stichprobe mit $n = 800$ groß genug, um die Binomialverteilung durch die Normalverteilung approximieren zu können. Hier gilt $np \geq 10$ und $n(1-p) \geq 10$.

Die Approximation der Binomialverteilung durch die Normalverteilung kann durch die Stetigkeitskorrektur verbessert werden (vgl. S. 173/174 im Buch). Hier mit Anwendung der Stetigkeitskorrektur (entweder mit 0.5 absolut, wenn die Summe der binomialverteilten Zufallsvariablen interessiert, oder als Anteil von $n$, d.h. $\frac{0.05}{800}$).

Die Abweichung zum Ergebnis ohne Stetigkeitskorrektur ist sehr gering.

Mit Stetigkeitskorrektur und dem Weg über die Anteilswerte:

$$P\left(H_n > \frac{90}{800}\right) = 1 - P\left(H_n \leq \frac{90}{800}\right) = 1 - P(H_n \leq 0.1125)$$

$$= 1 - F_{st}\left(\frac{0.1125 + \frac{0.5}{800} - 0.1}{\frac{\sqrt{0.1*0.9}}{\sqrt{800}}}\right) = 1 - F_{st}\left(\frac{0.013125}{\frac{0.3}{28.2843}}\right) = 1 - F_{st}(1.237) = 1 - 0.891 = 0.109$$

Somit beträgt die Wahrscheinlichkeit, dass in der Stichprobe mehr als 90 schlechte Stücke befinden, ca. 10.9%.

Der genaue Wert der Binomialverteilung $F_{Bi}(X = 90, n = 800,\ p = 0.1)$ ist 0.8906465. Die genaue Wahrscheinlichkeit über die Binomialverteilung ist also 0.1093535 (vgl. Lösung mit R unten). Die Abweichung zu 0.109 ist also sehr gering. Wie zu erwarten kann also die Binomialverteilung hier durch die Normalverteilung approximiert werden.

**Lösung b)**

Hier mit den Absolutwerten für die binomialverteilte Zufallsvariable $X$ (analog für den Anteilswert $H$, nur dort alles durch $n = 800$ geteilt):

Gesucht ist $P(X < 60) = P(X \leq 59)$. Der Erwartungswert ist $E(X) = 0.1 * 800 = 80$.

$$P(X \leq 59) \approx F_{st}\left(\frac{59 + 0.5 - 80}{\sqrt{0.1*0.9}*\sqrt{800}}\right) = F_{st}\left(\frac{-20.5}{8.4853}\right) = F_{st}(-2.42) = 1 - F_{st}(2.42)$$

$$= 1 - 0.9922 = 0.0078$$

Der genaue Wert der Binomialverteilung $F_{Bi}(X = 59, n = 800,\ p = 0.1)$ ist 0.00615964. Die Abweichung zu 0.0078 ist also auch hier gering.

**Lösung mit R**

```
> # Ausgabe der genauen Werte über die Binomialverteilung:
> # a)
> # Über die Summe der Werte der Massenfunktion
> 1 - sum(dbinom(0:90, 800, 0.1))
[1] 0.1093535
> # oder
> # Über die Verteilungsfunktion
> 1 - pbinom(90, 800, 0.1)
[1] 0.1093535
> # b)
> # Über die Summe der Werte der Massenfunktion
> sum(dbinom(0:59, 800, 0.1))
[1] 0.00615964
> # oder
> # Über die Verteilungsfunktion
> pbinom(59, 800, 0.1)
[1] 0.00615964
```

**Aufgabe 8.6: Blutspende**

*In der Blutbank der Uniklinik in Leipzig gehen pro Tag 300 Blutspenden ein.*

*a) Angenommen, die Häufigkeit für Blutgruppe AB in der Grundgesamtheit aller potentiellen Spender sei 4%. Bestimmen Sie Mittelwert und Standardabweichung der Zahl der Spender mit Blutgruppe AB in der Stichprobe.*

*b) Begründen Sie, warum eine Normalverteilung zur Approximation der Verteilung der Zahl der Spender mit Blutgruppe AB benutzt werden kann.*

*c) Wie wahrscheinlich ist es, 10 oder mehr Spender mit Blutgruppe AB pro Tag zu beobachten?*

**Lösung a)**

Die Zufallsvariable $X$ (Zahl „Erfolge“ bzw. Zahl Personen mit Blutgruppe AB in der Stichprobe) ist binomialverteilt mit $p = 0.04$ und $n = 300$.
Erwartungswert der binomialverteilten Summe ist:
$E(X) = np = 300 * 0.04 = 12 \geq 10$ (Bedingung $np \geq 10$ ist wie $n(1-p) \geq 10$ erfüllt).
Die Standardabweichung der Summe ist:
$\sigma = \sqrt{n * p * (1-p)} = \sqrt{300 * 0.04 * 0.96} = 3.39$

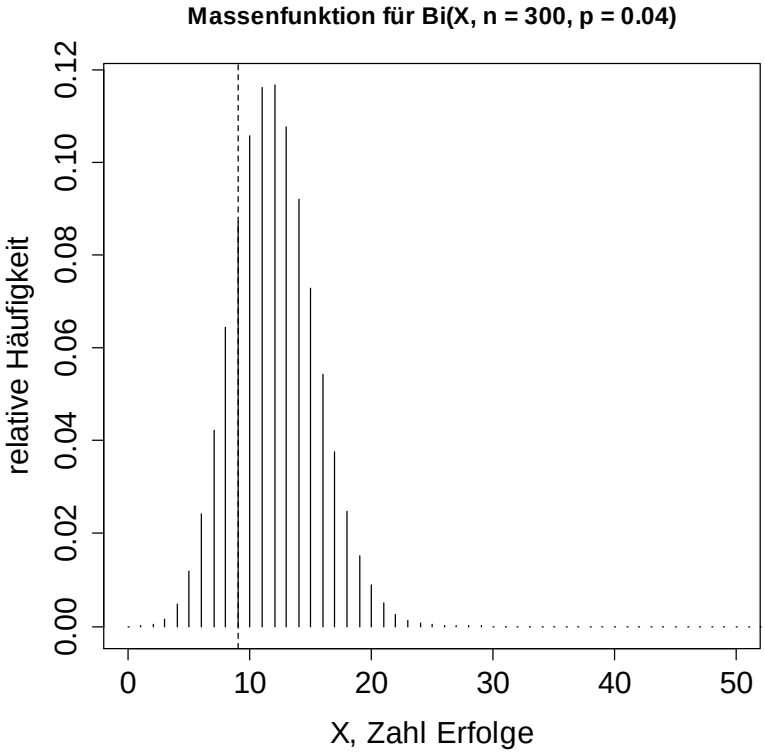
**Lösung b)**

Der zentrale Grenzwertsatz stellt keinerlei Anforderung an die Ausgangsverteilung der zugrundeliegenden Zufallsvariablen. Wie auch immer die Verteilung der $X_i$ (der Elemente in der Grundgesamtheit) beschaffen sein mag, die Verteilungsfunktion des arithmetischen Mittels (multipliziert mit $n$ = Summe) einer Zufallsstichprobe konvergiert unter bestimmten Bedingungen gegen die Normalverteilung.
Hier liegt eine große Stichprobe vor. Wir können eine zufällige Auswahl unterstellen und der Auswahlsatz $\frac{n}{N}$ ist hinreichend klein.

**Lösung c)**

Die Lösung mit Stetigkeitskorrektur ist:
$P(X \geq 10) = 1 - P(X \leq 9) = 1 - P\left(z \leq \frac{9+0.5-12}{3.39}\right) = 1 - P(z \leq -0.74)$
$= P(z \leq 0.74) = F_{st}(0.74) = 0.7704$
Die Wahrscheinlichkeit für 10 oder mehr Blutspender mit Blutgruppe AB in einer Stichprobe mit $n = 300$ beträgt ca. 77%. Die Abweichung zur genauen Lösung über die Binomialverteilung (vgl. Lösung mit R) ist relativ klein.

Die Abbildung unten zeigt die Massenfunktion. Wir suchen die relative Häufigkeit rechts von der gestrichelten senkrechten Linie.




**Lösung mit R**

```
> # Plot der Massenfunktion für Bi(X, n = 300, p = 0.04)
> x <- 0:300
> plot(x, dbinom(0:300, 300, 0.04), type = "h",
+ xlim = c(0, 50),
+ xlab = "X, Zahl Erfolge", ylab = "relative Häufigkeit",
+ main = "Massenfunktion für Bi(X, n = 300, p = 0.04)",
+ cex.lab = 1.5, cex.axis = 1.5)
> abline(v = 9, lty = 2)
>
> # genauer Wert mit R
> # Wahrscheinlichkeit für 9 oder weniger Spender mit AB
> sum(dbinom(0:9, 300, 0.04))
[1] 0.237043
> pbinom(9, 300, 0.04)
[1] 0.237043
> # Wahrscheinlichkeit für 10 oder mehr Spender mit AB
> 1 - sum(dbinom(0:9, 300, 0.04))
[1] 0.762957
> 1 - pbinom(9, 300, 0.04)
[1] 0.762957
```

**Aufgabe 8.7: No-shows**

*Viele Airlines überbuchen ihre Flüge, da regelmäßig ein Teil der gebuchten Passagiere nicht zum Flug erscheint. Ein Airbus A380 hat 550 Sitzplätze für Passagiere. Angenommen, eine Airline verkauft 570 Tickets und der Anteil der no-shows ist 5%. Wie wahrscheinlich ist es, dass die Sitzplätze nicht ausreichen und Passagiere umgebucht oder entschädigt werden müssen?*

*a) Nutzen Sie die Normalverteilung zur Approximation der Binomialverteilung, um die Wahrscheinlichkeit zu bestimmen, dass mindestens 551 Passagiere erscheinen.*

*b) Sollte die Airline die Zahl der verkauften Tickets für solche Flüge ändern? Welche Vor- und Nachteile hätte die Erhöhung der Zahl verkaufter Tickets? Erklärung.*

**Lösung a)**

Erwartungswert der binomialverteilten Summe:

$E(X) = \mu = np = 570 * 0.95 = 541.5$

Bedingung $np \geq 10$ und $n(1-p) \geq 10$ ist erfüllt.

Standardabweichung:

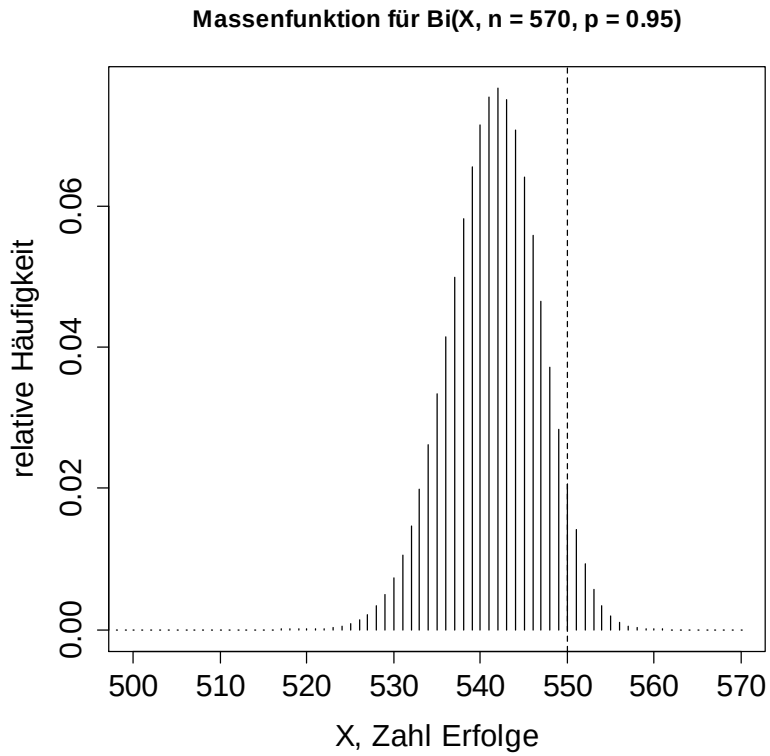
$\sigma = \sqrt{n * p * (1-p)} = \sqrt{541.5 * 0.05 * 0.95} = 5.072$

$P(X \geq 551) = 1 - P(X \leq 550) = 1 - P\left(\frac{X+Sk-\mu}{\sigma} \leq \frac{550+0.5-541.5}{5.072}\right)$

$= 1 - P(z \leq 1.77) = 1 - F_{st}(1.77) = 1 - 0.9616 = 0.0384$

Die Wahrscheinlichkeit, dass mindestens 551 Passagiere erscheinen, beträgt 3.8%.

In der Abbildung der Massenfunktion ist diese Wahrscheinlichkeit gleich der Fläche rechts von der gestrichelten vertikalen Linie.




**Lösung b)**

Ein Vorteil der Reduzierung auf 550 verkaufte Tickets könnte natürlich die Steigerung der Vertrauenswürdigkeit der Airline sein, da durch eventuelle Überbuchungen Vertrauensverluste und Reputationsschäden resultieren können. Allerdings überwiegt in diesem Fall wohl der wirtschaftliche Nutzen, da die Wahrscheinlichkeit von Überbuchungen mit 3.8%

relativ gering ist und mit der Reduzierung der verkauften Tickets auf längere Zeit Kapazitäten verschenkt werden und Verluste entstehen.

**Lösung mit R**

```
> # Plot der Massenfunktion für Bi(X, n = 570, p = 0.95)
> x <- 0:570  # Wert für die x-Achse
> plot(x, dbinom(0:570, 570, 0.95), type = "h",
+ xlim = c(500, 570),
+ xlab = "X, Zahl Erfolge", ylab = "relative Häufigkeit",
+ main = "Massenfunktion für Bi(X, n = 570, p = 0.95)",
+ cex.lab = 1.5, cex.axis = 1.5)
> abline(v = 550, lty = 2)
>
> # genauer Wert mit R
> # Wahrscheinlichkeit für 550 oder weniger Passagiere
> sum(dbinom(0:550, 570, 0.95))
[1] 0.9639856
> pbinom(550, 570, 0.95)
[1] 0.9639856
>
> # Werte der Massenfunktion im Intervall [551, 570]
> dbinom(551:570, 570, 0.95)
 [1] 1.415608e-02 9.257871e-03 5.725483e-03 3.338143e-03 1.828460e-03
 [6] 9.372502e-04 4.475917e-04 1.981275e-04 8.081050e-05 3.015963e-05
[11] 1.021449e-05 3.107968e-06 8.390962e-07 1.978720e-07 3.992461e-08
[16] 6.701128e-09 8.982111e-10 9.013738e-11 6.019720e-12 2.006573e-13
>
> # Wahrscheinlichkeit für 551 oder mehr Passagiere
> 1 - sum(dbinom(0:550, 570, 0.95))
[1] 0.03601438
> 1 - pbinom(550, 570, 0.95)
[1] 0.03601438
```

# Kapitel 9: Schätzung unbekannter Parameter

**Aufgabe 9.1: Kugellager**

*Ein Automat produziert Kugellager. Wir ziehen eine Zufallsstichprobe von $n = 225$ Kugellagern aus der Tagesproduktion dieses Automaten. In der Stichprobe finden wir als durchschnittliches Kugellagergewicht $\bar{x} = 0.824kg$ und $s = 0.005kg$.*

*a) Welche Standardabweichung hat der Stichprobenmittelwert? Ist dies eine Schätzung oder ein wahrer Wert?*
*b) Geben Sie eine zahlenmäßige Punktschätzung für den wahren Mittelwert und*
*c) ein 95%-Konfidenzintervall für den wahren Mittelwert μ des Kugellagergewichts in der Grundgesamtheit an. Interpretieren Sie Ihr Ergebnis.*

**Lösung a)**

Gegeben ist die Standardabweichung der Stichprobe, $s$. Die Standardabweichung der Grundgesamtheit, $\sigma$, ist dagegen unbekannt. Zunächst muss $\sigma$ geschätzt werden. Die geschätzte Standardabweichung der Grundgesamtheit ist

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}}s = \sqrt{\frac{225}{224}} * 0.005 = 0.00502.$$

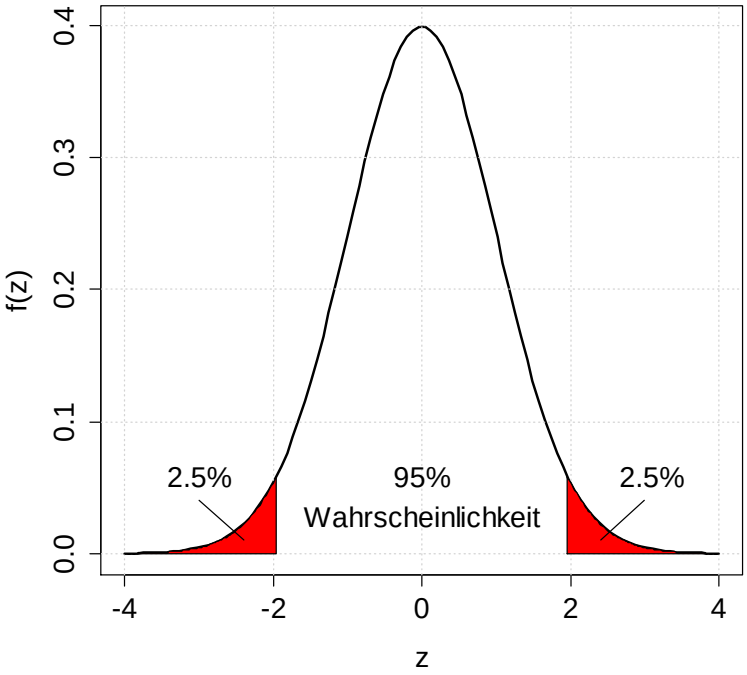
Die (geschätzte) Standardabweichung des Stichprobenmittelwerts berechnet sich dann mit $\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}}$. Mit anderen Worten, wenn $n$ sehr groß wird (gegen unendlich strebt), wird die Standardabweichung des Stichprobenmittels immer kleiner und geht letztlich gegen Null. Hier ist $\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{0.00502}{\sqrt{225}} = 0.000334\ kg$. Dieser Wert ist eine Schätzung, weil schließlich auch

die Standardabweichung der Stichprobe nur eine Schätzung der wahren Standardabweichung der Grundgesamtheit ist.

**Lösung b)**
Der Erwartungswert des Stichprobenmittelwertes ist gleich dem wahren Mittelwert der Grundgesamtheit, d.h. $E(\bar{X}) = \mu$. Also ist es sinnvoll, als Punktschätzung für den wahren Mittelwert der Grundgesamtheit den Mittelwert der Stichprobe anzugeben, also $\bar{x} = 0.824\ kg$.

**Lösung c)**
Wir können unterstellen, dass die Verteilung des Mittelwerts gegen die Normalverteilung konvergiert, wenn die Zahl der Beobachtungen hinreichend hoch ist (zentraler Grenzwertsatz). Das ist hier mit $n = 225$ der Fall. Wir können also nach Standardisierung die Standardnormalverteilung $N(0,1)$ nutzen.
Wir suchen das 95%-Konfidenzintervall für $\mu$, d.h., links und rechts der kritischen Grenzen liegen 2.5% der Wahrscheinlichkeit. Wir suchen also das 2.5%-Quantil von $N(0,1)$, $z_{[0.025]}$, und das 97.5%-Quantil von $N(0,1)$, $z_{[0.975]}$.
Es gilt $z_{[0.025]} = -z_{[0.975]}$ und hier $z_{[0.025]} = -1.96$ bzw. $z_{[0.975]} = 1.96$.
Das Konfidenzintervall berechnet sich folgendermaßen:
$KI(1-\alpha) = \left[\bar{x} - z_{[0.975]}\hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{x} + z_{[0.975]}\hat{\sigma}_{\bar{X}}\right]$
$KI(0.95) = [0.824 - 1.96 * 0.000334 \leq \mu \leq 0.824 + 1.96 * 0.000334]$
$= [0.82335 \leq \mu \leq 0.82465]$
Mit 95%-iger Wahrscheinlichkeit liegt der wahre Mittelwert im Intervall von 0.82335 kg bis 0.82465 kg. Streng genommen ist das nicht richtig, denn das wahre mittlere Gewicht liegt entweder in dem Intervall oder eben nicht. Eine Wahrscheinlichkeitsaussage für ein konkretes Intervall ist daher nicht möglich. Korrekt wäre folgende Interpretation: Wenn wir 100 Zufallsstichproben dieser Größe ziehen würde, ergäben sich 100 verschiedene Konfidenzintervalle. Von diesen 100 Konfidenzintervallen werden im Mittel 95 den wahren Mittelwert beinhalten und 5 nicht. Unser Konfidenzintervall, [0.82335, 0.82465], ist eines dieser 100 Intervalle.

Abbildung für die Dichtefunktion der $N(0,1)$-Verteilung inkl. kritischer Grenzen des 95%-Konfidenzintervalls:




**Lösung mit R:**

```
# R Skript für N(0,1) und 95%-KI
x=seq(-4,4,length=100)
y=dnorm(x,mean=0,sd=1)
plot(x,y,type="l",lwd=2, xlim = c(-4, 4),
xlab = "z", ylab = "f(z)",
cex = 1.4, cex.lab = 1.4, cex.axis = 1.4)
grid()

x1=seq(-4,-1.96,length=100)
y1=dnorm(x1,mean=0,sd=1)
polygon(c(-4,x1,-1.96),c(0,y1,0),col="red")

x2=seq(1.96,4,length=100)
y2=dnorm(x2,mean=0,sd=1)
polygon(c(1.96,x2,4),c(0,y2,0),col="red")

text(0, 0.06, "95%", cex = 1.4)
text(0, 0.03, "Wahrscheinlichkeit", cex = 1.4)

text(-3, 0.06, "2.5%", cex = 1.4)
text(3.1, 0.06, "2.5%", cex = 1.4)
segments (-3, 0.04, -2.4, 0.01)
segments (3, 0.04, 2.4, 0.01)
```

**Aufgabe 9.2: Stichprobenumfang**

*Der Mittelwert μ einer Normalverteilung mit der Varianz $\sigma^2 = 9$ soll geschätzt werden. Eine Stichprobe vom Umfang n = 100 bringt den Mittelwert 53.97.*

*a) Geben Sie ein 95%-Konfidenzintervall für μ an. Interpretieren Sie Ihr Ergebnis.*

*b) Wie groß müsste der Stichprobenumfang genommen werden, damit man ein 95%-Konfidenzintervall der Länge 0.4 erhält?*

**Lösung a)**
Die wahre Varianz der Grundgesamtheit ist mit $\sigma^2 = 9$ gegeben. Man soll ein Konfidenzintervall für den Mittelwert der Grundgesamtheit $\mu$ bilden. Ausgangspunkt ist der Mittelwert der Stichprobe. Wir benötigen dann noch die Standardabweichung des Stichprobenmittelwerts.
Die Standardabweichung des Stichprobenmittelwerts berechnet sich mit
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{100}} = 0.3.$
Das 95%-Konfidenzintervall für den Mittelwert $\mu$ ist dann analog zu Aufgabe 9.1 zu bestimmen. Wir suchen das 95%-Konfidenzintervall, d.h., links und rechts der kritischen Grenzen liegen 2.5% der Wahrscheinlichkeit. Wir suchen also das $z$, welches zum Wert der Verteilungsfunktion der $N(0,1)$-Verteilung $F_{St}(z) = 0.975$ korrespondiert. Das ist $z_{[0.975]} = 1.96$, also das 97.5%-Quantil.
$KI(1-\alpha) = \left[\bar{x} - z_{[0.975]}\sigma_{\bar{X}} \leq \mu \leq \bar{x} + z_{[0.975]}\sigma_{\bar{X}}\right]$
$KI(0.95) = [53.97 - 1.96 * 0.3 \leq \mu \leq 53.97 + 1.96 * 0.3] = [53.382 \leq \mu \leq 54.558]$
Mit 95%-iger Wahrscheinlichkeit liegt der wahre, aber unbekannte Mittelwert im Intervall von 53.382 bis 54.558.
Streng genommen ist das so nicht richtig, denn der wahre Mittelwert liegt entweder in dem Intervall oder eben nicht. Eine Wahrscheinlichkeitsaussage für ein Intervall ist daher nicht möglich. Korrekt wäre folgende Interpretation: Wenn wir 100 Zufallsstichproben dieser Größe ziehen würde, ergäben sich 100 verschiedene Konfidenzintervalle. Von diesen 100 Konfidenzintervallen werden im Mittel 95 den wahren Mittelwert beinhalten und 5 nicht. Unser Konfidenzintervall, [53.382, 54.558], ist eines dieser 100 Intervalle.

