R-Einführungsübung

1. Visualisierung von Daten und Berechnung von Kenngrössen

Das Dataframe d.fuel enthält die Daten verschiedener Fahrzeuge aus einer amerikanischen Untersuchung der 80er-Jahre. Jede Zeile (row) enthält die Daten eines Fahrzeuges (ein Fahrzeug entspricht einer Beobachtung).

a) Lesen Sie die Daten ein mit den folgenden R-Befehlen:
`t.file <- "http://stat.ethz.ch/Teaching/Datasets/NDK/d.fuel.dat"`
`d.fuel <- read.table(t.file,header=T,sep=",")`[1]
Das Argument `sep=","` braucht es, weil die Kolonnen im File `d.fuel.dat` durch Kommata getrennt sind. Sie können übrigens den Inhalt des Files `d.fuel.dat` mit einem Internetbrowser anschauen, indem Sie die obige URL in Ihren Browser eintippen.

b) Betrachten Sie die eingelesenen Daten.
Im File `d.fuel.dat` wurden die Zeilen durchnumeriert und daher steht in der ersten Spalte die Nummer der Zeile. Die Spalten (columns) enthalten die folgenden Variablen:

weight:	Gewicht in Pounds (1 Pound = 0.45359 kg)
mpg:	Reichweite in Miles Per Gallon (1 gallon = 3.789 l; 1 mile = 1.6093 km)
type:	Autotyp

c) Wählen Sie nur die fünfte Zeile des Dataframe d.fuel aus. Welche Werte stehen in der fünften Zeile?

d) Wählen Sie nun die erste bis fünfte Beobachtung des Datensatzes aus.
So lässt sich übrigens bei einem unbekannten Datensatz ein schneller Überblick über die Art des Dataframe gewinnen.

e) Zeigen Sie gleichzeitig die 1. bis 3. und die 57. bis 60. Beobachtung des Datensatzes an.

f) Berechnen Sie den Mittelwert der Reichweiten aller Autos in Miles/Gallon.

g) Berechnen Sie den Mittelwert der Reichweite der Autos 7 bis 22.

h) Erzeugen Sie einen neuen Vektor t.kml, der alle Reichweiten in km/l, und einen Vektor t.kg, der alle Gewichte in kg enthält.

i) Berechnen Sie den Mittelwert der Reichweiten in km/l und denjenigen der Fahrzeuggewichte in kg.

j) Zeichnen Sie ein Streudiagramm, welches den Verbrauch pro 100km als Funktion des Gewichtes in kg darstellt.

k) Machen Sie eine Stamm-Blatt-Darstellung der Benzinverbräuche pro 100 km. Bestimmen Sie den minimalen und maximalen Verbrauch.
R-Hinweis: `stem()`

[1] Alternativ können Sie den Dataframe `d.fuel.dat` von der Internetseite `http://stat.ethz.ch/Teaching/Datasets/NDK` in einen für diesen Zweck erstellten Ordner `Datasets` in Ihrem Home-Directory kopieren (speichern unter `T:/ndk../Datasets/d.fuel.txt`) und dann von dort in R einlesen mit dem Befehl `d.fuel <- read.table("T:/ndk../Datasets/d.fuel.txt",header=T,sep=",")`

l) Zeichnen Sie zuerst ein Histogramm des Verbrauchs der Autos (pro 100km) mit den Defaulteinstellungen und dann ein Histogramm mit 15 Klassen statt nur 8, einer x-Achse von 0 bis 15 und einem Titel.

m) Zeichnen Sie einen Boxplot der Benzinverbräuche.
R-Hinweis: `boxplot()`

n) Vergleichen Sie die Standardabweichung und den MAD der Benzinverbräuche miteinander (vgl. Stat. Datenanalyse, Kap. 2.3).
R-Hinweis: `mad(), sd()`

o) Vergleichen Sie den Mittelwert und den Median der Benzinverbräuche in l/100km.

2. Korrelationen (R-Funktion: `cor()`)

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

b) Zeichnen Sie die Streudiagramme t.y vs. t.x und t.y1 vs t.x1 .

c) Berechnen Sie die Korrelationskoeffizienten zwischen t.x und t.y bzw. zwischen t.x1 und t.y1. Warum sind die beiden Korrelationen so verschieden (vgl. Stat. Datenanalyse, Abschnitt 3.2.h)?