**Lösung b)**
Die Formel für das Konfidenzintervall ist (wenn $\sigma$ bzw. $\sigma_{\bar{X}}$ bekannt sind)
$KI(1-\alpha) = \bar{x} \pm z_{[1-\alpha/2]}\sigma_{\bar{X}} = \bar{x} \pm z_{[1-\alpha/2]}\frac{\sigma_X}{\sqrt{n}}$
Für eine Länge des Intervalls von 0.4 und $1 - \alpha = 0.95$ ergibt sich
$z_{[0.975]}\frac{\sigma}{\sqrt{n}} = 0.2 \Leftrightarrow 1.96\frac{3}{\sqrt{n}} = 0.2 \Longrightarrow n = 865$ (aufgerundet).
Es ist aufzurunden, weil es nur ganze Beobachtungszahlen gibt, somit müsste der Stichprobenumfang 865 betragen.
Für eine Länge des Intervalls von 0.4 und $1 - \alpha = 0.99$ ergibt sich wegen $z_{[0.995]} = 2.575$
$z_{[0.995]}\frac{\sigma}{\sqrt{n}} = 0.2 \Leftrightarrow 2.575\frac{3}{\sqrt{n}} = 0.2 \Longrightarrow n = 1492$ (aufgerundet).
Wie man sieht, führt die Erhöhung der Vertrauenswahrscheinlichkeit von 95% auf 99% zu einer massiven Erhöhung der notwendigen Stichprobengröße $n$. Ursache hierfür ist die Quadratwurzel im Nenner der (geschätzten) Standardabweichung des Stichprobenmittels.

**Aufgabe 9.3: Meinungsforschung**

*Ein Meinungsforschungsinstitut befragt im Juli 2013 in Deutschland 1110 zufällig ausgewählt Wahlberechtigte, ob sie bei der Wahl zum Deutschen Bundestag am 22. September 2013 für die CDU und Angela Merkel stimmen werden. 466 der Befragten antworten mit „ja", der Rest mit „nein".*

*a) Man berechne ein 90%-Konfidenzintervall für den wahren Anteil der Wahlberechtigten in der Grundgesamtheit, die für die CDU und Angela Merkel stimmen. Interpretation. Wie groß ist der absolute Fehler des Konfidenzintervalls?*

*b) Erläutern Sie, welche Grenzwertsätze es uns ermöglichen, ein KI wie in a) zu bestimmen.*

*c) Bei einer anderen Befragung im Juli 2013 mit einer Zufallsstichprobe von $n = 1200$ ergibt sich ein Wert für die Zustimmung zur CDU und Angela Merkel von 42.50% bei einem absoluten Fehler von 2.8%. Berechnen Sie die Vertrauenswahrscheinlichkeit dieses Konfidenzintervalls.*

*d) Erläutern Sie den Zielkonflikt zwischen Präzision und Sicherheit der Aussage beim Aufstellen von Konfidenzintervallen.*

**Lösung a)**

Für den Stichprobenanteilswert gilt: $h = \frac{466}{1110} = 0.42$.

Der Standardfehler, welcher der geschätzten Standardabweichung des Stichprobenanteilswertes entspricht, berechnet sich wie folgt:

$\hat{\sigma}_H = SE = \sqrt{h(1-h)/n} = \sqrt{0.42 * 0.58/1110} = 0.0148$

Das 90%-Konfidenzintervall für den wahren Anteilswert $p$ ist dann analog der Aufgaben zuvor zu bestimmen. Wir suchen das 90%-KI, d.h., links und rechts der kritischen Grenzen liegen 5% der Wahrscheinlichkeit. Wir suchen also 95%-Quantil der $N(0,1)$-Verteilung. Das ist $z_{[0.95]} = 1.645$.

$KI(1-\alpha) = \left[\bar{h} - z_{[1-\alpha/2]}\,\hat{\sigma}_H \;\leq p \leq \bar{h} + z_{[1-\alpha/2]}\hat{\sigma}_H\right]$

$KI(0.90) \;= [0.42 \pm 1.645 * 0.0148\,] = [0.396 \;\leq p \leq 0.444]$

Wir können zu 90% darauf vertrauen, dass der wahre aber unbekannte Anteilswert der Grundgesamtheit in dem o.g. Intervall liegt. Technisch korrekte Interpretation: Wenn man 100 Stichproben nach dem o.g. Verfahren realisiert, werden von den 100 resultierenden Konfidenzintervallen im Mittel 90 den wahren Anteilswert beinhalten und 10 nicht. Das Konfidenzintervall $[0.396 \;\leq p \leq 0.444]$ ist eines dieser 100 Intervalle. Das Vertrauen bezieht sich also nicht auf das einzelne Intervall, sondern die Methode zur Bestimmung solcher Intervalle.

Absoluter Fehler des Konfidenzintervalls (Margin of Error):

$z_{[1-\alpha/2]}\hat{\sigma}_H = ME = \frac{(0.444-0.396)}{2} = 0.024$ oder 2.4%

Man könnte das 95%-Konfidenzintervall oben also auch so schreiben:

$p = h \pm ME = 0.42 \pm 0.024$ bei $n = 1110$

Aus diesen Informationen ließe sich (näherungsweise) das Vertrauensniveau $1-\alpha$ bestimmen.

**Lösung b)**

Der zentrale Grenzwertsatz. Bei einer hinreichend großen Zufallsstichprobe der Größe $n$ aus einer Grundgesamtheit mit wahrem Anteilswert $p$ nimmt die Verteilung des Stichprobenanteilswertes $H$ näherungsweise die Form einer Normalverteilung $N(p, \sqrt{\frac{p(1-p)}{n}})$ an.

**Lösung c)**

$\hat{\sigma}_H = SE = \sqrt{h(1-h)/n} = \sqrt{0.425 * 0.575/1200} = 0.0143$

$$z_{[1-\alpha/2]}\hat{\sigma}_H = ME = 0.028$$

Die Vertrauenswahrscheinlichkeit des Konfidenzintervalls ermittelt man folgendermaßen:

$0.028 = z_{[1-\alpha/2]}0.0143 \Longrightarrow z_{[1-\alpha/2]} = 1.96$

Der berechnete $z$-Wert ist das 97.5%-Quantil der $N(0,1)$-Verteilung, d.h. die Vertrauenswahrscheinlichkeit des Konfidenzintervalls vom Juli 2013 war 95%.

**Lösung d)**

Ein relativ präzises (schmales) KI geht einher mit einer relativ großen Unsicherheit darüber, ob der tatsächliche, aber unbekannte Anteilswert bzw. Mittelwert der Grundgesamtheit durch ein solches KI erfasst wird. Dies liegt daran, dass wenn man bei der Berechnung des Konfidenzintervalls den $z_{[1-\alpha/2]}$-Wert reduzieren möchte, dies immer einhergeht mit einer geringeren Vertrauenswahrscheinlichkeit $1-\alpha$. Analog führt die Forderung nach höherem Vertrauen zu einem breiteren Konfidenzintervall.

Es gibt also einen trade-off („Abwägen"): Höhere Präzision der Aussage führt zu höherer Unsicherheit und umgekehrt. Wir müssen also einen Preis (hier: höhere Unsicherheit) zahlen für eine präzisere Aussage. Glücklicherweise ist es möglich, bei einer hinreichend großen Zufallsstichprobe hinreichend sicher (z.B. $1-\alpha = 0.95$) und zugleich hinreichend präzise zu sein ($ME \leq 0.03$).

**Aufgabe 9.4: Gewerkschaft**

*Von den über 10000 gewerkschaftlich organisierten Arbeitnehmern eines Großbetriebs wurden 250 zufällig ausgewählt und nach ihrer Streikbereitschaft gefragt. Von diesen gaben 200 an, dass sie für einen Streik sind.*

*a) Berechnen Sie ein 95%-Konfidenzintervall für den Anteil der Streikwilligen unter allen Gewerkschaftsmitgliedern des Betriebs. Interpretation.*

*b) Das Ergebnis unter a) sei der Gewerkschaft noch zu unscharf. Man fordert, dass der absolute Fehler 0.03 nicht überschreiten darf, also auf $\pm 3$ Prozentpunkte genau geschätzt werden soll. Der Sicherheitsgrad soll nach wie vor 95% betragen. Wie viele Arbeitnehmer müssen befragt werden, damit ein solches Konfidenzintervall konstruiert werden kann?*

**Lösung a)**

Für den Stichprobenanteilswert gilt: $h = \frac{200}{250} = 0.8$.

Der Standardfehler, welcher der geschätzten Standardabweichung des Stichprobenanteilswertes entspricht, berechnet sich wie folgt:

$\hat{\sigma}_H = SE = \sqrt{h(1-h)/n} = \sqrt{0.8 * 0.2/250} = 0.0253$

Das 95%-Konfidenzintervall für den Anteilswert $p$ ist dann analog der Aufgaben zuvor zu bestimmen. Wir suchen das 95%-KI, d.h., links und rechts der kritischen Grenzen liegen 2.5% der Wahrscheinlichkeit. Wir benötigen also $z_{[0.975]} = 1.96$.

$KI(1-\alpha) = \left[h - z_{[0.975]}\,\hat{\sigma}_H \leq p \leq h + z_{[0.975]}\hat{\sigma}_H\right]$

$$KI(0.95) = [0.8 \pm 1.96 * 0.0253\,] = [0.7504 \leq p \leq 0.8496]$$

Der unbekannte Anteil $p$ der Streikwilligen liegt mit 95%-iger Wahrscheinlichkeit im Intervall [0.7505, 0.8496]. Korrekt: Wenn wir 100 Stichproben nach dieser Methode ziehen würden, ergäben sich 100 unterschiedliche Konfidenzintervalle. Von diesen werden im Mittel 95 den wahren Anteil beinhalten und 5 nicht. Unser Intervall [0.7504, 0.8496] ist eines dieser 100 Intervalle.

**Lösung b)**

Der absolute Fehler (ME, Margin of Error) soll $+/-$ 3 Prozentpunkte sein. Dies entspricht einer Breite des KI von 6% oder 0.06. Für die Hälfte des Konfidenzintervall gilt

$$ME = z_{[1-\alpha/2]}\hat{\sigma}_H \Leftrightarrow 0.03 = 1.96\sqrt{\frac{0.8(1-0.8)}{n}} \Longrightarrow n = 683$$

Damit ist die Stichprobengröße mindestens $n = 683$, damit ein solches Konfidenzintervall konstruiert werden kann. In diesem Fall unterstellen wir, dass $h = 0.8$ unverändert bleibt. Man könnte die Frage auch beantworten, ohne diese Annahme zu treffen. In diesen Fall gilt

$$0.03 = 1.96\sqrt{\frac{h(1-h)}{n}}$$

Wir haben also eine Gleichung mit zwei Unbekannten. Glücklicherweise lässt sich aber die Gleichung lösen, wenn wir für $h(1-h)$ den Maximalwert von $0.5 * (1-0.5) = 0.25$ annehmen. In diesem Fall können wir $n$ in einem worst-case-Szenario berechnen.

$$\left(\frac{0.03}{1.96}\right)^2 = \frac{0.25}{n} \Longrightarrow n = \frac{0.25}{\left(\frac{0.03}{1.96}\right)^2} = \frac{0.25}{0.00023428} = 1067.099$$

Um sicherzustellen, dass unabhängig von $h$ der Margin of Error 3% nicht überschreitet, müssten 1068 Teilnehmer befragt werden. Allerdings wäre in diesem Fall die Annahme eines kleinen Auswahlsatzes ($\frac{n}{N} \leq 0.05$) verletzt.

**Aufgabe 9.5: Eine kleine Stichprobe I**

*Eine kleine Stichprobe aus einer normalverteilten Grundgesamtheit lieferte den folgenden Beobachtungsbefund*

*12.7 13.3 13.0 12.9 13.1 12.8 13.1*

*Geben Sie ein Konfidenzintervall an, das den Mittelwert der Grundgesamtheit mit einer Wahrscheinlichkeit von 95% überdeckt. Interpretation.*

**Lösung**

Zunächst berechnen wir das arithmetische Mittel der Stichprobe ($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 12.9857$) und die Stichprobenstandardabweichung ($s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.1884$).

Die geschätzte Standardabweichung der Grundgesamtheit ist

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}}s = \sqrt{\frac{7}{6}}0.1884 = 0.2035$$

Die geschätzte Standardabweichung des Stichprobenmittelwerts ist dann

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{0.2035}{\sqrt{7}} = 0.0769.$$

Es handelt sich um eine kleine Stichprobe aus einer normalverteilten Grundgesamtheit. In diesem Fall ist der Quotient $\frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{X}}}$ t-verteilt mit $n-1$ Freiheitsgraden.

Wir suchen das 95%-Konfidenzintervall:

$$KI(1-\alpha) = \left[\bar{x} - t_{n-1\left[1-\frac{\alpha}{2}\right]}\hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{x} + t_{n-1\left[1-\frac{\alpha}{2}\right]}\hat{\sigma}_{\bar{X}}\right]$$

$t_{6[0.975]} = 2.447$ (siehe Tabelle)

$KI(0.95) = [12.9857 \pm 2.447 * 0.0769\ ] = [\ 12.797 \leq \mu \leq 13.174]$

Wir können zu 95% darauf vertrauen, dass der wahre aber unbekannte Mittelwert der Grundgesamtheit in dem o.g. Intervall liegt. Technisch korrekte Interpretation: Wenn man 100 Stichproben nach dem o.g. Verfahren realisiert, werden von den 100 resultierenden Konfidenzintervallen im Mittel 95 den wahren Mittelwert beinhalten und 5 nicht. Das o.g. Konfidenzintervall [12.797, 13.174] ist eines dieser 100 Intervalle.

**Lösung mit R:**

```
# Dateneingabe
x <- c(12.7, 13.3, 13.0, 12.9, 13.1, 12.8, 13.1)
mean(x)
n <- length(x)

# Standardabweichung der Stichprobe
s <- sqrt(sum((x-mean(x))^2)/n); s_x
sqrt(sum(x^2)/n - mean(x)^2)

# Geschätzte Standardabweichung der Grundgesamtheit
sqrt(7/6)*s
sd(x)

# Geschätzte Standardabweichung des Stichrobenmittel
sd(x)/sqrt(n)
SE <- sd(x)/sqrt(n); SE

# kritischer t-Wert
t <- qt(0.975, df = n-1); t

# 95%-KI
KI_unten <- mean(x) - t*SE
KI_oben <- mean(x) + t*SE
KI_unten; KI_oben

# oder
t.test(x)
```

**Aufgabe 9.6: Eine kleine Stichprobe II**

*In einem Versicherungsunternehmen soll die Profitabilität von Versicherungspolicen geprüft werden. In der Tabelle sind die Gewinne (in €) von 30 zufällig ausgewählten Policen eines Mitarbeiters gegeben (vgl. auch Datensatz police.csv). Es kann angenommen werden, dass die Grundgesamtheit normalverteilt ist.*

| | | | | |
|---|---|---|---|---|
| 222.80 | 3255.60 | 57.90 | 1415.65 | 3249.65 |
| 1756.23 | 3701.85 | 833.95 | 2756.34 | -397.70 |
| 1100.85 | -803.35 | 1390.70 | 2089.40 | -397.31 |
| 3340.66 | 3865.90 | 2447.50 | 2692.75 | 186.25 |
| 1006.50 | 463.35 | 1847.50 | 2495.70 | 590.85 |
| 445.50 | -66.20 | 865.40 | 2172.70 | 578.95 |

*a) Konstruieren Sie ein 95%-Konfidenzintervall für den mittleren Gewinn des Mitarbeiters. Interpretation.*

*b) Diskutieren Sie die Annahmen für die Berechnung des Konfidenzintervalls.*

**Lösung a)**

Mit $n \leq 30$ liegt eine kleine Stichprobe vor. Wir müssen also mit der t-Verteilung arbeiten. Die Abweichung zur Lösung mit der $N(0,1)$-Verteilung (vgl. unten) ist relativ groß.
Wir berechnen Mittelwert und Standardabweichungen.

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1438.86$

$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \bar{x}^2} = 1307.256$

$\hat{\sigma} = \sqrt{\frac{n}{n-1}}s = 1329.604$

$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{1329.604}{\sqrt{30}} = 242.751$

Das 95%-Konfidenzintervall für den Mittelwert $\mu$ ist dann analog zu den Aufgaben zuvor zu bestimmen. Wir suchen das 95%-KI, d.h., links und rechts der kritischen Grenzen liegen 2.5% der Wahrscheinlichkeit. Wir suchen also das 97.5%-Quantil der t-Verteilung bei 29 Freiheitsgraden, $t_{[0.975]df=29} = 2.045$.

$KI(1-\alpha) = \left[\bar{x} - t_{[1-\alpha/2]}\hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{x} + t_{[1-\alpha/2]}\hat{\sigma}_{\bar{X}}\right]$

$KI(0.95) = [1438.86 \pm 2.045 * 242.751\,] = [942.43 \leq \mu \leq 1935.29]$

Wir können zu 95% darauf vertrauen, dass der wahre aber unbekannte mittlere Gewinn der Grundgesamtheit in dem o.g. Intervall liegt. Technisch korrekte Interpretation: Wenn man 100 Stichproben nach dem o.g. Verfahren realisiert, werden von den 100 resultierenden Konfidenzintervallen im Mittel 95 den wahren Mittelwert beinhalten und 5 nicht. Das o.g. Intervall [942.43, 1935.29] ist eines dieser 100 Intervalle.

**Lösung mit R:**

```
# Aufgabe 9.6: Eine kleine Stichprobe II
gewinn <- c(222.80, 3255.60,   57.90, 1415.65, 3249.65, 1756.23,
            3701.85,  833.95, 2756.34, -397.70, 1100.85, -803.35,
            1390.70, 2089.40, -397.31, 3340.66, 3865.90, 2447.50,
            2692.75,  186.25, 1006.50,  463.35, 1847.50, 2495.70,
             590.85,  445.50,  -66.20,  865.40, 2172.70,  578.95)

# Mit Anwendung der t-Verteilung
# -------------------------------------
n <- length(gewinn); n
x_dach <- mean(gewinn) # 1438.86
s <- sqrt(mean(gewinn^2) - mean(gewinn)^2); s # 1307.225
sigma <- sd(gewinn); sigma # 1329.604
sqrt(n/(n-1))*s
SE <- sigma/sqrt(30); SE # [1] 242.7513

t <- qt(0.975, df = n-1); t # 2.045
KI_unten <- x_dach - t*SE
KI_oben <- x_dach + t*SE
KI_unten; KI_oben # [942.38, 1935.34]
t.test(gewinn)

# Mit Anwendung der N(0,1)-Verteilung
# -------------------------------------
z <- qnorm(0.975); z # 1.96
KI_unten_z <- x_dach - z*SE
KI_oben_z <- x_dach + z*SE
KI_unten_z; KI_oben_z # [963.0786, 1914.646]
```

Mit der t-Verteilung wird ein breiteres Intervall erzeugt im Vergleich zum Intervall mit der $N(0,1)$-Verteilung. Es gilt: $t_{[0.975]df=29} > z_{[0.975]}$. Bei gleicher Vertrauenswahrscheinlichkeit ist das (korrekte) Konfidenzintervall mit der t-Verteilung weniger präzise als wenn die $N(0,1)$-Verteilung zugrunde gelegt werden würde. Eigentlich muss immer, wenn die Standardabweichung der Grundgesamtheit geschätzt werden muss, die t-Verteilung benutzt werden. Bei $n > 30$ kann aber auch approximativ die Normalverteilung genutzt werden.

**Lösung b)**

Annahmen für die Anwendung des Konfidenzintervalls mit der t-Verteilung:

- Unabhängigkeit der Beobachtungen. Bedingung: Zufallsauswahl und $n/N \leq 0.05$. Die Größe der Grundgesamtheit $N$ müsste also mindestens 600 betragen.
- Normalverteilung in der Grundgesamtheit. Bedingung: Für $n \leq 30$ muss die Grundgesamtheit näherungsweise normalverteilt sein, für $n > 30$ ist dies nicht mehr nötig, man kann approximativ die $N(0,1)$-Verteilung anwenden.

# Kapitel 10: Hypothesentests für eine Stichprobe

**Aufgabe 10.1: Hypothesen**

*Erstellen Sie $H_0$ und $H_A$ für die folgenden Situationen:*

*a) Der StuRa einer Hochschule führt eine Befragung der Studenten zu ihrem Einkommen durch. Eine Erhebung im letzten Jahr ergab, dass das Durchschnittseinkommen (arithmetisches Mittel) bei 550€/Monat lag. Es stellt sich die Frage, ob sich dieser Wert geändert hat.*

*b) Ein Marktforschungsunternehmen befragt zufällig ausgewählte Kunden nach ihrer Zufriedenheit mit einem Produkt eines Herstellers (Skala von 1/sehr gut bis 5/ungenügend). Eine neue Marketingmaßnahme soll anlaufen, wenn die Zufriedenheit schlechter als 3 ist.*

*c) Ein Busunternehmen möchte die durchschnittliche Körpergröße seiner Kunden wissen. Der Sitzabstand in den Bussen wird vergrößert, wenn die Durchschnittsgröße den Wert 181cm überschreitet.*

*d) Ein Marktforschungsunternehmen testet für einen Cola-Hersteller einen neuen Cola-Geschmack. Die Einführung des neuen Geschmacks ist geplant, wenn über 60% der Kunden diesen mögen.*

*e) Eine Immobilienagentur gibt bekannt, dass der Anteil an Häusern, die länger als drei Monate zum Verkauf stehen, nun über 50% liegt.*

*f) Ein Managermagazin berichtet, dass in 2005 35% aller CEOs einen MBA-Abschluss hatten. Es stellt sich die Frage, ob sich dieser Anteil geändert hat.*

*g) Ein Smartphone-Hersteller gibt bekannt, dass seine Ausschussrate unter 4% liegt.*

*h) Ein Pkw-Hersteller versucht durch ein neuartiges Bauteile die Reparaturhäufigkeit eines seiner Modelle zu reduzieren. Bislang mussten 20% aller Pkw dieses Modells nach 50000km zur Reparatur.*

*i) Ein Online-Lieferservice behauptet, dass der Anteil der pünktlich gelieferten Produkte erhöht wurde. Bislang lag dieser bei 90%.*

*Hinweis: Stellen Sie $H_0$ und $H_A$ so auf, dass die Ablehnung von $H_0$ und damit die Annahme von $H_A$ zu einem ökonomisch interessanten Resultat führt.*

**Lösung**

a) $H_0: \bar{Y} = 550€/Monat$ vs. $H_A: \bar{Y} \neq 550€/Monat$ (zweiseitiger Test)
b) $H_0: \bar{Z} = 3$ vs. $H_A: \bar{Z} > 3$ (oberseitiger Test)
c) $H_0: \bar{G} = 181cm$ vs. $H_A: \bar{G} > 181cm$ (oberseitiger Test)
d) $H_0: p = 0.6$ vs. $H_A: p > 0.6$ (oberseitiger Test)
e) $H_0: p = 0.5$ vs. $H_A: p > 0.5$ (oberseitiger Test)
f) $H_0: p = 0.35$ vs. $H_A: p \neq 0.35$ (zweiseitiger Test)
g) $H_0: p = 0.04$ vs. $H_A: p < 0.04$ (unterseitiger Test)
h) $H_0: p = 0.2$ vs. $H_A: p < 0.2$ (unterseitiger Test)
i) $H_0: p = 0.9$ vs. $H_A: p > 0.9$ (oberseitiger Test)

Auch bei einem einseitigen Test formulieren wir die Nullhypothese auf Gleichheit (=). Wir können dies tun, da alle anderen Werte für $p$ bzw. $\bar{x}$ (also Werte, die bei einem oberseitigen Test kleiner oder bei einem unterseitigen Test größer als sind als der Wert der Nullhypothese) einen geringeren $P$-value haben als der für Gleichheit berechnete $P$-value. Vgl. Lehrbuch hierzu auch S. 218 und S. 226.