Musterlösung zur R-Einführungsübung

1. Visualisierung von Daten und Berechnung von Kenngrössen

a) Siehe Aufgabenstellung.

b) Um die Daten in Tabellenform zu sehen, tippt man den Namen des Objektes ein.

```
> d.fuel
    X weight mpg    type
1   1   2560  33   Small
2   2   2345  33   Small
3   3   1845  37   Small
4   4   2260  32   Small
5   5   2440  32   Small
:   :     :    :     :
:   :     :    :     :
59 59   3185  20     Van
60 60   3690  19     Van
```

c) Auswählen der fünften Beobachtung:

```
> d.fuel[5,]
  X weight mpg  type
5 5   2440  32 Small
```

d) Auswählen der 1. bis 5. Beobachtung:

```
> d.fuel[1:5,]
  X weight mpg  type
1 1   2560  33 Small
2 2   2345  33 Small
3 3   1845  37 Small
4 4   2260  32 Small
5 5   2440  32 Small
```

e) Auswählen der 1. bis 3. und 57. bis 60. Beobachtung:

```
> d.fuel[c(1:3,57:60),]
    X weight mpg  type
1   1   2560  33 Small
2   2   2345  33 Small
3   3   1845  37 Small
57 57   3735  19   Van
58 58   3415  20   Van
59 59   3185  20   Van
60 60   3690  19   Van
```


f) Die Werte der Reichweiten stehen in der dritten Spalte, die `mpg` heisst. Zur Berechnung des Mittelwertes gibt es verschiedene Möglichkeiten, welche sich in der Art der Datenselektion unterscheiden:

```
> mean(d.fuel[,3])
[1] 24.58333
> mean(d.fuel[,"mpg"])
[1] 24.58333
> mean(d.fuel$mpg)
[1] 24.58333
```

g) Auch hier gibt es wieder verschiedene Möglichkeiten. Eine davon ist:

```
> mean(d.fuel[7:22,"mpg"])
[1] 27.75
```

h) Umrechnung der Miles Per Gallon in Kilometer pro Liter und der Pounds in Kilogramm:

```
> t.kml <- d.fuel[,"mpg"]*1.6093/3.789
> t.kg  <- d.fuel[,"weight"]*0.45359
```

i) Mittelwert der Reichweite und des Gewichtes:

```
> mean(t.kml)
[1] 10.44127
> mean(t.kg)
[1] 1315.789
```

Der Mittelwert der Reichweite kann auch wie folgt berechnet werden (siehe Stat. Datenanalyse, Kapitel 2.6): `> mean(d.fuel[,"mpg"])*1.6093/3.789`

j) Verbrauch als Funktion des Gewichtes:

```
> plot(t.kg,100/t.kml)
```

k) Stem-and-leaf-Plot des Benzinverbrauchs:

```
> stem(100/t.kml)

  The decimal point is at the |

   6 | 479
   7 | 1111448
   8 | 1447777
   9 | 111114448888
  10 | 2222222277777
  11 | 22222288888
  12 | 444
  13 | 1111

> min(100/t.kml)
[1] 6.363351
> max(100/t.kml)
[1] 13.08022
```

l) Histogramm des Verbrauchs: `hist(100/t.kml)`
Mit 15 Klassen: `hist(100/t.kml,nclass=15)`

Mit x-Achse 0 bis 15: `hist(100/t.kml,nclass=15,xlim=c(0,15))`
Mit Titel: `hist(100/t.kml,nclass=15,xlim=c(0,15),main="Verteilung der Verbraeuche")`

m) Boxplot der Verbräuche zeichnen: `boxplot(100/t.kml)`

n) Vergleich der Standardabweichung mit dem MAD:

```
> sd(100/t.kml)
[1] 1.783549
> mad(100/t.kml)
[1] 1.751184
> mad(100/t.kml,constant=1)
[1] 1.181157
```

Der Befehl `mad` ohne `constant=1` berechnet einen skalierten MAD, sodass der mad bei normalverteilten Daten gerade der Standardabweichung entsprechen würde. Im Buch wurde der MAD ohne Skalierung eingeführt.

o) Vergleich des Mittelwertes mit dem Median:

```
> mean(100/t.kml)
[1] 9.912268
>  median(100/t.kml)
[1] 10.23669
```

2. Korrelationen

a) Erzeugen der Vektoren:

```
> t.x  <- (-10):10
> t.x1 <- 0:10
> t.y  <- t.x^2
> t.y1 <- t.x1^2
```

b)

```
> par(mfrow=c(1,2))   # zwei Grafiken im Grafikfenster
> plot(t.x,t.y)
> plot(t.x1,t.y1)
```

c)

```
> cor(t.x,t.y)
[1] 0
> cor(t.x1,t.y1)
[1] 0.9631427
```

Die Korrelation zwischen t.x und t.y ist 0, weil die Daten symmetrisch zur y-Achse liegen.
Im zweiten Fall ist die Korrelation hoch (0.96), obwohl die Daten keine lineare Beziehung aufweisen. Der Grund dafür ist, dass x und y monoton steigen. Wenn statt der üblichen Korrelation die Rangkorrelation verwendet worden wäre, würde der Koeffizient exakt 1.0 betragen.

Serie 2

Vorbesprechung: 27/28. Februar 2013

Aufgabe 1

Bei der Ermittlung der landwirtschaftlichen Nutzfläche von Bauernhöfen in einem Bezirk ergaben sich folgende Werte (in ha):

2.1 2.4 2.8 3.1 4.2 4.9 5.1 6.0 6.4 7.3 10.8 12.5 13.0 13.7 14.8 17.6 19.6 23.0 25.0 35.2 39.6

(**a**) Berechnen Sie die Summen $\sum x_i$ und $\sum x_i^2$.

(**b**) Bestimmen Sie den Median .

(**c**) Berechnen Sie den Mittelwert und die Standardabweichung .

Aufgabe 2

Gegeben sind die Datenpaare (x, y)

x	2	2	6	7	7	8	8	9
y	11	14	14	16	27	27	27	38

(**a**) Gesucht sind die Summen $\sum x_i,\ \sum x_i^2,\ \sum y_i,\ \sum y_i^2$ und $\sum x_i \cdot y_i$.

(**b**) Verifizieren Sie die Gleichheit der Formeln $s_{xx} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$ und $s_{xx} = \frac{1}{n-1}\sum_{j=1}^{n} x_j^2 - n \cdot \overline{x}^2$ anhand der Zahlen in der obigen Tabelle .

(**c**) Beweisen Sie die Gleichheit von Teilaufgabe (b) allgemein .

Aufgabe 3

Zeigen Sie, dass

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2.$$

den kleinsten Wert annimmt, falls die Parameter

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ a &= \bar{y} - \hat{b}\bar{x} \end{aligned}$$

sind.

Aufgabe 4

Bei einer Firma werden in einem Monat 400 Lebensversicherungsverträge abgeschlossen. Nachstehend ist die klassifizierte Häufigkeitsverteilung für die Versicherungssummen angegeben.

Versicherungssumme (Tausend Fr) von... bis unter...	Anzahl der Verträge
4-10	20
10-20	160
20-30	80
30-40	40
40-80	88
80-120	12

(**a**) Man zeichne ein Histogramm und die Summenkurve für die relativen Häufigkeiten .

(**b**) Untersuchen Sie wieviel Prozent der Versicherten mit höchstens 18'000.- versichert sind, sowie mit welchem Betrag die 20% Personen, die am höchsten versichert sind, mindestens versichert sind. Bestimmen Sie zudem Median und Mittelwert der Verteilung .

(**c**) Berechnen Sie die Standardisierung der Daten .

(**d**) Wenn n Daten $(x_i)_{1\leq i\leq n}$ standardisiert sind (d.h. wenn $\overline{x}=0,\ s_x=1$ gilt), wie gross ist dann $\sum_{i=1}^{n} x_i^2$?

Aufgabe 5

Der Geysir Old Faithful im Yellowstone National Park ist eine der bekanntesten heissen Quellen. Für die Zuschauer und den Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse.
Im File `http://stat.ethz.ch/Teaching/Datasets/geysir.dat` sind die Messungen vom 1.8.1978−8.8.1978 in 3 Spalten abgelegt: "`Tag`", "`Zeitspanne`" und "`Eruptionsdauer`".

(**a**) Zeichnen Sie Histogramme von der Zeitspanne zwischen zwei Ausbrüchen:

```
> geysir <- read.table("http://stat.ethz.ch/Teaching/Datasets/geysir.dat",
+ header = TRUE) ## Datensatz einlesen
> par(mfrow = c(2,2)) ## 4 Grafiken im Grafikfenster
> hist(geysir[,"Zeitspanne"])
> hist(geysir[,"Zeitspanne"], breaks = 20)
> hist(geysir[,"Zeitspanne"], breaks = seq(41, 96, by = 11))
```

Was fällt auf? Was ist der Unterschied zwischen den drei Histrogrammen?
Bemerkung: Wenn man die Anzahl Klassen mit `breaks` = 20 vorgibt, so wird dies nur als "Vorschlag" interpretiert und intern unter Umständen abgeändert.