***Aufgabe 10.2: Eisenstäbe I***

*Ein Hersteller behauptet, der Durchmesser von in Serie hergestellten Eisenstäben entspreche im Mittel dem Sollwert von* $10mm$*. Aus früheren Untersuchungen sei bekannt, dass der Durchmesser der Eisenstäbe normalverteilt ist und die produzierende Maschine mit einer Varianz von* $\sigma^2 = 0.49mm^2$ *arbeitet. Für die Verwendbarkeit der Stäbe beim Abnehmer ist zwar diese Streuung akzeptabel, dagegen wäre es unerwünscht, wenn der tatsächliche mittlere Durchmesser der in dieser Serie hergestellten Stäbe vom Sollwert nach oben oder nach unten abweicht. Im Interesse des Abnehmers soll bei einem Signifikanzniveau von 5% (bzw. 1%) die Behauptung des Herstellers durch einen Stichprobenbefund getestet werden. Die entnommene Stichprobe vom Umfang* $n = 100$ *Stäben liefert einen mittleren Durchmesser von* $\bar{x} = 9.85mm$*.*

**Lösung**

Es handelt sich um einen zweiseitigen Einstichproben-$z$-Test.

1. Schritt: Aufstellen von Null- und Alternativhypothese und Festlegung des Signifikanzniveaus

$H_0: \mu = 10mm$ vs. $H_A: \mu \neq 10mm$ für $\alpha = 0.05$

2. Schritt: Festlegung einer Testvariable und Bestimmung ihrer Verteilung

$Z = \frac{\bar{X}-\mu_0}{\sigma_{\bar{X}}}$ mit $Z \sim N(0,1)$ und $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Da die Standardabweichung der Grundgesamtheit bekannt ist, müssen wir diese nicht schätzen und können direkt mit der Standardabweichung des Stichprobenmittels rechnen. Die Testvariable $Z$ folgt einer $N(0,1)$-Verteilung, da $n > 30$.

3. Schritt: Berechnen des kritischen Wertes der Prüfgröße

$z_{[0.975]} = 1.96$

4. Schritt: Berechnen des empirischen Wertes der Testvariable (Prüfgröße)

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.49}}{\sqrt{100}} = 0.07mm$

$\bar{x} = 9.85mm$

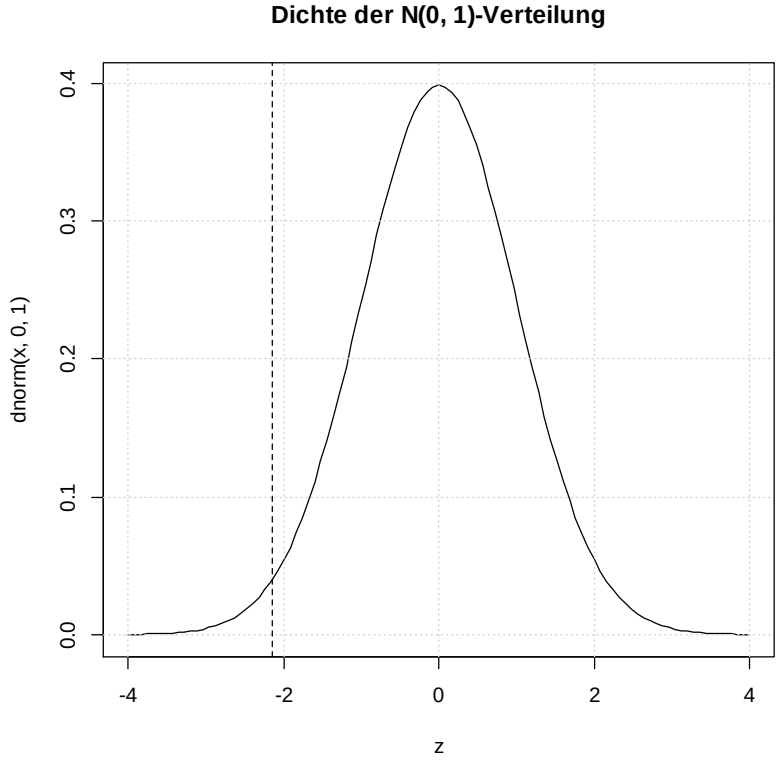
$z = \frac{\bar{x}-\mu_0}{\sigma_{\bar{X}}} = \frac{9.85-10}{0.07} = -2.143$

5. Schritt: Entscheidung über $H_0$ und Interpretation

Wir verwerfen $H_0$ und nehmen $H_A$ an, weil $|z| > z_{[0.975]}$. Die Behauptung des Herstellers muss also zu einem Signifikanzniveau von 5% verworfen werden.

Zum gleichen Ergebnis führt die Argumentation über den $P$-value: Der $P$-value liegt bei $2 \cdot \left(1 - F_{St}(|z|)\right) = 0.032$ und damit unter 5%. Die Wahrscheinlichkeit, Daten wie in der Stichprobe oder extremer zu beobachten, wenn $H_0$ wahr wäre, liegt somit bei nur $3.2\% < 5\%$. Wir verwerfen daher die Nullhypothese zugunsten der Alternativhypothese.

Die Dichte der Testvariable $Z \sim N(0,1)$ und der Wert der Prüfgröße $z = -2.143$ sind in der folgenden Abbildung dargestellt. Der $P$-value ist zwei Mal die Fläche links der gestrichelten Linie bei $= -2.143$. Interpretation des $P$-value: Wenn der tatsächliche mittlere Durchmesser der Stäbe gleich 10mm ist, dann ist die Wahrscheinlichkeit bei einer Zufallsstichprobe von 100 Stäben eine Abweichung von 10mm in Höhe von 0.15mm oder mehr zu finden, gleich 3.2%.




Bei einem Signifikanzniveau von 1% ist der kritische Wert der Prüfgröße $z_{[0.995]} = 2.58$. In diesem Fall kann $H_0$ nicht verworfen werden, da der beobachtete Wert der Prüfgröße nicht größer ist als der kritische Wert. $H_0$ wird also beibehalten. Entsprechend: Der $P$-value liegt über 1%, daher behalten wir die Nullhypothese bei.

**Lösung mit R:**

```
# Berechnung der Quantile
qnorm(0.975); qnorm(0.995) # Quantile der N(0,1)-Verteilung
2*(1 - pnorm(abs(-2.143))) # Berechnung des p-values

# Darstellung der Dichte von N(1,0)
curve(dnorm(x, 0, 1), xlab = "z",
main = "Dichte der N(0, 1)-Verteilung",
xlim = c(-4, 4))
grid()
abline(v = -2.143, lty = 2)
# Der p-value ist 2 Mal die Fläche
# links von der gestrichelten Linie bei z = -2.143
```

***Aufgabe 10.3: Eisenstäbe II***

*Aufgabenstellung wie bei Aufgabe 10.2 („Eisenstäbe I"), die Varianz $\sigma^2$ der Grundgesamtheit sei nun aber unbekannt. Aus einer kleinen (!) Stichprobe vom Umfang mit $n = 25$ errechnet man $\bar{x} = 9.79mm$ sowie $s^2 = 0.64mm^2$.*

**Lösung**

Es handelt sich um einen zweiseitigen Einstichproben-$t$-Test

1. Schritt: Aufstellen von Null- und Alternativhypothese und Festlegen des Signifikanzniveaus

$H_0: \mu = 10mm$ vs. $H_A: \mu \neq 10mm$ für $\alpha = 0.05$

2. Schritt: Festlegung einer Testvariable und Bestimmung ihrer Verteilung

$T = \frac{\bar{X}-\mu_0}{SE} \sim T_{df=n-1}$ mit $SE = \hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}}$ mit $\hat{\sigma} = \sqrt{\frac{n}{n-1}}s$

Da die Standardabweichung der Grundgesamtheit unbekannt ist, müssen wir diese schätzen und können nicht direkt mit der Standardabweichung des Stichprobenmittels rechnen. Die Testvariable $T$ folgt einer $t$-Verteilung mit $n-1$ Freiheitsgraden (hier also $df = 24$), da die Grundgesamtheit annahmegemäß normalverteilt ist.

3. Schritt: Berechnen des kritischen Wertes der Prüfgröße

$t_{df=24[0.975]} = 2.064$ (man beachte, dieser Wert ist größer als der entsprechende $z$-Wert)

4. Schritt: Berechnen des empirischen Wertes der Testvariable (Prüfgröße)

$SE = \hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{25}{24}}\frac{\sqrt{0.64}}{\sqrt{25}} = 0.162mm$

$\bar{x} = 9.79mm$

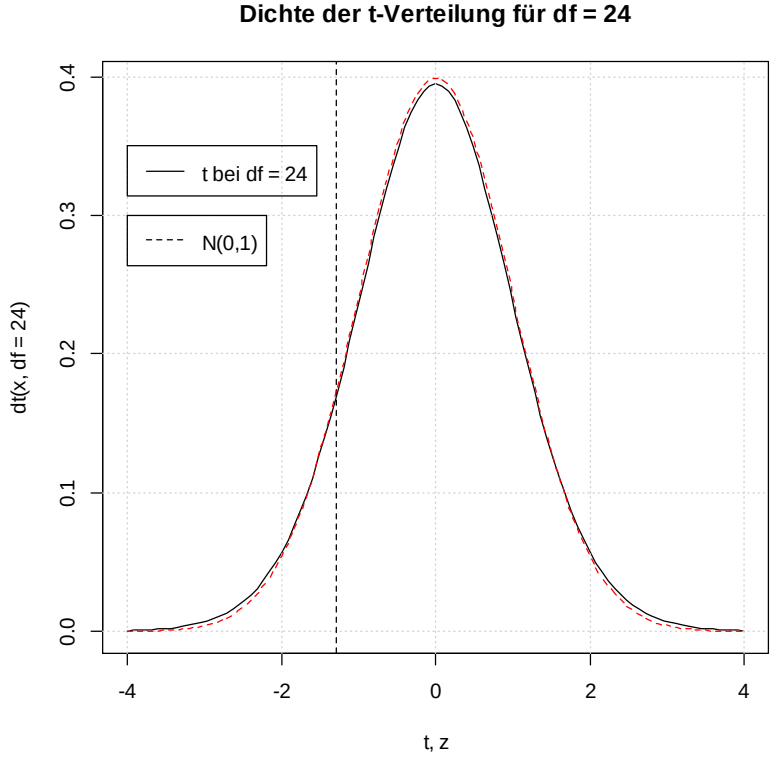
$t = \frac{\bar{x}-\mu_0}{SE} = \frac{9.79-10}{0.162} = -1.3$ und damit $|t| \not> t_{df=24[0.975]}$

5. Schritt: Entscheidung über $H_0$ und Interpretation

Bei einem Signifikanzniveau von 5% ist der kritische Wert der Prüfgröße $t_{df=24[0.975]} = 2.064$. In diesem Fall kann $H_0$ nicht verworfen werden, da der Betrag der Prüfgröße nicht größer ist als der kritische Wert. $H_0$ wird also beibehalten.

Zum gleichen Ergebnis führt die Argumentation über den $P$-value: Der $P$-value muss laut Tabelle der $t$-Verteilung bei über 20% liegen, denn das 90%-Quantil der $t$-Verteilung bei $df = 24$ ist $t_{df=24[0.9]} = 1.318$ und der beobachtete Wert für $t$ ist betragsmäßig kleiner als dieser Wert. Der hohe $P$-value lässt uns die Nullhypothese beibehalten.

Die Abbildung zeigt die Dichte der t-Verteilung bei $df = 24$. Dazu den Wert für $t = -1.3$, aus dem sich der $P$-value berechnen lässt mit $P\text{-value} = 2 \cdot \left(1 - F_{t,df=24}(|t|)\right) = 0.2059$. Zusätzlich ist rot gestrichelt die Dichte der $N(0,1)$-Verteilung eingezeichnet.



Am Ergebnis ändert sich beim Signifikanzniveau $\alpha = 0.01$ nichts. Wenn wir $H_0$ bei $\alpha = 0.05$ nicht verwerfen, dann tun wir dies auch bei $\alpha = 0.01$ nicht.

**Lösung mit R:**

```
# Berechnung der Quantile
qt(0.975, df = 24); qt(0.995, df = 24) # Quantile der t-Verteilung
2*(1 - pt(abs(-1.3), df = 24)) # Berechnung des p-values = 0.2059

# Darstellung der Dichte der t-Verteilung mit df = 24
curve(dt(x, df = 24), xlab = "t, z",
main = "Dichte der t-Verteilung für df = 24",
xlim = c(-4, 4))
grid()
abline(v = -1.3, lty = 2)
curve(dnorm(x, 0, 1), lty = 2, add = TRUE, col = "red")
legend(-4, 0.3, lty = 2, "N(0,1)")
legend(-4, 0.35, lty = 1, "t bei df = 24")
# Der p-value ist 2 Mal die Fläche
# links von der gestrichelten Linie bei t = -1.3
```

***Aufgabe 10.4: Flottenverbrauch***

*Ein Unternehmen mit einem großen Fuhrpark versucht durch Fahrhinweise für die Angestellten den Kraftstoffverbrauch zu reduzieren. Ziel ist ein Durchschnittsverbrauch von höchstens 10 Liter/100km. Um zu überprüfen, ob dieses Ziel eingehalten wird, werden zufällig 50 Fahrten ausgewählt. Dabei ergibt sich ein Durchschnittsverbrauch von 11.4 Liter/100km bei einer Standardabweichung von 2.3 Liter/100km. Lässt sich daraus schließen, dass das Ziel verfehlt wurde?*

*a) Erstellen Sie die passenden Hypothesen.*
*b) Prüfen Sie die Annahmen für einen Hypothesentest.*
*c) Führen Sie einen Hypothesentest durch. Das Signifikanzniveau sei 5%. Ermitteln Sie den P-value und interpretieren Sie diesen Wert.*

**Lösung**

a)

$H_0: \mu = 10\ l/100km$ vs. $H_A: \mu > 10\ l/100km$ (oberseitiger Test)

Beachte: Das ökonomisch interessante Ergebnis (hier: das Verfehlen des Ziels nach oben) wird als Alternativhypothese formuliert.

b)

Die Annahme einer Zufallsstichprobe ist erfüllt, da die Stichprobe aus 50 zufällig ausgewählten Fahrten besteht. Wir können außerdem davon ausgehen, dass $n = 50$ weniger als 5% der gesamten Anzahl an Fahrten des Unternehmens darstellt (Annahme von Unabhängigkeit der Beobachtungen bei Ziehen ohne Zurücklegen).

c)

Es handelt sich um einen oberseitigen Einstichproben-$t$-Test

1. Schritt: Aufstellen von Null- und Alternativhypothese und Festlegen des Signifikanzniveaus

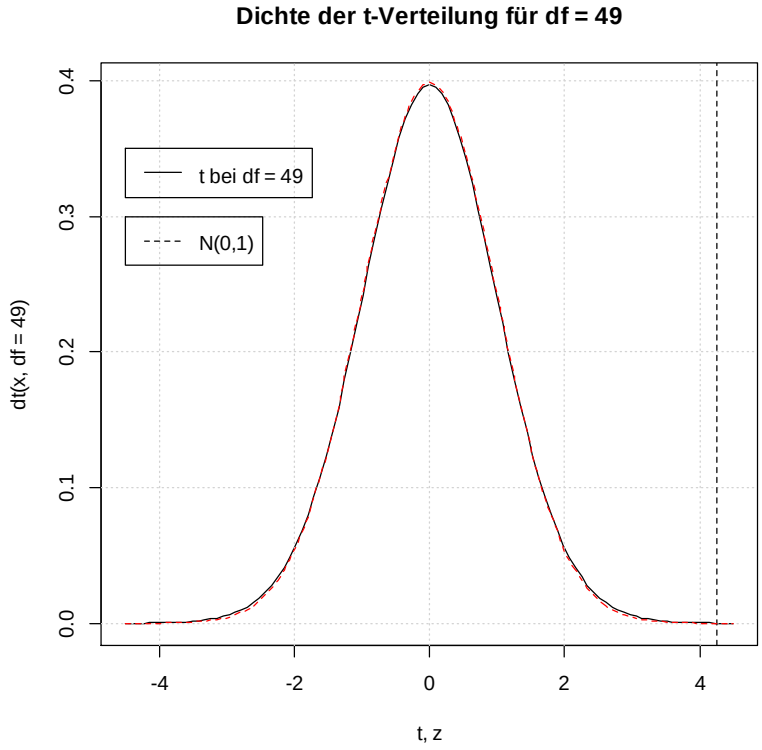
$H_0: \mu = 10\ l/100km$ vs. $H_A: \mu > 10\ l/100km$ für $\alpha = 0.05$

2. Schritt: Festlegung einer Testvariable und Bestimmung ihrer Verteilung

$Z = \frac{\bar{X} - \mu_0}{SE} \sim N(0,1)$ mit $SE = \hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}}$ und $\hat{\sigma}_X = \sqrt{\frac{n}{n-1}}s$

Da die Standardabweichung der Grundgesamtheit unbekannt ist, müssen wir diese schätzen und können nicht direkt mit der Standardabweichung des Stichprobenmittels rechnen. Die

Prüfgröße ergibt sich aus folgender Überlegung: Da $\sigma_X$ geschätzt werden muss, folgt die standardisierte Größe $\frac{\bar{X}-\mu_0}{SE}$ einer $t$-Verteilung mit $n-1$ Freiheitsgraden. Da aber $n > 30$ ist, können wir die $t$-Verteilung mit der $N(0,1)$-Verteilung approximieren. Wir arbeiten also im Folgenden mit der Testvariable $Z$ (mit R wird dagegen immer die $t$-Verteilung genutzt).
3. Schritt: Berechnen des kritischen Wertes der Prüfgröße
$z_{[0.95]} = 1.645$
4. Schritt: Berechnen des empirischen Wertes der Testvariable (Prüfgröße)
$SE = \hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{50}{49}}\frac{2.3}{\sqrt{50}} = 0.329\ l/100km$
$\bar{x} = 11.4\ l/100km$
$z = \frac{\bar{x}-\mu_0}{SE} = \frac{11.4-10}{0.329} = 4.255$
5. Schritt: Entscheidung über $H_0$ und Interpretation
Bei einem Signifikanzniveau von 5% ist der kritische Wert der Prüfgröße $z_{[0.95]} = 1.645$. In diesem Fall kann $H_0$ verworfen werden, da der beobachtete Wert der Prüfgröße mit $z = 4.255$ deutlich größer ist als der kritische Wert. $H_0$ wird also abgelehnt und $H_A$ angenommen. Der $P$-value ist extrem klein: $P$-value = $1 - F_{St}(4.255) \approx 0$. Die Wahrscheinlichkeit, Daten wie in der Stichprobe oder extremer zu beobachten, wenn die Nullhypotese wahr wäre, ist also nahe Null. Hier im Kontext: Wenn der mittlere Verbrauch tatsächlich 10 l/100km oder weniger ist, dann ist die Wahrscheinlichkeit bei einer Zufallsstichprobe von 50 Fahrten einen durchschnittlichen Verbrauch von 11.4 l/100km oder mehr zu finden, praktisch gleich Null. Dies lässt uns hinsichtlich der Gültigkeit der Null sehr skeptisch sein, und daher wird die Null verworfen und die Alternativhypothese akzeptiert. Wir können also davon ausgehen, dass das Ziel verfehlt wurde.
Die Abbildung zeigt die Dichte der t-Verteilung bei $df = 49$. Dazu den Wert für $t = z = 4.255$, aus dem sich der $P$-value berechnen lässt mit $P$-value $= 1 - F_{t,df=49}(|t|) = 0.000$. Zusätzlich ist rot gestrichelt die Dichte der $N(0,1)$-Verteilung eingezeichnet.