(**b**) Zeichnen Sie Histogramme (Anzahl Klassen variieren) von der Eruptionsdauer:

```
> hist(geysir[,"Eruptionsdauer"])
```

Was fällt auf? Vergleichen Sie mit der ersten Teilaufgabe.

Aufgabe 6

21 Labors bestimmten den Kupfergehalt von 9 verschiedenen Klärschlammproben. Die Daten stehen im Data Frame `klaerschlamm` zur Verfügung. Die erste Spalte bezeichnet das Labor, die restlichen 9 Spalten sind die verschiedenen Klärschlammprobe. Die Daten (in mg/kg) kann man mit dem Befehl

```
> url <- "http://stat.ethz.ch/Teaching/Datasets/klaerschlamm.dat"
> schlamm.all <- read.table(url, header = TRUE)
> schlamm <- schlamm.all[,-1] ## Labor-Spalte entfernen
```

einlesen.

(**a**) Erstellen Sie für jede Probe einen Boxplot, und berechnen Sie jeweils das arithmetische Mittel und den Median. Bei welchen Proben gibt es Ausreisser, und wo unterscheiden sich arithmetisches Mittel und Median wesentlich? Bei welchen der 9 Proben ist es plausibel, dass die wahre Konzentration unter 400 mg/kg liegt?
R-Hinweise: `summary(schlamm); boxplot(schlamm)` .

(**b**) Erstellen Sie für jedes Labor einen Boxplot der Messfehler. Unter dem Messfehler eines Labors bei einer Probe verstehen wir den gemessenen Wert minus den Median über alle Labors. Welche der 21 Labors haben systematische Fehler in ihrem Analyseverfahren? Welche haben grosse Zufallsfehler, und bei welchen Labors ist die Qualität der Analysen besonders gut?
R-Hinweise:

```
> ## Fuer jede Spalte Median berechnen
> med <- apply(schlamm, 2, median)
> ## Median von jeder *Spalte* abziehen
> schlamm.centered <- scale(schlamm, scale = FALSE, center = med)
> ## Boxplot zeichnen. Dazu zuerst data-frame transponieren
> boxplot(data.frame(t(schlamm.centered)))
```

Lösungen zu Serie 2

Vorbesprechung: 27/28. Februar 2013

Aufgabe 1

```
> x<-c(2.1,2.4,2.8,3.1,4.2,4.9,5.1,6.0,6.4,7.3,10.8,12.5,13.0,13.7,14.8,17.6,19.6,23.0,25.0,35.2,39.6)
```

(**a**) $\sum x_i =$

```
> sum(x)
```

```
[1] 269.1
```

$\sum x_i^2 =$

```
> sum(x^2)
```

```
[1] 5729.27
```

(**b**) Median:

```
> median(x,na.rm=FALSE)
```

```
[1] 10.8
```

(**c**) Mittelwert:

```
> round(mean(x),4)
```

```
[1] 12.8143
```

Standardabweichung:

```
> round(sd(x),4)
```

```
[1] 10.6793
```

Aufgabe 2

```
> x<-c(2,2,6,7,7,8,8,9)
> y<-c(11,14,14,16,27,27,27,38)
```

(**a**) $\sum x_i =$

```
> sum(x)
```

```
[1] 49
```

$\sum x_i^2 =$

```
> sum(x^2)
```

```
[1] 351
```

$\sum y_i =$

```
> sum(y)
```

```
[1] 174
```

$\sum y_i^2 =$

```
> sum(y^2)
```

```
[1] 4400
```

$\sum x_i \cdot y_i =$

```
> sum(x*y)
```

```
[1] 1209
```

(**b**) $s_{xx} =$

```
> n <- length(x)
> 1/(n-1)*sum((x-mean(x))^2)
```

```
[1] 7.267857
```

alternativ

```
> cov(x,x)

[1] 7.267857
```

$\frac{1}{n-1}\left(\sum_{j=1}^{n} x_j^2 - n \cdot \overline{x}^2\right) =$

```
> n <- length(x)
> 1/(n-1)*(sum(x^2)-n*mean(x)^2)

[1] 7.267857
```

(**c**)

$$\begin{aligned}\sum_{j=1}^{n} (x_j - \overline{x})^2 &= \sum_{j=1}^{n} \left(x_j^2 - 2x_j\overline{x} + \overline{x}^2\right) \\ &= \sum_{j=1}^{n} \left(x_j^2 + \overline{x}^2\right) - 2\overline{x}\sum_{j=1}^{n} x_j \\ &= n\overline{x}^2 + \sum_{j=1}^{n} x_j^2 - 2\overline{x}n\overline{x} \\ &= \sum_{j=1}^{n} x_j^2 - n \cdot \overline{x}^2 \, .\end{aligned}$$