**Lösung mit R:**

```
# Berechnung der Quantile
qt(0.95, df = 49) # Quantil der t-Verteilung
qnorm(0.95) # Quantil der z-Verteilung
1 - pt(abs(4.255), df = 49) # Berechnung des p-values < 0.001
1 - pnorm(abs(4.255)) # Berechnung des p-values < 0.001

# Darstellung der Dichte der t-Verteilung mit df = 49
curve(dt(x, df = 49), xlab = "t, z",
main = "Dichte der t-Verteilung für df = 49",
xlim = c(-4.5, 4.5))
grid()
abline(v = 4.255, lty = 2)
curve(dnorm(x, 0, 1), lty = 2, add = TRUE, col = "red")
legend(-4.5, 0.3, lty = 2, "N(0,1)")
legend(-4.5, 0.35, lty = 1, "t bei df = 49")
# Der p-value ist die Fläche
# rechts von der gestrichelten Linie bei t = 4.255
```

***Aufgabe 10.5: Kneipe***

*Sie vermuten, dass in Ihrer Stammkneipe der Wirt weniger Bier ausschenkt als eigentlich gefordert. Mit ein paar Freunden trinken Sie an einem Abend sieben 0.5-Liter-Bier und messen die Füllmengen nach. Die Füllmengen der Biergläser (in Liter) sind:*
*0.51 0.48 0.5 0.45 0.47 0.48 0.47*

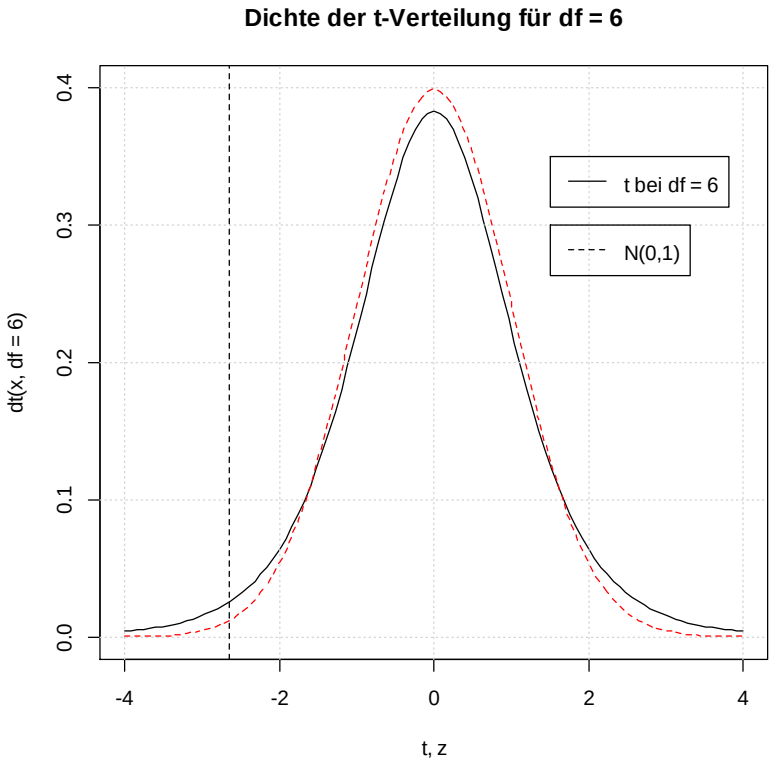
*a) Sie wollen zeigen, dass der Wirt tatsächlich zu wenig Bier ausschenkt. Führen Sie einen entsprechenden Hypothesentest mit einem Signifikanzniveau von 5% durch. Erläutern Sie Ihre Vorgehensweise. Interpretieren Sie Ihr Ergebnis.*
*b) Welche Annahme müssen Sie hinsichtlich der Grundgesamtheit treffen?*

**Lösung a)**
Es handelt sich um einen unterseitigen Einstichproben-$t$-Test.
1. Schritt: Aufstellen von Null- und Alternativhypothese und Festlegung des Signifikanzniveaus
$H_0: \mu = 0.5\ l$ vs. $H_A: \mu < 0.5\ l$ mit $\alpha = 0.05$
2. Schritt: Festlegen einer Testvariable und Bestimmung ihrer Verteilung
$T = \frac{\bar{X}-\mu_0}{SE} \sim T_{df=n-1}$ mit $SE = \hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}}$ und $\hat{\sigma}_X = \sqrt{\frac{n}{n-1}}s$
Da die Standardabweichung der Grundgesamtheit unbekannt ist, müssen wir diese schätzen und können nicht direkt mit der Standardabweichung des Stichprobenmittels rechnen. Die Testvariable ergibt sich aus folgender Überlegung: Da $\sigma$ geschätzt werden muss, folgt die standardisierte Größe $\frac{\bar{X}-\mu_0}{SE}$ einer $t$-Verteilung. Da $n \leq 30$ ist, müssen wir mit der $t$-Verteilung arbeiten.
3. Schritt: Berechnen des kritischen Wertes der Prüfgröße
$t_{df=6[0.05]} = -1.943$ (vgl. Tabelle der $t$-Verteilung)
4. Schritt: Berechnen des empirischen Wertes der Testvariable (Prüfgröße)
$s = \sqrt{\frac{1}{n}\sum_{i=1}^{7}(x_i - \bar{x})^2} = 0.185\ l$ (Standardabweichung der Stichprobe)
$SE = \hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{7}{6}}\frac{0.185}{\sqrt{7}} = 0.00756\ l$
$\bar{x} = 0.48\ l$ (Mittelwert der Beobachtungen)
$t = \frac{\bar{x}-\mu_0}{SE} = \frac{0.48-0.5}{0.00756} = -2.646$

5. Schritt: Entscheidung über $H_0$ und Interpretation
Da $t < t_{df=6[0.05]}$ muss die Nullhypothese verworfen und die Alternativhypothese angenommen werden. Wir wissen aus der Tabelle der $t$-Verteilung, dass der $P$-value unter 2.5% liegt, denn $t_{df=6[0.025]} = -2.447$ und der beobachtete Wert für $t$ ist kleiner als das 2.5%-Quantil der $t$-Verteilung bei 6 Freiheitsgraden.
Wir können also davon ausgehen, dass der Wirt weniger ausschenkt als 0.5 Liter.
Die Abbildung unten zeigt die Dichte der t-Verteilung bei $df = 6$. Dazu den Wert für $t = -2.65$, aus dem sich der $P$-value berechnen lässt mit $P$-value $= F_{t,df=6}(t) = 0.019$.
Zusätzlich ist rot gestrichelt die Dichte der $N(0,1)$-Verteilung eingezeichnet.

**Lösung b)**
Es muss gelten, dass die Grundgesamtheit normalverteilt ist. Dies kann hier als erfüllt angesehen werden, da zufällige Abweichungen (beim Zapfen) von einem Mittelwert häufig einer Normalverteilung folgen.

**Lösung mit R:**
```
bier <- c(0.51, 0.48, 0.5, 0.45, 0.47, 0.48, 0.47)

# Methode 1
mu_0 <- 0.5                          # Mittelwert der Nullhypothese
sigma_dach <- sd(bier); sigma_dach  # geschätzte SD der GG = 0.02
x_dach <- mean(bier); x_dach        # Stichprobenmittel = 0.48
n <- 7; n                            # Stichprobengröße = 7
SE <- sigma_dach/sqrt(n); SE        # SE = 0.007559289

t <- (x_dach - mu_0)/SE; t          # emp. Wert der Testvar. t = -2.645751
t_krit <- qt(df = n - 1, 0.05); t_krit  # krit. Wert = -1.94318
# t < t_krit => Ablehnung von H0

# Methode 2
pt(df = n-1, t)                          # P-value = 0.01912259
# P-value < 0.05 => Ablehnung von H0

# Methode 3: t-Test in R
t.test(bier, mu = 0.5, alternative = "less")

# Darstellung der Dichte der t-Verteilung mit df = 6
curve(dt(x, df = 6), xlab = "t, z", ylim = c(0, 0.4),
main = "Dichte der t-Verteilung für df = 6",
xlim = c(-4, 4))
grid()
abline(v = -2.6458, lty = 2)
curve(dnorm(x, 0, 1), lty = 2, add = TRUE, col = "red")
legend(1.5, 0.3, lty = 2, "N(0,1)")
legend(1.5, 0.35, lty = 1, "t bei df = 6")
# Der p-value ist die Fläche
# links von der gestrichelten Linie bei t = -2.6458
```




***Aufgabe 10.6: Umfrage***

*Über das Jahr 2007 wurden in den USA mehrere Befragungen zur Einschätzung der zukünftigen ökonomischen Lage durchgeführt. Im Mittel gaben 20% der Befragten an, dass sie optimistisch hinsichtlich der zukünftigen ökonomischen Lage sind. Im Januar 2008 wurden zufällig 2590 Wahlberechtigte ausgewählt und nur 13% gaben an, dass sie optimistisch sind. Ist dies Evidenz dafür, dass das Vertrauen der US-Bürger in die zukünftige ökonomische Lage gesunken ist?*

*a) Erstellen Sie die passenden Hypothesen.*
*b) Prüfen Sie, ob die Bedingungen für einen Hypothesentest erfüllt sind.*
*c) Führen Sie einen Hypothesentest durch. Das Signifikanzniveau sei 5%. Ermitteln Sie den P-value und interpretieren Sie diesen Wert.*

**Lösung a)**
Es handelt sich um einen unterseitigen Anteilswert-$z$-Test.
Aufstellen von Null- und Alternativhypothese und Festlegung des Signifikanzniveaus
$H_0: p = 0.2$ vs. $H_A: p < 0.2$ mit $\alpha = 0.05$

**Lösung b)**
Folgende Bedingungen müssen erfüllt sein (i) Es muss eine Zufallsstichprobe vorliegen, (ii) $\frac{n}{N} \leq 0.05$ und (iii) $np_0 \geq 10$ und $n(1-p_0) \geq 10$.
(i) ist erfüllt,
(ii) In den USA gibt es ca. 230 Mio. Wahlberechtigte (hier $N$). Damit ist $\frac{n}{N} \leq 0.05$ erfüllt.
(iii) Es gilt $np_0 = 2590 \cdot 0.2 \geq 10$ und $n(1-p_0) = 2590 \cdot 0.8 \geq 10$.

**Lösung c)**
1. Schritt: Aufstellen von Null- und Alternativhypothese (vgl. oben)
$H_0: p = 0.2$ vs. $H_A: p < 0.2$ mit $\alpha = 0.05$
2. Schritt: Festlegen der Testvariable und Bestimmung ihrer Verteilung

$Z = \frac{H-p_0}{\sigma_H}$ mit $\sigma_H = \sqrt{\frac{p_0(1-p_0)}{n}}$

Da wir in $H_0$ einen wahren Wert für $p$ annehmen ($p = p_0$), können wir mit $p_0$ auch die Standardabweichung des Stichprobenmittels $\sigma_H$ berechnen.

3. Schritt: Berechnen des kritischen Wertes der Prüfgröße

$z_{[0.05]} = -1.645$ (vgl. Tabelle der $N(0,1)$-Verteilung)

4. Schritt: Berechnen des empirischen Wertes der Testvariable (Prüfgröße)

$\sigma_H = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.2(1-0.2)}{2590}} = 0.00786$ (Standardabweichung des Stichprobenmittels)

$z = \frac{h-p_0}{\sigma_H} = \frac{0.13-0.2}{0.00786} = -8.906$

5. Schritt: Entscheidung über $H_0$ und Interpretation

Wir können $H_0$ ablehnen, da $z < z_{[0.05]}$. Es gibt starke Evidenz dafür, dass das Vertrauen der US-Bürger in die zukünftige ökonomische Lage gesunken ist.

Der $P$-value ist hier extrem klein ($F_{St}(-8.906) < 0.001$). Es ist also extrem unwahrscheinlich, Daten wie in der Stichprobe oder extremer zu beobachten, wenn die Nullhypothese wahr wäre.

**Lösung mit R:**

```
# Methode 1
p0 <- 0.2; p0                 # Anteilswert der Nullhypothese
n <- 2590; n                  # Stichprobengröße
sigma_H <- sqrt(p0*(1-p0)/n); sigma_H  # SD des St.pr.anteils
# 0.007859775
h <- 0.13; h        # Stichprobenanteil

z <- (h - p0)/sigma_H; z      # Prüfgröße
# -8.906107

z_krit <- qnorm(0.05); z_krit     # krit. Wert
# -1.644854
# z < z_krit => Ablehnung von H0

# Methode 2
pnorm(z)                          # P-value
# 2.642785e-19 (2.642 * 10 hoch -19)
# P-value < 0.05 => Ablehnung von H0

# Methode 3
# Anteilswert-z-Test in R
prop.test(0.13*2590, 2590, p = 0.2, alternative = "less")
# Die die Approximation der diskreten Binomialverteilung
# durch die stetige Normalverteilung kommt es zu einer
# kleinen Abweichung beim p-value
```

# Kapitel 11: Hypothesentests für zwei Stichproben und Verteilungen qualitativer Daten

***Aufgabe 11.1: Ausgabenverhalten***

*Auf einer Freizeitmesse wird das Ausgabeverhalten von Frauen und Männern erhoben. Es ergeben sich folgende Daten (vgl. auch* `two_sample_t_test.csv`*).*

| | *Stichprobengröße* | *Mittelwert in €* | *Standardabweichung in €* |
|---|---|---|---|
| *Frauen* | $n_X = 100$ | $\bar{x} = 101.347$ | $\hat{\sigma}_X = 12.798$ |
| *Männer* | $n_Y = 120$ | $\bar{y} = 105.991$ | $\hat{\sigma}_Y = 17.617$ |

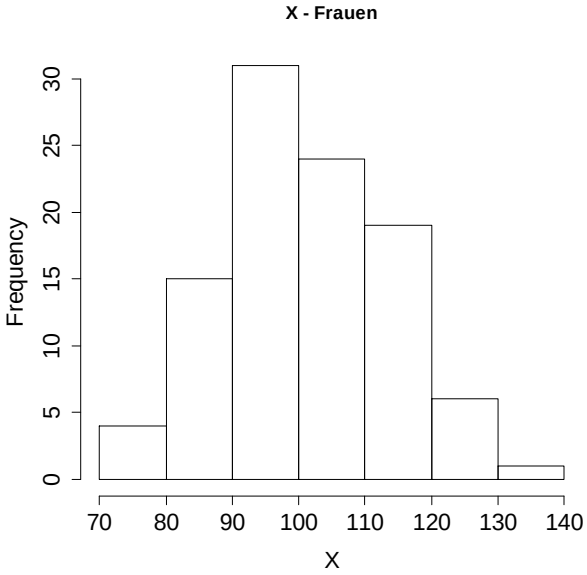
*a) Prüfen Sie die Annahmen zur Durchführung eines Zweistichproben-t-Tests.*
*b) Unterscheiden sich die mittleren Ausgaben für Frauen und Männer? Führen Sie einen Zweistichproben-t-Test durch. Das Signifikanzniveau sei 5%. Interpretation.*
*c) Berechnen Sie das zugehörige 95%-Konfidenzintervall. Interpretation.*

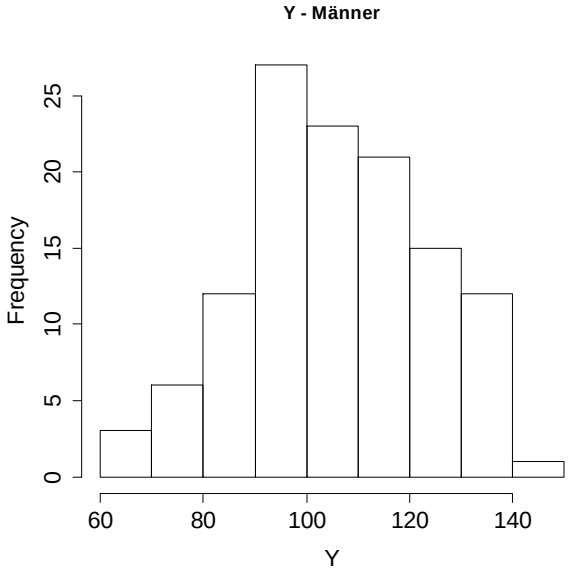
**Lösung a)**

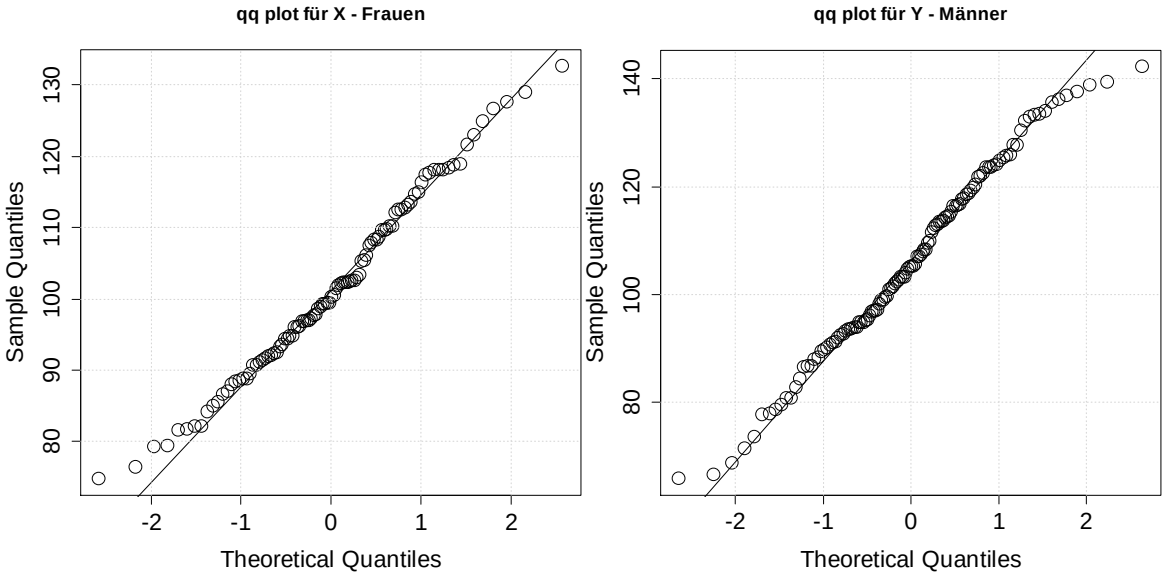
Folgende Annahmen sind für den Zweistichproben-t-Test zu prüfen:

1. Unabhängigkeit der Beobachtungen: Hierzu haben wir hier keine Informationen. Wir nehmen an, dass die Frauen als Zufallsstichprobe aus der Grundgesamtheit der Frauen gezogen wurden. Gleiches gilt für die Männer.
2. Normalverteilungsannahme: Zu prüfen ist, ob die Werte in der Stichprobe normalverteilt sind. Hierzu ist ein Histogramm oder – besser – ein qq-plot (vgl. unten) mit R zu erstellen. Die Darstellung ergibt, dass diese Annahme hier als erfüllt angesehen werden kann.
3. Unabhängige Gruppen: Es darf keinerlei Beeinflussung der Gruppen untereinander geben. Hier sollte also sichergestellt sein, dass sich die Frauen und Männer nicht kennen bzw. beeinflussen.

Die Histogramme bzw. qq-plots zeigen, dass Annahme 2 als erfüllt angesehen werden kann.






**Lösung b)**

Zunächst sind Null- und Alternativhypothese zu formulieren und die Testvariable anzugeben. Die Nullhypothese lautet

$H_0: \mu_X - \mu_Y = 0$

bei einem Signifikanzniveau von $\alpha = 0.05$. Die Alternativhypothese ist zweiseitig

$H_0: \mu_X - \mu_Y \neq 0$

Die Testvariable für diesen Test ist

$T = \frac{(\bar{X}-\bar{Y})}{SE(\bar{X}-\bar{Y})}$ mit $T \sim T_{df}$

wobei der Standardfehler ($SE$) berechnet wird mit

$$SE(\bar{X} - \bar{Y}) = \sqrt{\frac{\hat{\sigma}_X^2}{n_X} + \frac{\hat{\sigma}_Y^2}{n_Y}}$$

Die Testvariable $T$ folgt einer $t$-Verteilung mit $df$ Freiheitsgraden, wobei

$$df = \frac{\left(\frac{\hat{\sigma}_X^2}{n_X} + \frac{\hat{\sigma}_Y^2}{n_Y}\right)^2}{\frac{1}{n_X - 1}\left(\frac{\hat{\sigma}_X^2}{n_X}\right)^2 + \frac{1}{n_Y - 1}\left(\frac{\hat{\sigma}_Y^2}{n_Y}\right)^2}$$

Diese Berechnungen sind mit R durchzuführen (vgl. unten). Der beobachtete Unterschied im mittleren Ausgabenverhalten ist $\bar{x} - \bar{y} = -4.644$. Wir erhalten hier für den Standardfehler $SE = 2.055$. Die Prüfgröße (d.h. die Realisation der Testvariablen) ist $t = \frac{101.347 - 105.991}{2.055} = -2.259$. Die Zahl der Freiheitsgrade ist $df = \frac{17.844}{0.083} = 214$. Der kritische Wert der $t$-Verteilung bei $df = 214$ ist $t_{[0.975]df=214} = 1.971$. Wir können daher die Nullhypothese verwerfen, da $|t| > t_{[0.975]df=214}$.

Der exakte $P$-value mit R zu berechnen. Es ergibt sich $2\big(1 - F_{t,df=214}(|t|)\big) = 0.0248$. Wenn wir, da $df > 100$, die $t$-Verteilung durch die $N(0,1)$-Verteilung approximieren, erhalten wir nach Tab. 14.1 einen $P$-value von $2(1 - F_{St}(|-2.26|)) = 2(1 - 0.9881) = 0.0238$.

Wir können also die Nullhypothese, nach der es zwischen dem mittleren Ausgabenverhalten von Männern und Frauen keinen Unterschied gibt, verwerfen. Wenn es tatsächlich keinen Unterschied im mittleren Ausgabenverhalten gäben würde, liegt die Wahrscheinlichkeit, bei einer solchen Zufallsstichprobe eine Differenz der Mittelwerte von 4.644 oder größer zu erhalten, bei ca. 2.5%. Auf Grund dieser geringen Wahrscheinlichkeit, verwerfen wir die

Nullhypothese und nehmen die Alternativhypothese an, nach der sich das mittlere Ausgabenverhalten zwischen Frauen und Männern unterscheidet.

**Lösung c)**

Das zum Zweistichproben-$t$-Test korrespondierende Konfidenzintervall für den Unterschied der Mittelwerte ist

$$KI(1-\alpha)_{\mu_X-\mu_Y} = (\bar{x}-\bar{y}) \pm t_{df[1-\frac{\alpha}{2}]} SE(\bar{X}-\bar{Y})$$

Wir erhalten mit R

$$KI(0.95)_{\mu_X-\mu_Y} = (-4.644) \pm 1.971 * 2.055 = [-0.59, +8.70]$$

Wir können also zu 95% darauf vertrauen, dass ein mit dieser Methode erzeugtes Intervall, die wahre mittlere Ausgabendifferenz zwischen Männern und Frauen beinhaltet. Unser Intervall [-0.59€, +8.70€] ist ein solches Intervall.

**Lösung mit R**

```
getwd()
data <- read.csv("two_sample_t_test.csv")
head(data)
attach(data)
# ---------------------------------------
# a)
# X: Frauen, Y: Männer
# Histogramm und qq-Plot
hist(X, main = "X - Frauen", cex.axis = 1.5, cex.lab = 1.5)
qqnorm(X, cex.axis = 1.5, cex.lab = 1.5, cex = 2,
main = "qq plot für X - Frauen"); qqline(X); grid()
hist(Y, main = "Y - Männer", cex.axis = 1.5, cex.lab = 1.5)
qqnorm(Y, cex.axis = 1.5, cex.lab = 1.5, cex = 2,
main = "qq plot für Y - Männer"); qqline(B); grid()
# ---------------------------------------
# b)
# Methode 1
# Beobachtungen pro Vektor
n_X <- sum(table(X)); n_X
n_Y <- sum(table(Y)); n_Y
mean_X <- mean(X, na.rm = TRUE)
mean_Y <- mean(Y, na.rm = TRUE)
var_X <- var(X, na.rm = TRUE); var_X
var_Y <- var(Y, na.rm = TRUE); var_Y

SE <- sqrt(var_X/n_X + var_Y/n_Y); SE
t <- (mean_X - mean_Y)/SE; t

# T folgt einer t-Verteilung mit df als Zahl Freiheitsgrade:
# df = Z / N
Z <- (var_X/n_X + var_Y/n_Y)^2
N <- (1/(n_X - 1))*(var_X/n_X)^2 + (1/(n_Y - 1))*(var_Y/n_Y)^2
df <- Z/N; df
# kritischer Wert der t-Verteilung
qt(0.975, df = 214)
# p-value exakt
p.value <- 2*(1 - pt(abs(t), df)); p.value
p.value <- 2*(1 - pt(abs(t), 214)); p.value
# p-value approx.
p.value <- 2*(1 - pnorm(abs(t))); p.value

# Methode 2
t.test(X, Y)
# ---------------------------------------
# c)
# Methode 1
KI_unten <- (mean_X - mean_Y) - qt(0.975, df = df)*SE
KI_oben <- (mean_X - mean_Y) + qt(0.975, df = df)*SE
KI_unten; KI_oben
```

```
# Methode 2
t.test(X, Y)
```

***Aufgabe 11.2: Tapete***

*Auf einer Möbelmesse werden 16 zufällig ausgewählte Besucher nach ihrer Zahlungsbereitschaft für Tapete (in € pro Packung) befragt, wobei die Tapete ökologisch („Öko“) oder konventionell („kein Öko“) hergestellt wurde. Bis auf die Art der Herstellung („Öko“ oder „kein Öko“) gibt es keine Produktunterschiede. Die Tabelle zeigt die Werte der Zahlungsbereitschaften (vgl. auch `paired_t_test.csv`).*

| *Besucher* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *kein Öko* | *36* | *18* | *16* | *83* | *49* | *7* | *27* | *89* | *45* | *30* | *40* | *29* | *32* | *15* | *29* | *38* |
| *Öko* | *37* | *21* | *17* | *85* | *48* | *8* | *29* | *92* | *49* | *32* | *41* | *30* | *32* | *13* | *29* | *38* |

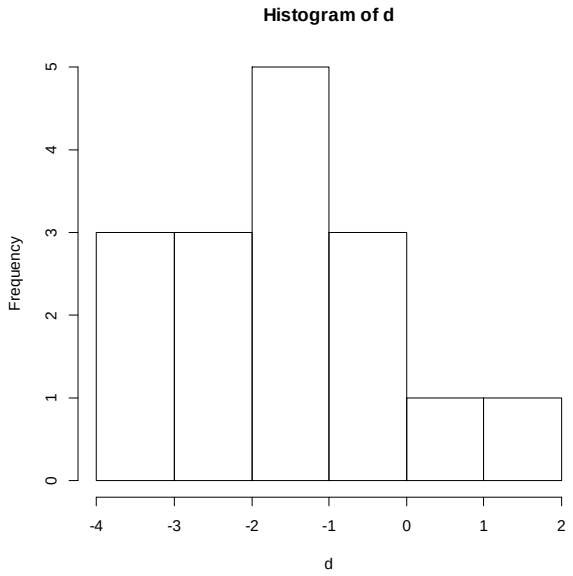
*a) Prüfen Sie die Annahmen zur Durchführung eines gepaarten $t$-Tests.*
*b) Führen Sie einen gepaarten $t$-Test durch. Das Signifikanzniveau sei 5%. Interpretation.*
*c) Berechnen Sie das zugehörige 95%-Konfidenzintervall. Interpretation.*

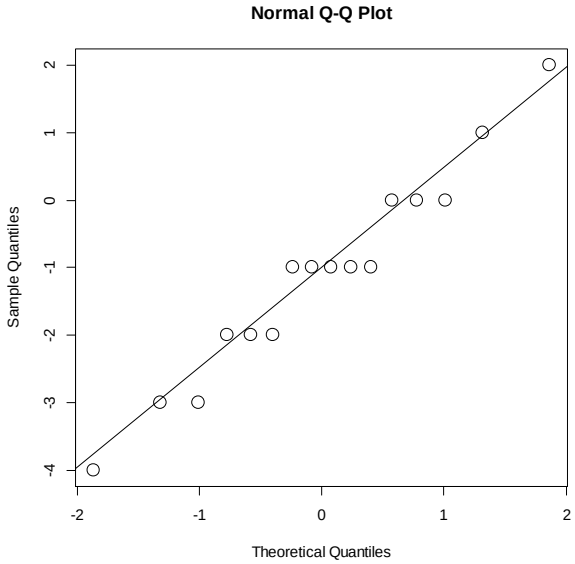
**Lösung a)**

Folgende Annahmen sind für den gepaarten t-Test zu prüfen:

1. Gepaarte Daten: Jeder Besucher gibt seine Zahlungsbereitschaft für konventionell erzeugte Tapete und Öko-Tapete ab. Wir können hier also von gepaarten Daten ausgehen.
2. Unabhängigkeit der Differenzen: Wir können davon ausgehen, dass sich die zufällig ausgesuchten Besucher nicht beeinflusst haben.
3. Normalverteilungsannahme: Zu prüfen ist, ob die Differenzen in der Stichprobe normalverteilt sind, entweder über ein Histogramm oder – besser – ein qq-plot (vgl. unten). Diese Annahme kann hier als erfüllt angesehen werden.

Mit einem Histogramm bzw. qq-plot ist zu zeigen, dass Annahme 3 als erfüllt angesehen werden kann.




**Lösung b)**

Wir bilden zunächst die paarweisen Differenzen $d_i = x_i - y_i$ mit $X$ als Zahlungsbereitschaft für kein Öko und $Y$ als Zahlungsbereitschaft für Öko. Dann formulieren wir Null- und Alternativhypothese und wählen die Testvariable.

Die Nullhypothese lautet
$H_0: \mu_D = 0$
bei einem Signifikanzniveau von $\alpha = 0.05$. Damit ist $\mu_D$ die wahre mittlere Differenz zwischen der Zahlungsbereitschaft für kein Öko ($X$) und für Öko ($Y$).
Die Alternativhypothese ist zweiseitig
$H_0: \mu_D \neq 0$
Die Testvariable für diesen Test ist
$T = \frac{\bar{D}-0}{SE(\bar{D})}$ mit $T \sim T_{df}$
wobei der Standardfehler ($SE$) berechnet wird mit
$SE(\bar{D}) = \frac{\hat{\sigma}_D}{\sqrt{n}}$
Hier hierbei ist $\hat{\sigma}_D$ die geschätzte Standardabweichung der paarweisen Differenzen.
Die paarweisen Differenzen $d_i$ sind

| *Besucher* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *kein Öko* | *36* | *18* | *16* | *83* | *49* | *7* | *27* | *89* | *45* | *30* | *40* | *29* | *32* | *15* | *29* | *38* |
| *Öko* | *37* | *21* | *17* | *85* | *48* | *8* | *29* | *92* | *49* | *32* | *41* | *30* | *32* | *13* | *29* | *38* |
| $d_i$ | -1 | -3 | -1 | -2 | 1 | -1 | -2 | -3 | -4 | -2 | -1 | -1 | 0 | 2 | 0 | 0 |

Wir berechnen dem Mittelwert der Differenzen mit $d = -1.125$ und die geschätzte Standardabweichung der Differenzen in der Grundgesamtheit mit $\hat{\sigma}_D = \sqrt{\frac{1}{n-1}(d_i - \bar{d})^2} = 1.5438$. Der Standardfehler ist $SE(\bar{D}) = \frac{1.5438}{\sqrt{16}} = 0.3860$. Der empirische Wert der Testvariable ist somit $t = \frac{-1.125}{0.3860} = -2.915$.

Die Zahl der Freiheitsgrade ist $df = n - 1 = 15$. Der kritische Wert der $t$-Verteilung bei $df = 15$ ist $t_{[0.975]df=15} = 2.131$. Wir können daher die Nullhypothese verwerfen, da $|t| > t_{[0.975]df=15}$. Der exakte $P$-value ist $2\left(1 - F_{t,df=15}(|t|)\right) = 0.0107$ (vgl. R Code unten).
Wir können also die Nullhypothese, nach der sich die mittlere Zahlungsbereitschaft für konventionelle Tapete und Öko-Tapete nicht unterscheidet, verwerfen. Wenn es tatsächlich keinen Unterschied in den mittleren Zahlungsbereitschaften gäben würde, liegt die Wahrscheinlichkeit, bei einer solchen Zufallsstichprobe eine mittlere Differenz der Zahlungsbereitschaften von 1.125 oder größer zu erhalten, bei ca. 1.07%. Auf Grund dieser geringen Wahrscheinlichkeit, verwerfen wir die Nullhypothese und nehmen die Alternativhypothese an, nach der sich die mittlere Zahlungsbereitschaften unterscheiden.

**Lösung c)**
Die Formel zur Berechnung des 95%-Konfidenzintervalls ist
$KI(0.95)_{\mu_D} = \bar{d} \pm t_{[0.975]df=n-1} SE(\bar{D})$
Hier erhalten wir
$KI(0.95)_{\mu_D} = -1.125 \pm 2.131 * 0.3860 = [-1.95, -0.30]$
Wir können also zu 95% darauf vertrauen, dass die wahre mittlere Differenz der Zahlungsbereitschaften in EUR für konventionelle und ökologische Tapete im Intervall $[-1.95, -0.30]$ liegt.

**Lösung mit R**
```
data <- read.csv("paired_t_test.csv")
attach(data)
```

```
head(data)
d <- kein_Öko - Öko; d
# Histogramm und qq-Plot
hist(d)
summary(d)
qqnorm(d, cex = 2)
qqline(d)

sd(d)
n <- length(d); n
SE <- sd(d)/sqrt(n); SE
t <- mean(d)/SE; t
qt(0.975, df = 15)

t.test(d, m0 = 0)
# oder
t.test(kein_Öko, Öko, paired = TRUE)
```

***Aufgabe 11.3: Lehrerfolg***

*Ein Dozent unterrichtet zwei Gruppen (A und B) von Studierenden zum gleichen Thema und möchte untersuchen, ob durch den Einsatz einer neuen Lehrmethode sich der Lehrerfolg verbessert. Gruppe A besucht wie üblich eine Vorlesung und arbeitet parallel mit einem Lehrbuch. Gruppe B erhält zusätzlich die Möglichkeit, Multiple-Choice-Aufgaben mit Lösungshinweisen online zu bearbeiten. Beide Gruppen schreiben die gleiche Klausur. Die folgende Tabelle zeigt die erreichten Punkte in der Klausur.*

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *A* | *42* | *60* | *45* | *49* | *39* | *50* | *36* | *47* | *55* | *43* | *46* | *43* | |
| *B* | *48* | *58* | *59* | *40* | *54* | *35* | *38* | *37* | *58* | *44* | *41* | *89* | *77* |

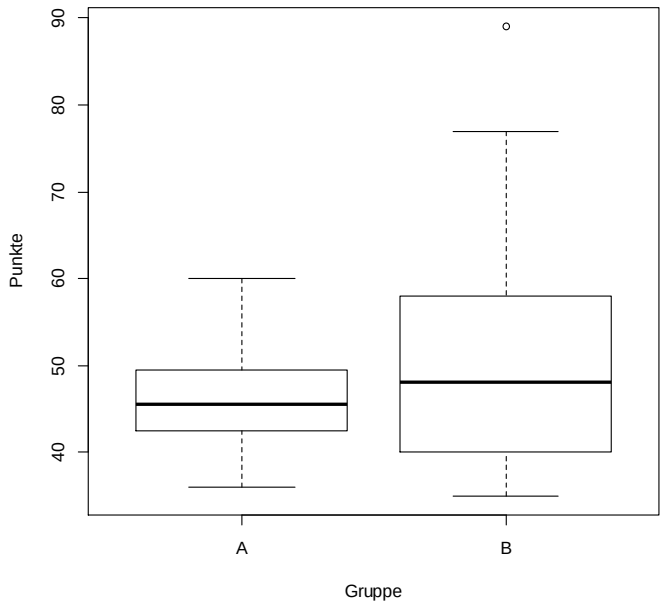
*a) Prüfen Sie die Annahmen zur Durchführung eines MWU-Tests.*
*b) Führen Sie einen MWU-Test durch. Das Signifikanzniveau sei 5%. Interpretation.*

**Lösung a)**

Folgende Annahmen sind für den MWU-Test zu prüfen:

1. Identisch verteilte Teilstichproben. Um diese Annahme zu überprüfen, sollte ein Boxplot mit R dargestellt werden. Der Boxplot unten spricht für diese Annahme.
2. Unabhängigkeit der Teilstichproben: Wir können davon ausgehen, dass sich die Studierenden in den Gruppen nicht beeinflusst haben.

Die Normalverteilungsannahme (wie beim t-Test) ist nicht nötig.



**Lösung b)**
Auch hier ist zunächst die Null- bzw. Alternativhypothese aufzustellen.
Die Nullhypothese lautet
$H_0: x_{[0.5]} = y_{[0.5]}$
und die (zweiseitige) Alternativhypothese ist hier
$H_A: x_{[0.5]} \neq y_{[0.5]}$
Die Nullhypothese besagt also, dass beide Verteilungen den gleichen Median haben. Hier ist $X = A$ und $Y = B$. $A$ hat $m = 12$ Beobachtungen und $B$ hat $n = 13$ Beobachtungen. Wir fassen zunächst beide Gruppen zusammen, ordnen die Beobachtungen und weisen jeder Beobachtung einen Rang zu. Bei gleichen Rängen wird der mittlere Rang der zu vergebenen Ränge zugewiesen.
Wir erhalten folgende Tabelle:

| | B | A | B | B | A | B | B | A | A | A | B | A | A | A | B | A | A | B | A | B | B | A | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| data | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43.0 | 43.0 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 54 | 55 | 58.0 | 58.0 | 59 | 60 | 77 | 89 |
| r | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9.5 | 9.5 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20.5 | 20.5 | 22 | 23 | 24 | 25 |

Wir berechnen die Rangsumme für beide Gruppen und erhalten hier $R_A = \sum Ränge_A = 148$ und $R_B = \sum Ränge_B = 177$. Dann berechnen wir
$U_A = mn + \frac{m(m+1)}{2} - R_A = 86$
und
$U_B = mn + \frac{n(n+1)}{2} - R_B = 70$
Wir prüfen die Bedingung $mn = U_X + U_Y$, mit $12 * 13 = 156 = 86 + 70$. Wir berechnen dann die Prüfgröße $U$ mit $U = \min\{U_A, U_B\} = 70$. Diese Prüfgröße vergleichen wir mit dem kritischen Wert der MWU-Teststatistik (vgl. Lehrbuch). Der kritische Wert ist $U_{krit}(m = 12, n = 13) = 41$. Man verwirft die Null, wenn der Wert von $U$ kleiner oder gleich $U_{krit}$ ist. Dies ist hier nicht der Fall, wir können also $H_0$ nicht verwerfen. Es gibt also keinen statistisch signifikanten Unterschied der Mediane beider Gruppen. Der p-value ist (vgl. Lösung mit R unten) 0.6832 und damit deutlich größer als 5%.

**Lösung mit R**

```
A <- c(42, 60, 45, 49, 39, 50, 36, 47, 55, 43, 46, 43)
B <- c(48, 58, 59, 40, 54, 35, 38, 37, 58, 44, 41, 89, 77)
m <- length(A)
n <- length(B)

summary(A); summary(B)
# Boxplot
boxplot(A, B, names = c("A", "B"),
ylab = "Punkte", xlab = "Gruppe")

data <- c(A, B) # Kombiniere beide Vektoren
names(data) <- c(rep("A", m), rep("B", n))
data <- sort(data) # Sortieren der Daten von klein zu groß
r <- rank(data) # Ermittlung der Ränge
rbind(data, r) # Kombiniere Daten und Ränge

U_A <- m*n + m*(m+1)/2 - sum(r[names(data) == "A"]); U_A
U_B <- m*n + n*(n+1)/2 - sum(r[names(data) == "B"]); U_B
# check: 12*13=156=86+70=156

U <- min(U_A, U_B); U # Ausgabe der Teststatistik U
# Syntax für den MWU test ist Wilcox.test()
wilcox.test(A, B)
```

***Aufgabe 11.4: Powernapping***

*Ein Powernapping (ein kurzer 15-Minuten-Schlaf) soll angeblich die geistige Leistungsfähigkeit steigern. 16 zufällig ausgewählte Mitarbeiter einer Firma lösen einen Mathetest vor einem solchen Schlaf („vorher"). Nach dem Schlaf lösen die Teilnehmer einen zweiten Mathetest mit anderen aber ähnlich schwierigen Fragen. Die Tabelle zeigt die erreichten Punkte (vgl. auch* `Wilcoxon_matched_pairs.csv`*) in den Tests.*

| *Mitarbeiter* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *vorher* | *30* | *26* | *14* | *38* | *40* | *8* | *26* | *24* | *38* | *22* | *59* | *51* | *32* | *11* | *49* | *86* |
| *nachher* | *37* | *21* | *17* | *85* | *48* | *8* | *29* | *92* | *49* | *32* | *41* | *30* | *32* | *13* | *29* | *38* |

*a) Prüfen Sie die Annahmen zur Durchführung eines Wilcoxon-Tests für gepaarte Stichproben.*

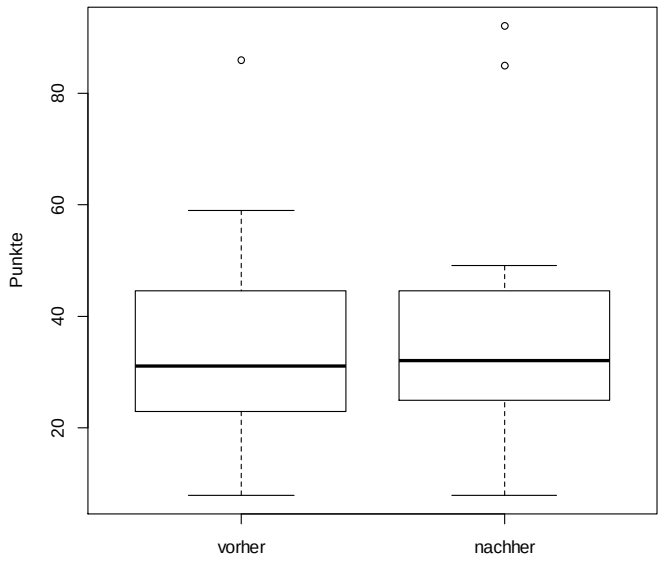
*b) Führen Sie einen Wilcoxon-Test für gepaarte Stichproben durch. Das Signifikanzniveau sei 5%. Interpretation.*

**Lösung a)**

Folgende Annahmen sind für den Wilcoxon-Test für gepaarte Stichproben zu prüfen:

1. Intervallskalierte Daten: Erfüllt, da Angaben mit Punkten (hier als Einheiten interpretiert) vorliegen.
2. Unabhängigkeit der paarweisen Beobachtung: Erfüllt, da eine Zufallsstichprobe vorliegt.

Die Normalverteilungsannahme der Differenzen (wie beim Zweistichproben-t-Test) ist nicht nötig. Auch hier ist zunächst die Null- bzw. Alternativhypothese festzulegen.
Die Nullhypothese lautet
$H_0$: $\Delta_{[0.5]} = 0$ mit $\Delta_i = x_i - y_i$
und die (zweiseitige) Alternativhypothese ist hier
$H_A$: $\Delta_{[0.5]} \neq 0$
Die Nullhypothese besagt also, dass der Median der paarweisen Differenzen Null ist. Hier ist $X = vorher$ und $Y = nachher$. $X$ und $Y$ haben jeweils $n = 16$ Beobachtungen.
Der Boxplot der Beobachtungen sieht unauffällig aus. Der Median von „nachher" ist nur geringfügig größer als der von „vorher".




Wir bilden zunächst die Differenzen der Beobachtungen (`d = vorher - nachher`).
Dann berechnen wir den Betrag der Differenzen (`d_abs`) und den zugehörigen Rang (`rang`).

Ränge bei Gleichheit, wenn also d = 0 ist, werden nicht mitgezählt (hier gibt es zwei Paare mit d = 0). Wir berechnen also den Rang ohne Gleichheit (`rang_ohne0`). Wir berechnen dann daraus die Summe der Rangwerte $\sum R_+$, die zu positiven Differenzen gehören, und die Summe der Rangwerte $\sum R_-$, die zu negativen Differenzen gehören. Es ergibt sich $\sum R_+ = 47$ und $\sum R_- = 58$.
Die Berechnung ergibt sich über folgende Tabelle.

```
   vorher nachher   d d_abs rang rang_ohne0
1      30      37  -7     7  7.0        5.0
2      26      21   5     5  6.0        4.0
3      14      17  -3     3  4.5        2.5
4      38      85 -47    47 14.0       12.0
5      40      48  -8     8  8.0        6.0
6       8       8   0     0  1.5       -0.5
7      26      29  -3     3  4.5        2.5
8      24      92 -68    68 16.0       14.0
9      38      49 -11    11 10.0        8.0
10     22      32 -10    10  9.0        7.0
11     59      41  18    18 11.0        9.0
12     51      30  21    21 13.0       11.0
13     32      32   0     0  1.5       -0.5
14     11      13  -2     2  3.0        1.0
15     49      29  20    20 12.0       10.0
16     86      38  48    48 15.0       13.0
```

Zur Probe berechnen wir (mit der Zahl der Beobachtungen ohne Gleichheit bzw. Rang 0, also mit $n = 14$) $\sum R_+ + \sum R_- = \frac{n(n+1)}{2} \Leftrightarrow 47 + 58 = 105 = \frac{14*15}{2}$.
Wir berechnen dann die Prüfgröße $W$ mit $W = \min\{\sum R_+, \sum R_-\} = 47$. Diese Prüfgröße vergleichen wir mit dem kritischen Wert der Teststatistik für den Wilcoxon-Test für gepaarte Stichproben (vgl. Lehrbuch). Der kritische Wert ist $W_{krit}(n = 14) = 21$. Man verwirft die Nullhypothese, wenn der Wert von $W$ kleiner oder gleich $W_{krit}$ ist. Dies ist hier nicht der Fall, wir können also $H_0$ nicht verwerfen. Es gibt also keinen statistisch signifikanten Unterschied der Differenz der Mediane beider Gruppen. Der p-value ist (vgl. Lösung mit R unten) 0.7536 und damit deutlich größer als 5%.

**Lösung mit R**

```
getwd()
data <- read.csv("Wilcoxon_matched_pairs_test.csv")
head(data)
names(data)
attach(data)
n <- length(vorher); n

# Boxplot: Variable vorher hat den kleineren Median
boxplot(vorher, nachher, ylab = "Punkte",
names = c("vorher", "nachher"))

# Differenzen und Absolutwert der Differenzen
d <- vorher - nachher; d
d_abs <- abs(d); d_abs

# Rang der absoluten Differenzen
rang <- rank(d_abs); rang

# Rang 0 (Gleichheit) wird nicht gezählt
rang_ohne0 <- rank(d_abs) - sum(d_abs == 0); rang_ohne0

tab <- data.frame(vorher, nachher, d, d_abs, rang, rang_ohne0); tab
sum(rang_ohne0[d < 0])
```

```
sum(rang_ohne0[d > 0])

# Probe: Achtung => Ohne Rang 0
n_neu <- n - 2; n_neu
# 47+58 = n_neu*(n_neu+1)/2

wilcox.test(vorher, nachher, paired = TRUE)
```

***Aufgabe 11.5: Würfel***

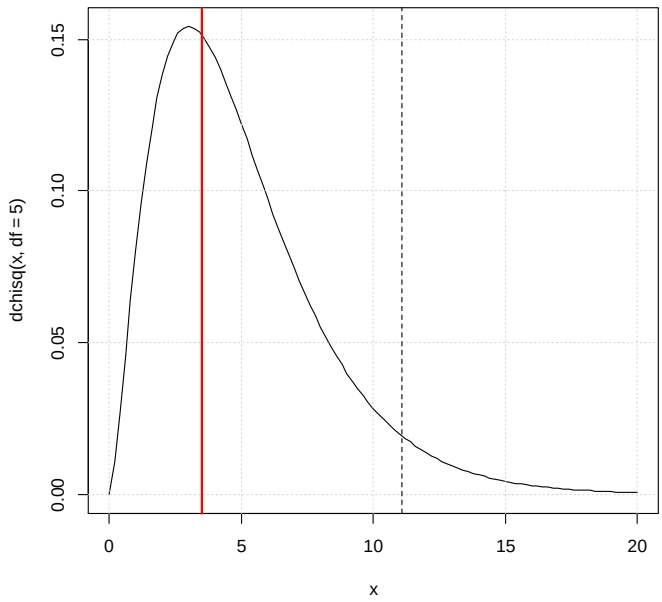
*Es soll untersucht werden, ob ein Würfel „fair" ist, d.h., ob jede der sechs Seiten des Würfels mit der gleichen Wahrscheinlichkeit auftritt. Sie würfeln 200 x und erhalten folgende Häufigkeitsverteilung (vgl. auch `Zahl.csv`).*

| *Zahl* | *1* | *2* | *3* | *4* | *5* | *6* |
|---|---|---|---|---|---|---|
| *Häufigkeit* | *28* | *37* | *37* | *27* | *33* | *38* |

*Untersuchen Sie diese Frage mit einem Anpassungstest. Das Signifikanzniveau sei 5%. Interpretation.*

**Lösung**

Wie bei jedem Test sind zunächst Nullhypothese, Alternativhypothese und Testverteilung zu bestimmen. Beim Anpassungstest will man zeigen, ob die Verteilung eines qualitativen Merkmals einer bestimmten Verteilung folgt oder eben nicht. Hier geht es darum, ob ein Würfel mit gleicher Wahrscheinlichkeit die Zahlen von 1 bis 6 generiert.
Die Nullhypothese ist daher
$H_0: F = F_{Gleichverteilung}$
und die Alternativhypothese ist
$H_A: F \neq F_{Gleichverteilung}$
Die Teststatistik ist
$\sum_{j=1}^{m} \frac{(n_j - E_j)^2}{E_j} \sim \chi^2_{m-1}$
mit $n_j$ als Zahl der Beobachtungen pro Klasse. Die Teststatistik folgt einer $\chi^2$-Verteilung mit $m - 1$ Freiheitsgraden.
Hier liegen $m = 6$ Klassen vor mit den Beobachtungen $\{28, 37, 37, 27, 33, 38\}$. Die erwarteten absoluten Häufigkeiten unter Gültigkeit von $H_0$ sind $E_j = \frac{200}{6} = 33.33$. Wir berechnen also die Prüfgröße mit
$\frac{(28-33.33)^2}{33.33} + \frac{(37-33.33)^2}{33.33} + \frac{(37-33.33)^2}{33.33} + \frac{(27-33.33)^2}{33.33} + \frac{(33-33.33)^2}{33.33} + \frac{(38-33.33)^2}{33.33} = 3.52$
Die Dichte der $\chi^2$-Verteilung mit 5 Freiheitsgraden ist hier dargestellt. Die rote vertikale Linie zeigt den Wert der Prüfgröße (=3.52) an, die gestrichelte vertikale Linie den kritischen Wert, hier das 95%-Quantil bei $df = 5$, welches bei 11.07 liegt.




Die Nullhypothese wird verworfen, wenn die Prüfgröße größer ist als der kritische Wert. Dies ist hier nicht der Fall. Wir können also die Nullhypothese, nach der der Würfel „fair“ ist, also die Zahlen von 1 bis 6 mit gleicher Wahrscheinlichkeit generiert, nicht verwerfen.
Die p-value berechnet sich über die Prüfgröße und unter Nutzung der $\chi^2$-Verteilung mit 5 Freiheitsgraden. Hier ist:

$$\text{p-value} = 1 - F_{\chi^2}\left(\sum_{j=1}^{m} \frac{(n_j - E_j)^2}{E_j}, df = 5\right) = 1 - F_{\chi^2}(3.52, df = 5) = 0.6204$$

Der p-value liegt über 5%, also kann die Nullhypothese nicht verworfen werden.

**Lösung mit R**

```
data <- read.csv("Zahl.csv")
head(data)
attach(data)
names(data)

# absolute Häufigkeitsverteilung
table(Zahl)

# Berechnung der Prüfgröße
nj <- c(28, 37, 37, 27, 33, 38)
Ej <- c(33.33, 33.33, 33.33, 33.33, 33.33, 33.33)
chi2 <- sum(((nj-Ej)^2)/Ej); chi2

# Plot der Dichte
curve(dchisq(x, df = 5), xlim = c(0, 20))
grid()
abline(v = 3.52, lwd = 2, col = "red")
abline(v = qchisq(0.95, df = 5), lty = 2)

# Test
chisq.test(table(Zahl))
# p-value
1 - pchisq(3.52, df = 5)
```

***Aufgabe 11.6: Journalpräferenz und Geschlecht***

*Testen Sie mit Hilfe eines Unabhängigkeitstests, ob die Merkmale „Journal“ und „Geschlecht“ in Bsp. 2.1 (vgl. `Bsp._2.1.csv`) unabhängig voneinander sind. Das Signifikanzniveau sei 5%. Interpretation der Testergebnisse.*

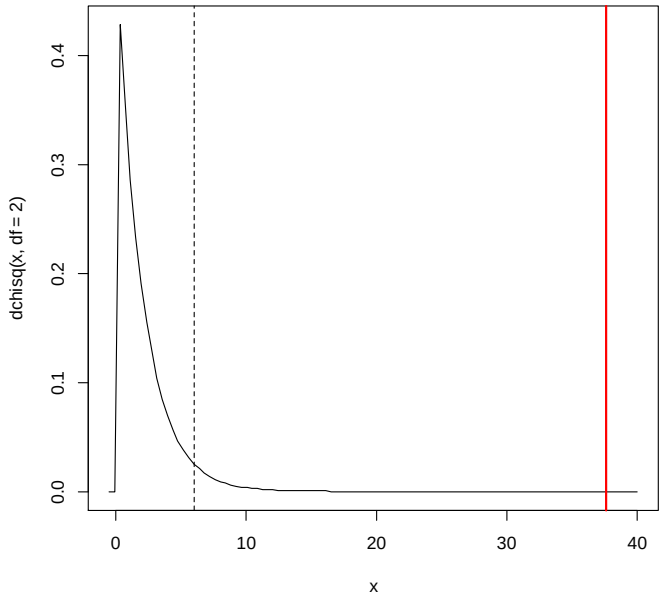
**Lösung**
Beim Unabhängigkeitstest will man zeigen, ob zwei qualitative Merkmale ($X$ und $Y$) voneinander abhängig sind oder eben nicht.
Die Nullhypothese ist daher
$H_0: X\ und\ Y\ sind\ unabhängig$
und die Alternativhypothese ist
$H_A: X\ und\ Y\ sind\ nicht\ unabhängig \Leftrightarrow X\ und\ Y\ sind\ abhängig$
Die Teststatistik ist
$\sum\sum \frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(k-1)(l-1)}$
mit $n_{ij}$ als Zahl der Beobachtungen für Zeile $i$ und Spalte $j$ und $E_{ij}$ als Werte bei Unabhängigkeit. Die Teststatistik folgt einer $\chi^2$-Verteilung mit $(k-1)(l-1)$ Freiheitsgraden, wobei $k$ die Zeilenzahl und $l$ die Spaltenzahl der Kontingenztabelle ist. Wenn der Wert der Teststatistik (also die Prüfgröße) überraschend groß ist, spricht dies gegen die Nullhypothese. In diesem Fall sind die Abweichungen zwischen $n_{ij}$ und $E_{ij}$ relativ groß. Wir verwerfen dann die Hypothese der Unabhängigkeit beider Merkmale.
Zunächst erstellen wir eine Kontingenztabelle für die gemeinsame Verteilung der Merkmale Journal und Geschlecht. Es ergibt sich

| | Journal | | | |
|---|---|---|---|---|
| Geschlecht | Cosmopolitan | Economist | Sports Illustrated | Sum |
| Frau | 45 | 5 | 10 | 60 |
| Mann | 5 | 10 | 25 | 40 |
| Sum | 50 | 15 | 35 | 100 |

Die Kontingenztabelle liefert die Werte für $n_{ij}$, also die absoluten Häufigkeiten für Zeile $i$ und Spalte $j$. Hier ist z.B. $n_{22} = 10$, d.h., 10 Teilnehmer sind Männer und präferieren Economist. Die Gesamtzahl an Beobachtungen ist $n = 100$.
Die erwarteten Werte bei Unabhängigkeit, $E_{ij}$, ergeben sich mit
$E_{ij} = \frac{Zeilensumme_i * Spaltensumme_j}{n}$
Wir erhalten für $E_{ij}$ folgende Werte

| | Journal | | |
|---|---|---|---|
| Geschlecht | Cosmopolitan | Economist | Sports Illustrated |
| Frau | 30 | 9 | 21 |
| Mann | 20 | 6 | 14 |

Für die Prüfgröße erhalten wir
$\sum\sum \frac{(n_{ij}-E_{ij})^2}{E_{ij}} = 37.6$
Die Dichte der $\chi^2$-Verteilung mit 2 Freiheitsgraden (= 1*2) ist hier dargestellt. Die rote vertikale Linie zeigt den Wert der Prüfgröße (=37.6) an, die gestrichelte vertikale Linie den kritischen Wert, hier das 95%-Quantil bei $df = 2$, welches bei 5.991 liegt.




Die Nullhypothese wird verworfen, wenn die Prüfgröße größer ist als der kritische Wert. Dies ist hier der Fall. Wir können also die Nullhypothese, nach Geschlecht und Journalpräferenz unabhängig voneinander sind, verwerfen. Wir können also davon ausgehen, dass Frauen eine andere Zeitschriftenpräferenz als Männer haben.
Die p-value berechnet sich über die Prüfgröße und unter Nutzung der $\chi^2$-Verteilung mit 2 Freiheitsgraden. Hier ist:

p-value = $1 - F_{\chi^2}\left(\sum\sum \frac{(n_{ij}-E_{ij})^2}{E_{ij}}, df = 2\right) = 1 - F_{\chi^2}(37.6, df = 2) < 0.001$

Der p-value liegt unter 5%, also kann die Nullhypothese verworfen werden.

**Lösung mit R**

```
data <- read.csv("Bsp._2.1.csv")
head(data)
attach(data)
addmargins(table(Geschlecht, Journal))

chi <- chisq.test(table(Geschlecht, Journal))
chi$expected
chi

curve(dchisq(x, df = 2), xlim = c(-0.5, 40))
abline(v = qchisq(0.95, df = 2), lty = 2)
abline(v = 37.6, lwd = 2, col = "red")

# p-value
1 - pchisq(37.599, df = 2)
```

***Aufgabe 11.7: Vegetarische Ernährung***

*Bei einer Befragung zum Essverhalten werden 145 Frauen und 130 Männer u.a. danach gefragt, ob sie sich überwiegend vegetarisch ernähren. Es ergibt sich die folgende Häufigkeitsverteilung der Angaben (vgl.* `Vegetarier.csv`*).*

| $n_{ij}$ | *Präferenz* | | |
|---|---|---|---|
| | *Kein Vegetarier* | *Vegetarier* | $\sum$ |
| *Frau* | *100* | *45* | *145* |
| *Mann* | *105* | *25* | *130* |
| $\sum$ | *205* | *70* | *275* |

*Mit einem Homogenitätstest soll untersucht werden, ob sich die Präferenz für vegetarische Ernährung zwischen Frauen und Männern unterscheidet. Das Signifikanzniveau sei 5%.*

**Lösung**

Die Frage ist hier, ob sich die Verteilung eines Merkmals (Präferenz für vegetarische Ernährung) zwischen zwei Grundgesamtheiten (Frauen und Männer) unterscheidet. Wie ist dementsprechend die Nullhypothese zu formulieren?

Gemäß der Nullhypothese haben beide Grundgesamtheiten die gleiche Präferenz:

$H_0$: $p_{kein\,Veg,Frau} = p_{kein\,Veg,Mann}$ und $p_{Veg,Frau} = p_{Veg,Mann}$

Die Alternativhypothese $H_A$ ist, dass mindestens eine der Hypothesen in $H_0$ falsch ist.

Wir überlegen nun, welche absoluten Beobachtungen unter Gültigkeit von $H_0$ zu erwarten sind. Von 275 Teilnehmern sind 70 (25.45%) Vegetarier und 205 (74.55%) nicht. Diese Relation soll sich zwischen Männern und Frauen nicht unterscheiden – wenn $H_0$ wahr ist. Wir können also erwarten, dass von 145 Frauen 37 (exakt: 36.91) Vegetarier sein sollten und 108 (108.09) nicht. Entsprechend sollten von den 130 Männern 33 (33.09) Vegetarier sein und 97 (96.91) nicht.

Es ergibt sich folgende Tabelle für die Werte bei Gültigkeit von $H_0$, $E_{ij}$:

```
          Präferenz
Geschlecht  kein Veg      Veg Sum
      Frau 108.09091 36.90909 145
      Mann  96.90909 33.09091 130
      Sum  205.00000 70.00000 275
```

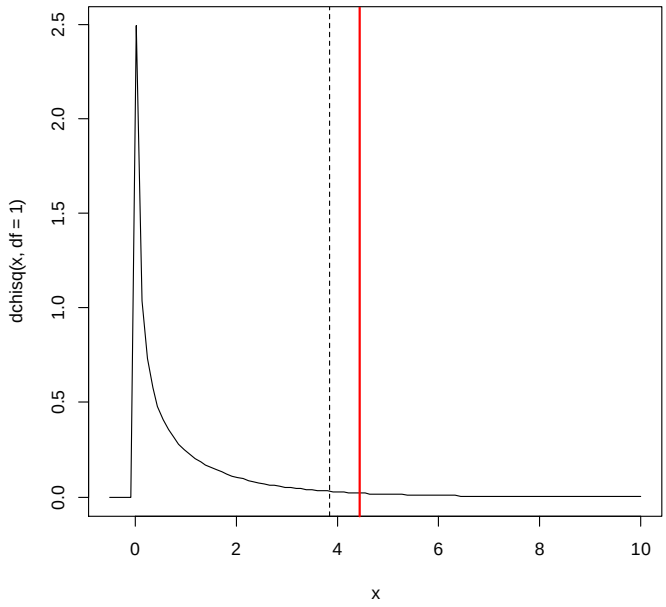
Die erwarteten Werte bei Gültigkeit von $H_0$, $E_{ij}$, ergeben sich analog zum Test auf Unabhängigkeit mit der Formel:

$$E_{ij} = \frac{Zeilensumme_i * Spaltensumme_j}{n}$$

Der Test ist ansonsten analog zum Test auf Unabhängigkeit. Für die Prüfgröße erhalten wir

$$\sum\sum \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = 4.43$$

Die Zahl der Freiheitsgrade ist hier 1, da es sich um eine 2x2 Kontingenztabelle handelt. Die Dichte der $\chi^2$-Verteilung mit 1 Freiheitsgrad (= 1*1) ist hier dargestellt. Die rote vertikale Linie zeigt den Wert der Prüfgröße (=4.43) an, die gestrichelte vertikale Linie den kritischen Wert, hier das 95%-Quantil bei $df = 1$, welches bei 3.841 (vgl. Lösung mit R oder Tab. 14.4) liegt.




Die Nullhypothese wird verworfen, wenn die Prüfgröße größer ist als der kritische Wert (der Test ist immer oberseitig). Dies ist hier der Fall. Wir können also die Nullhypothese, nach der Männer und Frauen die gleiche Präferenz für vegetarische Ernährung haben, verwerfen. Wir können hier also davon ausgehen, dass Frauen eine stärkere Präferenz für vegetarische Ernährung haben als Männer.
Die p-value berechnet sich über die Prüfgröße und unter Nutzung der $\chi^2$-Verteilung mit 1 Freiheitsgrad. Hier ist:

p-value = $1 - F_{\chi^2}\left(\sum\sum\frac{(n_{ij}-E_{ij})^2}{E_{ij}}, df = 1\right) = 1 - F_{\chi^2}(4.43, df = 1) = 0.035$

Der p-value liegt unter 5%, also kann die Nullhypothese verworfen werden.

**Lösung mit R**

```
getwd()
data <- read.csv("Vegetarier.csv")
head(data)

addmargins(table(data))
chi <- chisq.test(table(data))
addmargins(chi$expected)
chi

curve(dchisq(x, df = 1), xlim = c(-0.5, 10))
abline(v = qchisq(0.95, df = 1), lty = 2)
abline(v = 4.43, lwd = 2, col = "red")

# p-value
1 - pchisq(4.43, df = 1)
```

## Kapitel 12: Hypothesentests für lineare Regression und Korrelation

***Aufgabe 12.1: Autos***

*Der Datensatz* `Autos.csv` *enthält Angaben zu Modell, Leistung (in PS), Verbrauch (kombiniert, in l/100km) und $CO_2$-Emissionen (in $gCO_2$/km) von 30 Neuwagen im Jahr 2016.*

*a) Liegt eine statistisch signifikante Korrelation zwischen Leistung und Verbrauch vor? Führen Sie einen Hypothesentest für die Korrelation (das Signifikanzniveau sei 5%) durch. Interpretation.*

*b) Berechnen Sie das 95%-Konfidenzintervall für die wahre Korrelation zwischen Leistung und Verbrauch. Interpretation.*
*c) Führen Sie die Schritte a) und b) für Leistung und $CO_2$-Emissionen durch.*

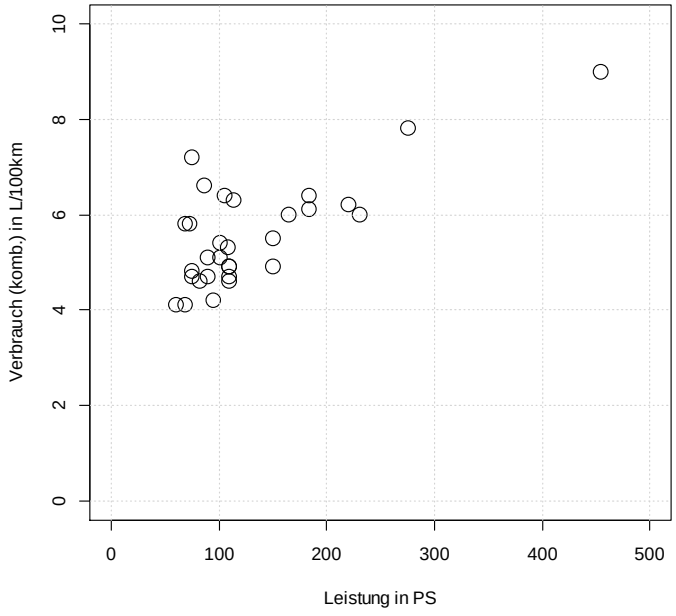
**Lösung a)**
Sowohl Leistung als auch Verbrauch sind quantitative Merkmale. Als Korrelationsmaß ist daher der Pearson-Korrelationskoeffizient $r_{XY}$ (hier als $r$ bezeichnet) zu verwenden. Wir gehen von einer Zufallsstichprobe aus. Die Nullhypothese ist dann:
$H_0: \rho = 0$ vs. $H_A: \rho \neq 0$
mit $\rho$ als wahre Korrelation in der Grundgesamtheit. Der Korrelationskoeffizient $r$ in der Stichprobe ist Realisation einer Zufallsvariablen $R$. Unter Gültigkeit von $H_0$ gilt

$$R\sqrt{\frac{n-2}{1-R^2}} \sim T_{df=n-2}$$

Die Realisation der Testvariablen, $t = r\sqrt{\frac{n-2}{1-r^2}}$, kann dann benutzt werden, um einen p-value zu berechnen bzw. über den Vergleich mit dem kritischen Wert der t-Verteilung zu einer Entscheidung über $H_0$ zu gelangen.
Wir berechnen hier zunächst den Korrelationskoeffizienten (mit $X$ als Leistung und $Y$ als Verbrauch) mit

$$r_{XY} = r = \frac{\hat{c}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{66.87126}{81.13976 * 1.119709} = 0.736039$$

Mit $r = 0.736$ ist die Korrelation positiv und relativ stark. Das Streudiagramm zeigt einen relativ starken, positiven linearen Zusammenhang. Höhere Leistung geht einher mit höherem Verbrauch und umgekehrt.




Der Wert der Testvariablen ist

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.736039\sqrt{\frac{30-2}{1-(0.736039)^2}} = 5.753475$$

Der kritische Wert der t-Verteilung bei $df = n - 2 = 28$ ist hier $t_{[0.975]df=28} = 2.048$. Da $t > t_{[0.975]df=28}$ ist, können wir die Nullhypothese, nach der es in der Grundgesamtheit keine Korrelation gibt, verwerfen. Der p-value ist $2(1 - F_t(t, df = 28)) \approx 0$. Die Korrelation zwischen Leistung und Verbrauch ist also statistisch signifikant.

**Lösung b)**
Das 95%-Konfidenzintervall für die wahre Korrelation zwischen Leistung und Verbrauch wird mit Fisher's $r$-to-$z$ Transformation bestimmt. Wir berechnen zunächst das 95%-KI für $z$. Hierzu berechnen wir

$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = 0.94178$

und die Standardabweichung

$s(z) = \sqrt{\frac{1}{n-3}} = 0.1924501$

Mit diesen Werten lässt sich das 95%-KI für den wahren z-Wert in der Grundgesamtheit bestimmen.

$KI(0.95)_{z_{GG}} = z \pm 1.96 * s(z) = [0.5645848, 1.318975]$

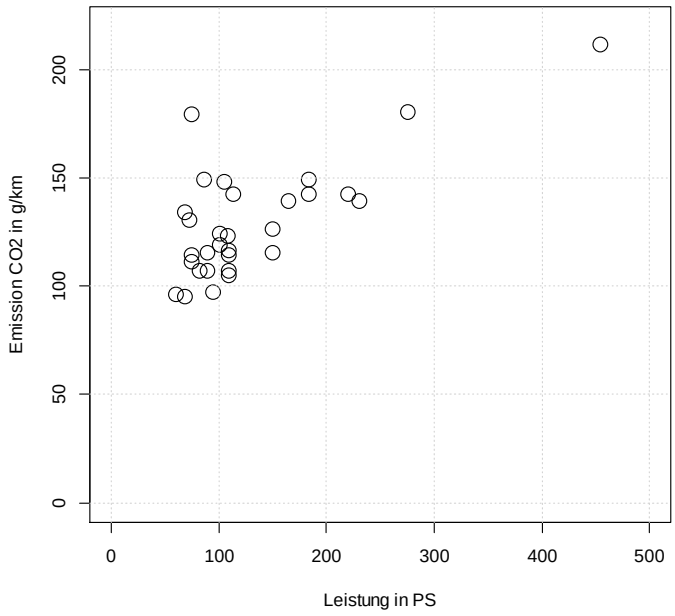
Die untere und obere Grenze des Intervalls können dann wiederum transformiert werden, um die untere und obere Grenze des KI für $\rho$ zu erhalten. Die Transformation ergibt sich mit

$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \Leftrightarrow r = \frac{e^{2z}-1}{e^{2z}+1}$

Wir erhalten dann

$KI(0.95)_{\rho} = [0.5113712, 0.8665289]$

**Lösung c)**
Betrachten wir nun die Korrelation zwischen Leistung und $CO_2$-Emission. Das Streudiagramm zeigt wiederum einen relativ starken, positiven linearen Zusammenhang.



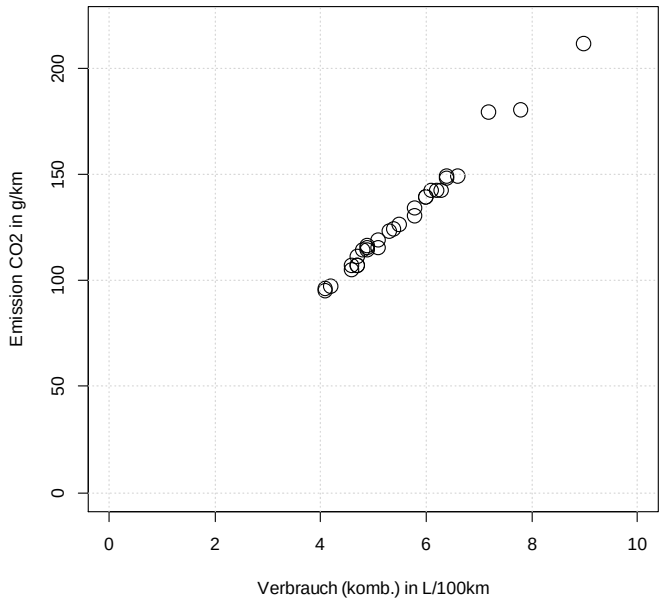

Der Korrelationskoeffizienten (mit $X$ als Leistung und $Y$ als $CO_2$-Emission) ist

$r = 0.7231656$

Wir erhalten für die Prüfgröße $t = 5.5404$. Der p-value ist wiederum nahe Null. Das 95%-KI ist [0.4907548 0.8595044]. Die Berechnung ist analog zu a) und b).

Die Ursache dafür, dass die Korrelation zwischen Leistung und $CO_2$-Emission einerseits und Leistung und Verbrauch andererseits praktisch identisch ist, liegt darin, dass zwischen Verbrauch und $CO_2$-Emission ein fixes Verhältnis existiert. Bei der Verbrennung von 1 Liter Benzin entstehen immer 2333 g/$CO_2$ (unabhängig von der gewählten Technologie) Dies zeigt auch das Streudiagramm zwischen Verbrauch und $CO_2$-Emission. Zum Beispiel ergeben sich

aus einem Verbrauch von 10L/100km $CO_2$-Emissionen in Höhe von 10*2333g$CO_2$/100km = 233g$CO_2$/km.
Kennt man also den Verbrauch eines Pkw, ist auch die $CO_2$-Emission bekannt und umgekehrt. Das Streudiagramm unten zeigt die fast perfekte Korrelation zwischen Verbrauch und $CO_2$-Emission.




## Lösung mit R

```
getwd()
data <- read.csv("Autos.csv")
head(data)
attach(data)

summary(Leistung)
summary(Verbrauch)
summary(CO2)

plot(Leistung, Verbrauch,
xlim = c(0, 500), ylim = c(0, 10),
xlab = "Leistung in PS", ylab = "Verbrauch (komb.) in L/100km",
cex = 2)
grid()

# -----------------------------------
# a)
sd(Leistung)
sd(Verbrauch)
cov(Leistung, Verbrauch)
cov(Leistung, Verbrauch)/(sd(Leistung)*sd(Verbrauch))

r <- cov(Leistung, Verbrauch)/(sd(Leistung)*sd(Verbrauch)); r

n <- length(Leistung); n
t <- r*sqrt((n-2)/(1-r^2)); t
qt(0.975, df = n-2)

# p-value
2*(1 - pt(t, df = n-2))

cor.test(Leistung, Verbrauch)

plot(Leistung, CO2,
xlim = c(0, 500), ylim = c(0, 220),
xlab = "Leistung in PS", ylab = "Emission CO2 in g/km",
```

```
cex = 2)
grid()
# ------------------------------------
# b)
z <- 0.5*log((1+r)/(1-r)); z
s <- sqrt(1/(n-3)); s
KI_z_u <- z - qnorm(0.975)*s; KI_z_u
KI_z_o <- z + qnorm(0.975)*s; KI_z_o
e <- exp(1); e
KI_u <- (e^(2*KI_z_u) - 1)/(e^(2*KI_z_u) + 1); KI_u
KI_o <- (e^(2*KI_z_o) - 1)/(e^(2*KI_z_o) + 1); KI_o
# ------------------------------------
# c)
plot(Verbrauch, CO2,
xlim = c(0, 10), ylim = c(0, 220),
xlab = "Verbrauch (komb.) in L/100km", ylab = "Emission CO2 in g/km",
cex = 2)
grid()

cor.test(Leistung, CO2)
cor.test(Verbrauch, CO2)
```

***Aufgabe 12.2: Zeugnisnoten***

*Der Datensatz* `Noten.csv` *enthält Zeugnisnoten von 20 Schülern für die Fächer Musik und Mathematik.*

*a) Schätzen Sie die Korrelation zwischen beiden Merkmalen. Interpretation.*

*b) Liegt eine statistisch signifikante Korrelation vor? Führen Sie einen Hypothesentest für die Korrelation (das Signifikanzniveau sei 5%) durch. Interpretation.*

**Lösung a) und b)**

Noten sind qualitative Merkmale, die ordinal skaliert sind. Als Korrelationsmaß ist daher der Spearman-Korrelationskoeffizient $r^{Sp}$ zu verwenden. Wir gehen von einer Zufallsstichprobe aus. Die Nullhypothese ist dann:

$H_0$: $\rho^{Sp} = 0$ vs. $H_A$: $\rho^{Sp} \neq 0$

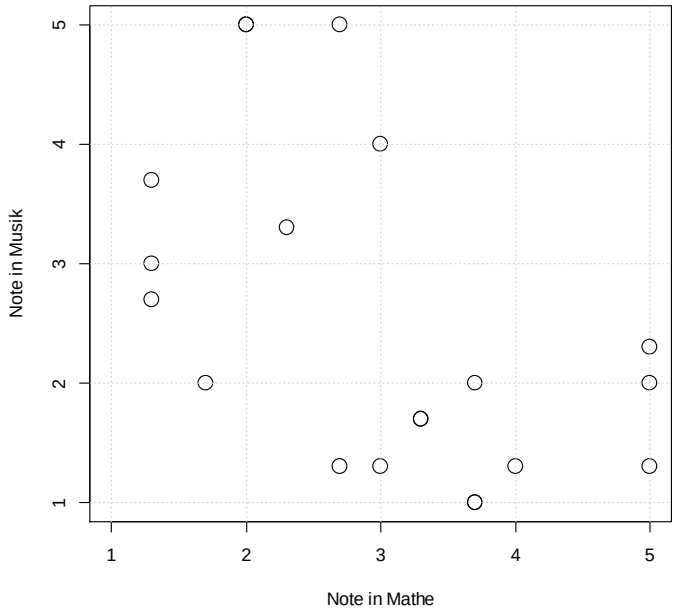
mit $\rho^{Sp}$ als wahre Korrelation in der Grundgesamtheit. Der Korrelationskoeffizient $r^{Sp}$ in der Stichprobe ist Realisation einer Zufallsvariablen $R^{Sp}$. Unter Gültigkeit von $H_0$ gilt

$$R^{Sp}\sqrt{\frac{n-2}{1-\left(R^{Sp}\right)^2}} \sim T_{df=n-2},$$

wenn $n > 10$ ist.

Die Realisation der Testvariablen, $t = r^{Sp}\sqrt{\frac{n-2}{1-\left(r^{Sp}\right)^2}}$, kann benutzt werden, um einen p-value zu berechnen bzw. über den Vergleich mit dem kritischen Wert der t-Verteilung zu einer Entscheidung über $H_0$ zu gelangen.

Betrachten wir zunächst das Streudiagramm der Noten. Dieses legt nahe, dass zwischen den Rängen ein mittelstarker, negativer linearer Zusammenhang existiert.




Wir berechnen hier zunächst den Korrelationskoeffizienten (mit $X$ als Note in Mathe und $Y$ als Note in Musik). Hierzu müssen wir zunächst die Ränge (mit `rg_Ma` und `rg_Mu` als Ränge für Mathe bzw. Musik) bestimmen. Es ergibt sich folgende Tabelle:

| | Mathe | Musik | rg_Ma | rg_Mu |
|---|---|---|---|---|
| 1 | 3.7 | 1.0 | 15.0 | 1.5 |
| 2 | 1.7 | 2.0 | 4.0 | 10.0 |
| 3 | 1.3 | 3.0 | 2.0 | 14.0 |
| 4 | 4.0 | 1.3 | 17.0 | 4.5 |
| 5 | 2.0 | 5.0 | 5.5 | 19.0 |
| 6 | 3.3 | 1.7 | 12.5 | 7.5 |
| 7 | 3.3 | 1.7 | 12.5 | 7.5 |
| 8 | 3.7 | 1.0 | 15.0 | 1.5 |
| 9 | 3.0 | 1.3 | 10.5 | 4.5 |
| 10 | 2.0 | 5.0 | 5.5 | 19.0 |
| 11 | 3.0 | 4.0 | 10.5 | 17.0 |
| 12 | 2.7 | 1.3 | 8.5 | 4.5 |
| 13 | 5.0 | 2.3 | 19.0 | 12.0 |
| 14 | 3.7 | 2.0 | 15.0 | 10.0 |
| 15 | 2.3 | 3.3 | 7.0 | 15.0 |
| 16 | 1.3 | 2.7 | 2.0 | 13.0 |
| 17 | 5.0 | 1.3 | 19.0 | 4.5 |
| 18 | 1.3 | 3.7 | 2.0 | 16.0 |
| 19 | 2.7 | 5.0 | 8.5 | 19.0 |
| 20 | 5.0 | 2.0 | 19.0 | 10.0 |

Aus den Rängen bestimmen wir dann Kovarianz und die Standardabweichungen und berechnen den Spearman-Korrelationskoeffizienten

$$r^{Sp} = \frac{\hat{c}_{rg(X),rg(Y)}}{\hat{\sigma}_{rg(X)}\hat{\sigma}_{rg(Y)}} = \frac{-19.84211}{5.880387 * 5.871429} = -0.5746958$$

Mit $r^{Sp} = -0.575$ ist die Korrelation negativ und mittelstark. Höhere Notenränge für Mathe gehen einher mit niedrigeren Notenrängen in Musik und umgekehrt. Kurz: Wer relativ gut ist in Mathe, ist relativ schlecht in Musik.

Die Berechnung der Prüfgröße ist (für $n > 10$) analog zum Pearson-Korrelationskoeffizienten. Hier ist $n = 20$ und

$$t = r^{Sp}\sqrt{\frac{n-2}{1-\left(r^{Sp}\right)^2}} = -0.5746958\sqrt{\frac{18}{1-(-0.5746958)^2}} = -2.979381$$

Der kritische Wert der t-Verteilung bei $df = n - 2 = 18$ ist hier $t_{[0.975]df=18} = 2.101$. Da $|t| > t_{[0.975]df=18}$ ist, können wir die Nullhypothese, nach der es in der Grundgesamtheit keine Korrelation gibt, verwerfen. Der p-value ist $2\big(1 - F_t(|t|, df = 18)\big) = 0.008$. Die Korrelation zwischen den Notenränken für Mathe und Musik ist also statistisch signifikant.

**Lösung mit R**

```
data <- read.csv("Noten.csv")
head(data)
attach(data)

plot(Mathe, Musik,
xlab = "Note in Mathe", ylab = "Note in Musik",
cex = 2, xlim = c(1,5), ylim = c(1,5))
grid()

# Ränge
rg_Ma <- rank(Mathe)
rg_Mu <- rank(Musik)

# Tabelle
tab <- data.frame(Mathe, Musik, rg_Ma, rg_Mu); tab

cov(rank(Mathe), rank(Musik))
sd(rank(Mathe))
sd(rank(Musik))

r <- cov(rank(Mathe), rank(Musik))/(sd(rank(Mathe))*sd(rank(Musik))); r

# Berechnung t-Wert/Prüfgröße
n <- length(Mathe); n
t <- r*sqrt((n-2)/(1-r^2)); t

# kritischer Wert
qt(0.975, df = n-2)

2*(1-pt(abs(t), df = n-2))

# oder
cor.test(Mathe, Musik, method = "spearman")
```

***Aufgabe 12.3: Bildung und Stundenlohn***

*Betrachten Sie die Angaben in Bsp. 5.4. Schätzen Sie das Modell*

$$\widehat{log\,Lohn} = \hat{\beta}_0 + \hat{\beta}_1 Bildung$$

*Prüfen Sie die Modellannahmen und untersuchen Sie mit einem Hypothesentest, ob der Anstiegsparameter statistisch signifikant von Null verschieden ist. Das Signifikanzniveau sei 5%. Interpretation.*

**Lösung**

Wir wollen das Modell $\widehat{log\,Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ schätzen, mit *X* = Jahre in Bildung und *Y* = Stundenlohn in USD.

Wir schätzen die Parameter $\hat{\beta}_0$ und $\hat{\beta}_1$ mit den Formeln

$\hat{\beta}_1 = \frac{\hat{c}_{XY}}{\hat{\sigma}_X^2} = \frac{2.468972}{10.47619} = 0.2356746$

und

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2.773468 - 0.2356746 * 13 = -0.2903015$

Damit ist das geschätzte Modell

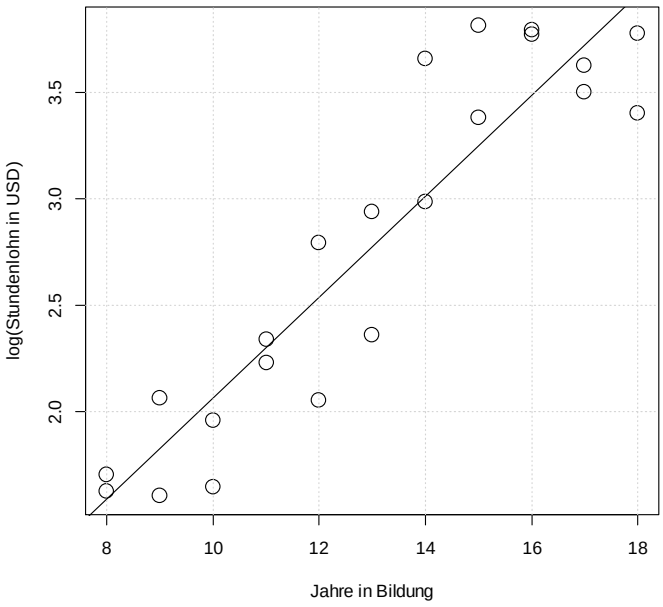
$\widehat{log\,Lohn} = -0.2903 + 0.2357 Bildung$

bzw.

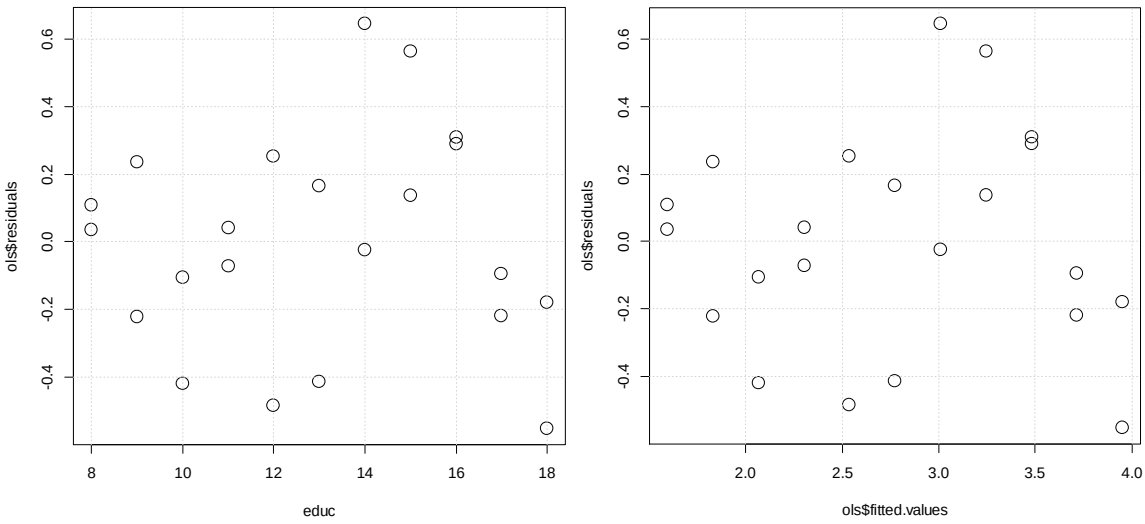
$\widehat{log\,Y} = -0.2903 + 0.2357 X$

Man beachte, dass es keine Rolle spielt, ob man mit den geschätzten Werten für die Grundgesamtheit arbeitet ($\hat{c}_{XY}$, $\hat{\sigma}_X^2$) oder aber mit den Werten der Stichprobe ($c_{XY}$, $\sigma_X^2$).
Wir prüfen nun die Modellannahmen:

1. Linearität: Es muss ein linearer Zusammenhang zwischen X und Y vorliegen. Das Streudiagramm zeigt, dass zumindest näherungsweise ein linearer Zusammenhang vorliegt.

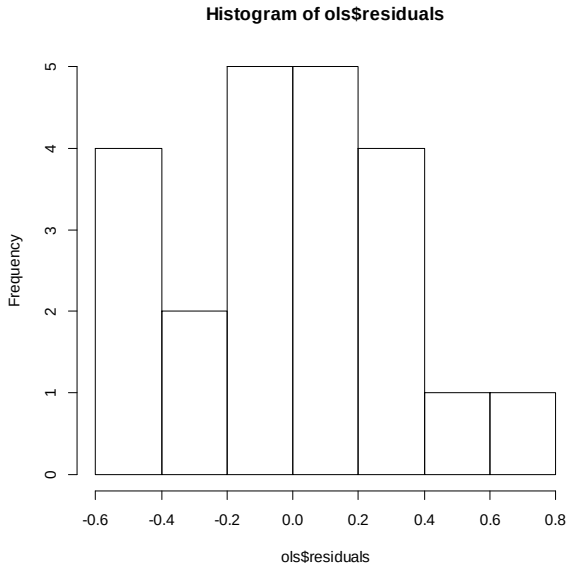


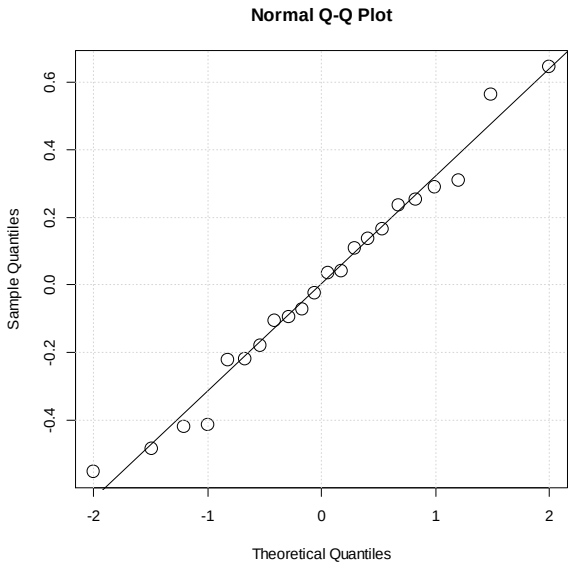

2. Unabhängigkeit der Fehlerterme: Die Residuen dürfen sich nicht gegenseitig beeinflussen. Da wir von einer Zufallsstichprobe ausgehen können, kann diese Annahme als erfüllt angesehen werden.
3. Konstante Varianz der Fehlerterme: Die Streuung der Residuen um die Regressionsgerade muss konstant sein. Hierfür erstellen wir ein Streudiagramm von $X$ bzw. der prognostizierten Werte und den Residuen. Hier darf keinerlei Struktur erkennbar sein. Die beiden folgenden Abbildungen zeigen dies.




4. Normalverteilung der Fehlerterme: Die Fehlerterme in der Grundgesamtheit müssen einer Normalverteilung folgen. Hierfür betrachten wir die Residuen in der Stichprobe.

Folgen diese näherungsweise einer Normalverteilung, kann diese Annahme als erfüllt angesehen werden. Wie beide Abbildungen zeigen, ist dies hier der Fall.




Wie wollen nun mit einem Hypothesentest prüfen, ob der Anstiegsparameter unseres Modells statistisch signifikant von Null verschieden ist.
Ausgangspunkt sind die Hypothesen:
$H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$
für den wahren Anstiegsparameter in der Grundgesamtheit. Als Testvariable kann
$\frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} \sim T_{df=n-2}$
verwendet werden. Der Standardfehler des Anstiegsparameters wird dabei geschätzt mit
$SE(\hat{\beta}_1) = \frac{\hat{\sigma}_E}{\hat{\sigma}_X \sqrt{n-1}}$
Hierbei ist
$\hat{\sigma}_E = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$
die geschätzte Standardabweichung der Fehlerterme, $n$ ist die Zahl der Beobachtungen und $\hat{\sigma}_X$ ist die geschätzte Standardabweichung von $X$.
Basierend auf den $n = 22$ Beobachtungen erhalten wir
$SE(\hat{\beta}_1) = \frac{\hat{\sigma}_E}{\hat{\sigma}_X \sqrt{n-1}} = \frac{0.3263489}{3.236694\sqrt{21}} = 0.02200244$
und für die Prüfgröße
$t = \frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} = \frac{0.2356746 - 0}{0.02200244} = 10.71129$
Dieser Wert ist größer als der kritische Wert, hier das 97.5%-Quantil der t-Verteilung bei 20 Freiheitsgraden, $t_{[0.975]df=20} = 2.085963$. Der p-value ist $2 * (1 - F_t(t, df = 20)) \approx 0$. Wir können also die Nullhypothese verwerfen und schließen, dass der Anstiegsparameter statistisch signifikant von Null verschieden ist. Bildung hat also einen statistisch signifikanten, positiven Effekt auf den (logarithmierten) Stundenlohn.

**Lösung mit R**

```
data <- read.csv("Bsp._5.4.csv")
head(data)
attach(data)

summary(educ)
summary(wage)
```

```
summary(log(wage))

# Streudiagramm
plot(educ, log(wage),
xlab = "Jahre in Bildung",
ylab = "log(Stundenlohn in USD)",
cex = 2)
abline(lm(log(wage) ~ educ))
grid()

# Schätzung des Modells
ols <- lm(log(wage) ~ educ)
summary(ols)

cov(log(wage), educ)
var(educ)

b1 <- cov(log(wage), educ)/var(educ); b1
b0 <- mean(log(wage)) - b1*mean(educ); b0

n <- length(educ); n
err <- ols$residuals
pred <- ols$fitted.values
SD_E <- sqrt(sum(err^2)/(n-2)); SD_E
# oder auch
sqrt(sum((log(wage)-pred)^2)/(n-2))

SE_b1 <- SD_E/(sd(educ)*sqrt(n-1)); SE_b1
t_b1 <- (b1 - 0)/SE_b1; t_b1
qt(0.975, df = n-2) # krit. Wert
p.value_b1 <- 2*(1-pt(t_b1, df = n-2)); p.value_b1

# Plot predicted values vs. residuals
plot(ols$fitted.values, ols$residuals, cex = 2)
grid()

# Plot X values vs. residuals
plot(educ, ols$residuals, cex = 2)
grid()

# Histogramm residuals
hist(ols$residuals)

# qqplot
qqnorm(ols$residuals, cex = 2)
qqline(ols$residuals)
grid()
```

***Aufgabe 12.4: Getreide***

*Der Datensatz* `cereals.csv` *enthält Angaben zu Preisen (P) und Mengen (Q) von Weizen und Gerste (vgl. auch* `cereals.txt`*).*
*Schätzen Sie für beide Getreidesorten die Modelle*

$$\hat{Q} = \hat{\beta}_0 + \hat{\beta}_1 P \quad \textit{und} \quad \widehat{\log Q} = \hat{\beta}_0 + \hat{\beta}_1 \log P$$

*Prüfen Sie die Modellannahmen und führen Sie einen Hypothesentest für den Anstiegsparameter durch. Interpretation und Vergleich beider Modelle.*

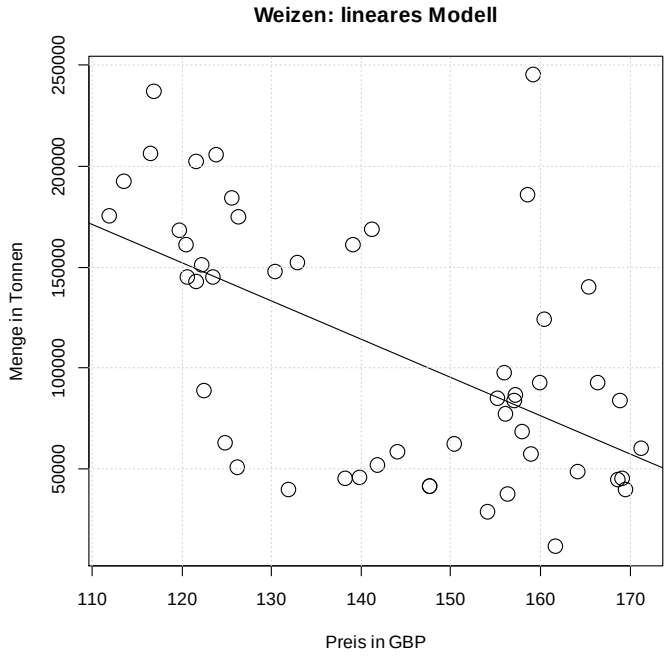
***Hinweis:*** *Im Buch wurden leider die Variablen vertauscht. Die Menge Q ist soll hier die abhängige und der Preis P die unabhängige Variable sein.*

**Lösung**

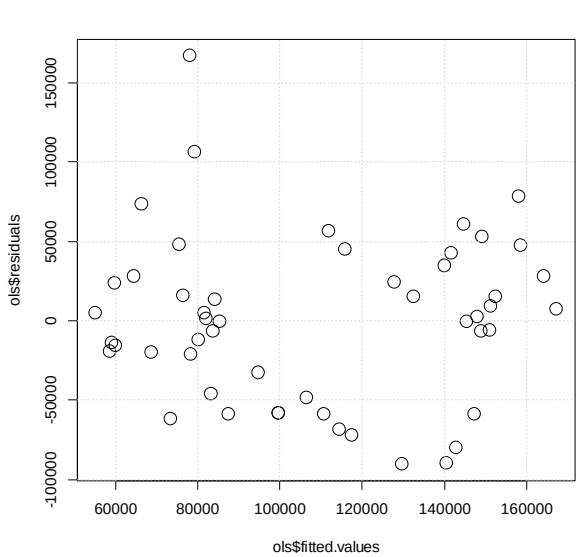
Wir lösen diese Aufgabe nur mit R. Zunächst schätzen wir das lineare Modell ($\hat{Q} = \hat{\beta}_0 + \hat{\beta}_1 P$) für Weizen. Der Regressionsoutput ist:

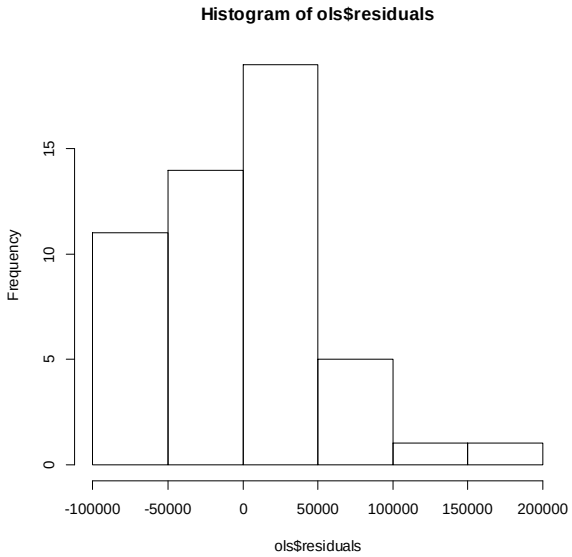
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 378961.8    57819.3   6.554 3.25e-08 ***
p_wheat      -1889.8      400.8  -4.715 2.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52470 on 49 degrees of freedom
Multiple R-squared:  0.3121,    Adjusted R-squared:  0.2981
F-statistic: 22.23 on 1 and 49 DF,  p-value: 2.041e-05
```

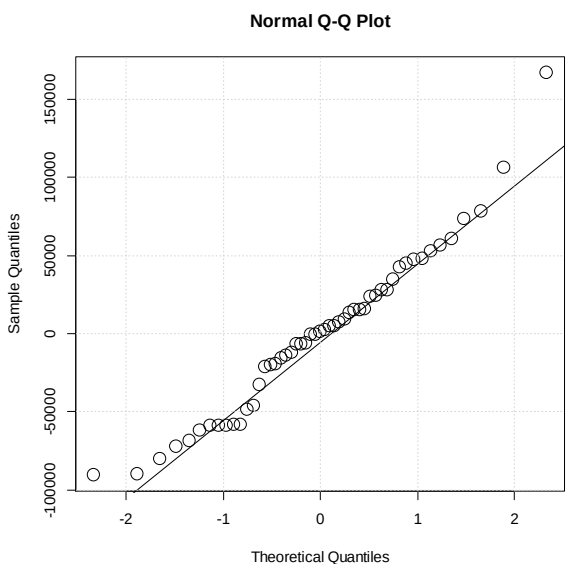
Das Streudiagramm zeigt, dass näherungsweise ein linearer Zusammenhang vorliegt.

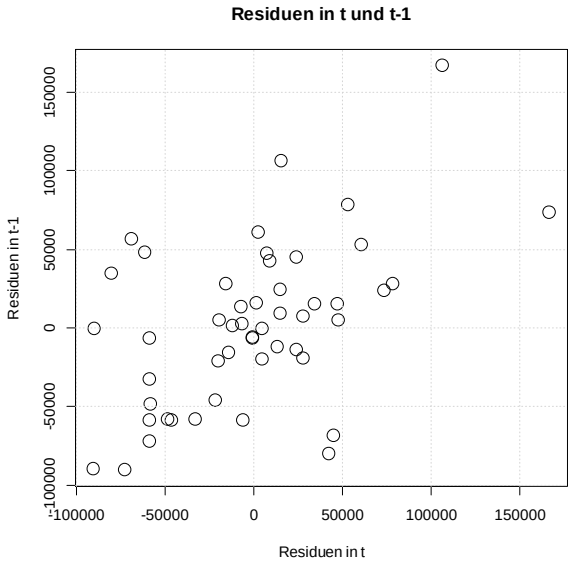



Eine Preiserhöhung um 1 GBP (Pfund Sterling) führt bei Weizen im Mittel zu einer Nachfragereduzierung von 1889.8 Tonnen. Betrachten wir die Abbildungen zum Prüfen der Modellannahmen.





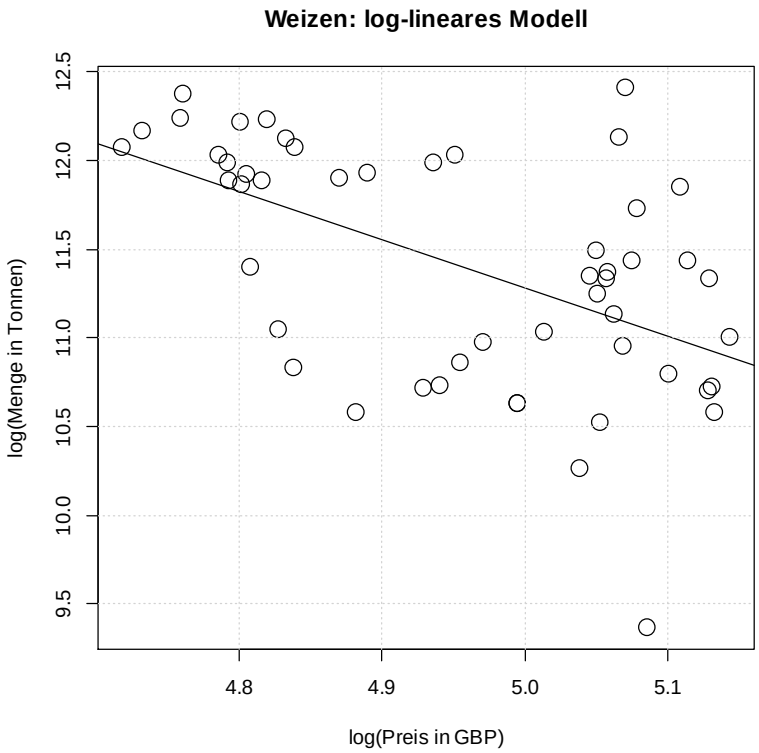




Die Erklärungskraft des Modells ist mit einem $R^2 = 0.31$ relativ hoch. Die Annahmen der linearen Regression sind relativ gut erfüllt. Allerdings liegen hier Zeitreihenbeobachtungen vor (Wochen). Wenn wir die Residuen in $t$ in einem Streudiagramm mit den Residuen in $t-1$ darstellen (Abbildung rechts oben), dann zeigt sich, dass es eine positive Korrelation zwischen den Residuen in $t$ und $t-1$ gibt („Autokorrelation“, d.h., die Fehler korrelieren mit sich selbst). Damit ist die Annahme unabhängiger Residuen verletzt. Dies ist zwar unkritisch für den geschätzten Anstiegsparameter, macht aber den Signifikanztest ungültig. Der p-value ist also mit Vorsicht zu genießen.

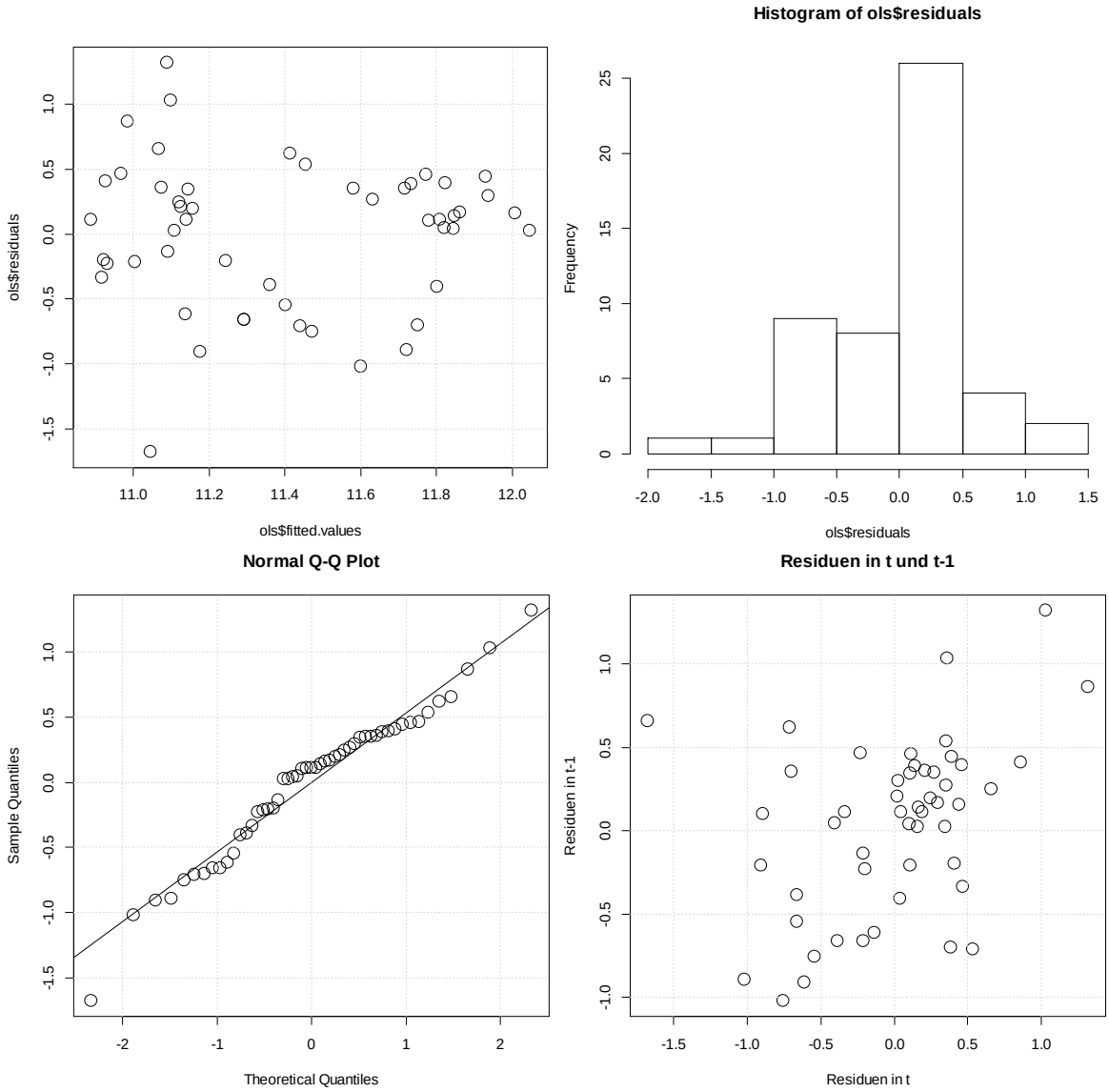
Nun betrachten wir das log-lineare Modell ($\widehat{log\,Q} = \hat{\beta}_0 + \hat{\beta}_1\, log\, P$) für Weizen. Der Regressionsoutput ist:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.9048     3.0367   8.201 9.42e-11 ***
log(p_wheat)  -2.7250     0.6126  -4.448 4.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5696 on 49 degrees of freedom
Multiple R-squared:  0.2876,    Adjusted R-squared:  0.2731
F-statistic: 19.79 on 1 and 49 DF,  p-value: 4.98e-05
```

Das Streudiagramm zeigt, dass näherungsweise ein linearer Zusammenhang vorliegt.




Eine Preiserhöhung um 1% führt im Mittel bei Weizen zu einer Nachfragereduzierung von 2.7%. Es liegt also eine elastische Nachfragereaktion vor. Betrachten wir die Abbildungen zum Prüfen der Modellannahmen.




Die Erklärungskraft des Modells ist mit einem $R^2 = 0.29$ geringfügig schwächer als die des linearen Modells. Die Annahmen der linearen Regression sind relativ gut erfüllt. Allerdings zeigt sich auch hier, dass es eine positive Korrelation zwischen den Residuen in $t$ und $t-1$ gibt („Autokorrelation“, d.h., die Fehler korrelieren mit sich selbst). Damit ist die Annahme unabhängiger Residuen verletzt. Dies ist zwar unkritisch für den geschätzten Anstiegsparameter, macht aber den Signifikanztest ungültig. Damit ist auch hier der p-value mit Vorsicht zu verwenden.

Welches Modell sollte nun verwendet werden? Beide Modelle liefern ein ähnliches $R^2$ und schneiden auch hinsichtlich der Modellannahmen ähnlich ab. Diese Entscheidung hängt von der Verwendung des Anstiegsparameters ab. Üblicherweise ist die konstante Nachfrageelastizität $E_{Q,P} = \frac{\frac{\delta Q}{Q}}{\frac{\delta P}{P}}$ von Interesse. In diesem Fall sollte das log-lineare Modell verwendet werden, denn es gilt:

$$\hat{\beta}_1 = E_{Q,P} = \frac{\frac{\delta Q}{Q}}{\frac{\delta P}{P}}$$

Wir können also den Anstiegsparameter des log-linearen Modells direkt als Nachfrageelastizität interpretieren.

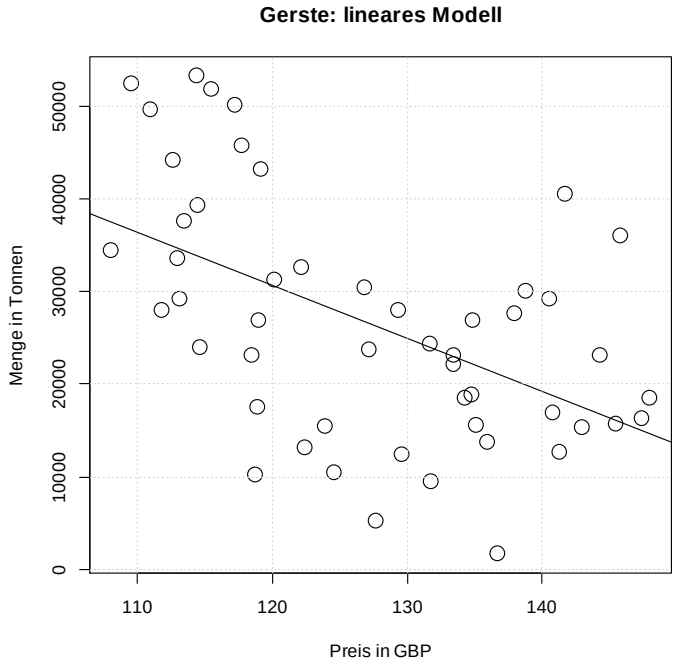
Wenn dagegen eine Prognose über das absolute Nachfrageverhalten nötig ist, sollte besser mit dem linearen Modell gearbeitet werden. In diesem Fall ist:

$\hat{\beta}_1 = \frac{\delta Q}{\delta P}$

Betrachten wir nun Gerste und das lineare Modell.

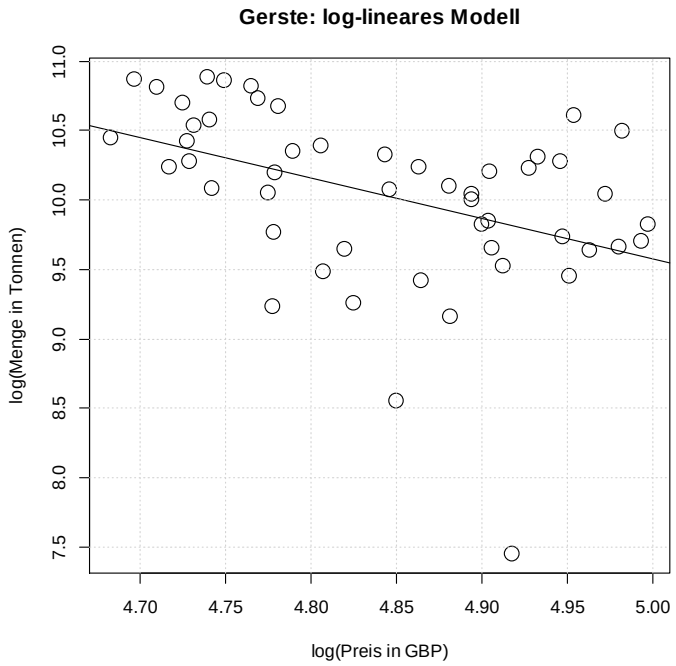
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99719.4    17432.9   5.720  6.3e-07 ***
p_barley      -575.0      136.4  -4.217 0.000106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11290 on 49 degrees of freedom
Multiple R-squared:  0.2663,    Adjusted R-squared:  0.2513
F-statistic: 17.78 on 1 and 49 DF,  p-value: 0.0001063
```

Eine Preiserhöhung um 1 GBP führt bei Gerste im Mittel zu einer Nachfragereduzierung von 575.0 Tonnen. Die Nachfragekurve verläuft also flacher als bei Weizen. Wir verzichten hier auf eine Diskussion der Annahmen (vgl. R).




Für Gerste und das log-lineare Modell ergibt sich:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    24.1153     4.3016   5.606 9.41e-07 ***
log(p_barley)  -2.9082     0.8881  -3.274  0.00195 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5775 on 49 degrees of freedom
Multiple R-squared:  0.1795,    Adjusted R-squared:  0.1628
F-statistic: 10.72 on 1 and 49 DF,  p-value: 0.001946
```




Eine 1%-ige Erhöhung des Preises für Gerste reduziert die Nachfrage um im Mittel 2.9%. Die Nachfrage ist also noch etwas elastischer als für Weizen.
Hinsichtlich der Annahmen gilt das Gleiche wie für Weizen.

### Lösung mit R

```
data <- read.csv("cereals.csv")
head(data)
attach(data)
names(data)
# -----------------------
# Weizen, lineares Modell
ols <- lm(q_wheat ~ p_wheat)
summary(ols)

plot(ols$fitted.values, ols$residuals, cex = 2)
grid()

hist(ols$residuals)
qqnorm(ols$residuals, cex = 2); qqline(ols$residuals)
grid()

e <- ols$residuals
length(e)
plot(e[-51], e[-1],
main = "Residuen in t und t-1",
xlab = "Residuen in t",
ylab = "Residuen in t-1",
cex = 2)
grid()
cor.test(e[-51], e[-1])

# Exkurs zum lag t-1
x <- c(1:5, 6, 3, 9)
x[-1]
# Exkurs Ende

plot(p_wheat, q_wheat,
main = "Weizen: lineares Modell",
xlab = "Preis in GBP",
ylab = "Menge in Tonnen", cex = 2)
abline(lm(q_wheat ~ p_wheat))
```

```
grid()

# -----------------------
# Weizen, log-lineares Modell
ols <- lm(log(q_wheat) ~ log(p_wheat))
summary(ols)

plot(ols$fitted.values, ols$residuals, cex = 2)
grid()

hist(ols$residuals)
qqnorm(ols$residuals, cex = 2); qqline(ols$residuals)
grid()

e <- ols$residuals
length(e)
plot(e[-51], e[-1],
main = "Residuen in t und t-1",
xlab = "Residuen in t",
ylab = "Residuen in t-1",
cex = 2)
grid()
cor.test(e[-51], e[-1])

plot(log(p_wheat), log(q_wheat),
main = "Weizen: log-lineares Modell",
xlab = "log(Preis in GBP)",
ylab = "log(Menge in Tonnen)", cex = 2)
abline(lm(log(q_wheat) ~ log(p_wheat)))
grid()

# -----------------------
# Gerste, lineares Modell
names(data)
ols <- lm(q_barley ~ p_barley)
summary(ols)

plot(ols$fitted.values, ols$residuals, cex = 2)
grid()

hist(ols$residuals)
qqnorm(ols$residuals, cex = 2); qqline(ols$residuals)
grid()

e <- ols$residuals
length(e)
plot(e[-51], e[-1],
main = "Residuen in t und t-1",
xlab = "Residuen in t",
ylab = "Residuen in t-1",
cex = 2)
grid()
cor.test(e[-51], e[-1])

plot(p_barley, q_barley,
main = "Gerste: lineares Modell",
xlab = "Preis in GBP",
ylab = "Menge in Tonnen", cex = 2)
abline(lm(q_barley ~ p_barley))
grid()

# -----------------------
# Gerste, log-lineares Modell
ols <- lm(log(q_barley) ~ log(p_barley))
summary(ols)

plot(ols$fitted.values, ols$residuals, cex = 2)
grid()
```

```
hist(ols$residuals)
qqnorm(ols$residuals, cex = 2); qqline(ols$residuals)
grid()

e <- ols$residuals
length(e)
plot(e[-51], e[-1],
main = "Residuen in t und t-1",
xlab = "Residuen in t",
ylab = "Residuen in t-1",
cex = 2)
grid()
cor.test(e[-51], e[-1])

plot(log(p_barley), log(q_barley),
main = "Gerste: log-lineares Modell",
xlab = "log(Preis in GBP)",
ylab = "log(Menge in Tonnen)", cex = 2)
abline(lm(log(q_barley) ~ log(p_barley)))
grid()
```