Aufgabe 3

Gegeben ist

$$F(a,b) \equiv \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 \, .$$

Folgende zwei Bedingungen müssen erfüllt sein:

$$\frac{\partial}{\partial a} F(a,b) = \sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) \stackrel{!}{=} 0 \tag{1}$$

$$\frac{\partial}{\partial b} F(a,b) = \sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) \cdot x \stackrel{!}{=} 0 \tag{2}$$

Aus Gleichung (1) folgt:

$$\sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) = 0$$

$$\Longleftrightarrow$$

$$\sum_{i=1}^{n} a = na = \sum_{i=1}^{n} \left(y_i - bx_i\right)$$

$$\Longleftrightarrow$$

$$a = \frac{\sum_{i=1}^{n} \left(y_i - bx_i\right)}{n} = \overline{y} - b\overline{x}$$

Aus Gleichung (2) folgt:

$$\sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) \cdot x = 0$$

$$\Longleftrightarrow$$

$$\sum_{i=1}^{n} \left(y_i - \overline{y} + b\overline{x} - bx_i\right) \cdot x_i = 0$$

$$\Longleftrightarrow$$

$$b \cdot \sum_{i=1}^{n} (x_i - \overline{x})\, x_i = \sum_{i=1}^{n} (y_i - \overline{y}) \cdot x$$

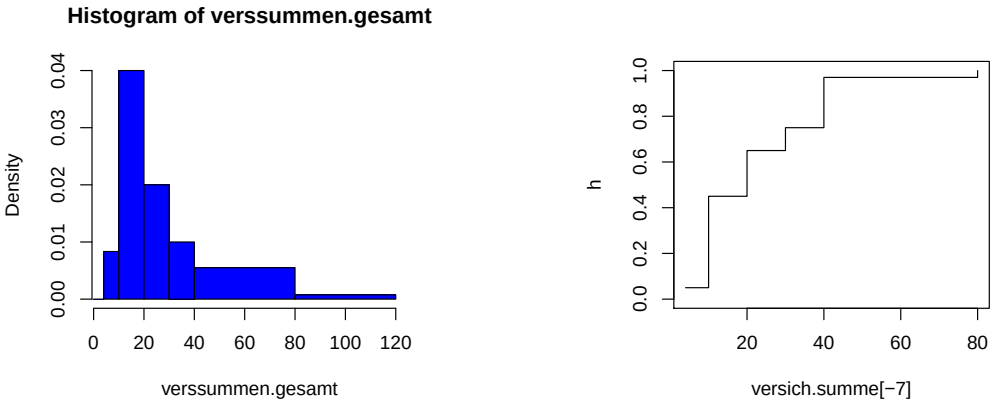
$$\Longleftrightarrow$$

$$b = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, x_i}{\sum_{i=1}^{n} (x_i - \overline{x})\, x_i} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}\,.$$

Aufgabe 4

```
> versich.summe <- c(4,10,20,30,40,80,120)
> mittlere.versich.summen <- c(7,15,25,35,60,100)
> anzahl.vertraege <- c(20,160,80,40,88,12)
> verssummen.gesamt <- rep(mittlere.versich.summen,anzahl.vertraege)
> brk <- c(0,4,10,20,30,40,80,120)
> h <- cumsum(anzahl.vertraege)/400

> hist(verssummen.gesamt,breaks=brk,col='blue',xlim=c(4,120))

> plot(versich.summe[-7],h,type="s",ylim=c(0,1))
```

(**a**) Histogramm und Summenkurve:



(**b**) Aus der empirischen Verteilungsfunktionskurve ergibt sich, dass 5% mit höchstens 18'000 Fr. versichert sind; die obersten 20% mit mind. 40'000 Fr. versichert. Der Median der Versicherungssumme ergibt

```
> median(verssummen.gesamt)

[1] 25
```

also einen Wert zwischen 20'000 und 30'000 Fr. Der Mittelwert der Versicherungssummen ist

```
> mean(verssummen.gesamt)

[1] 31.05
```

also ein Wert zwischen 30'000 und 40'000 Fr.

(**c**) Wir standardisieren den Vektor, der alle mittleren Versicherungssummen enthält, und lesen nur die Werte für jeden Verssicherungssummenbereich heraus

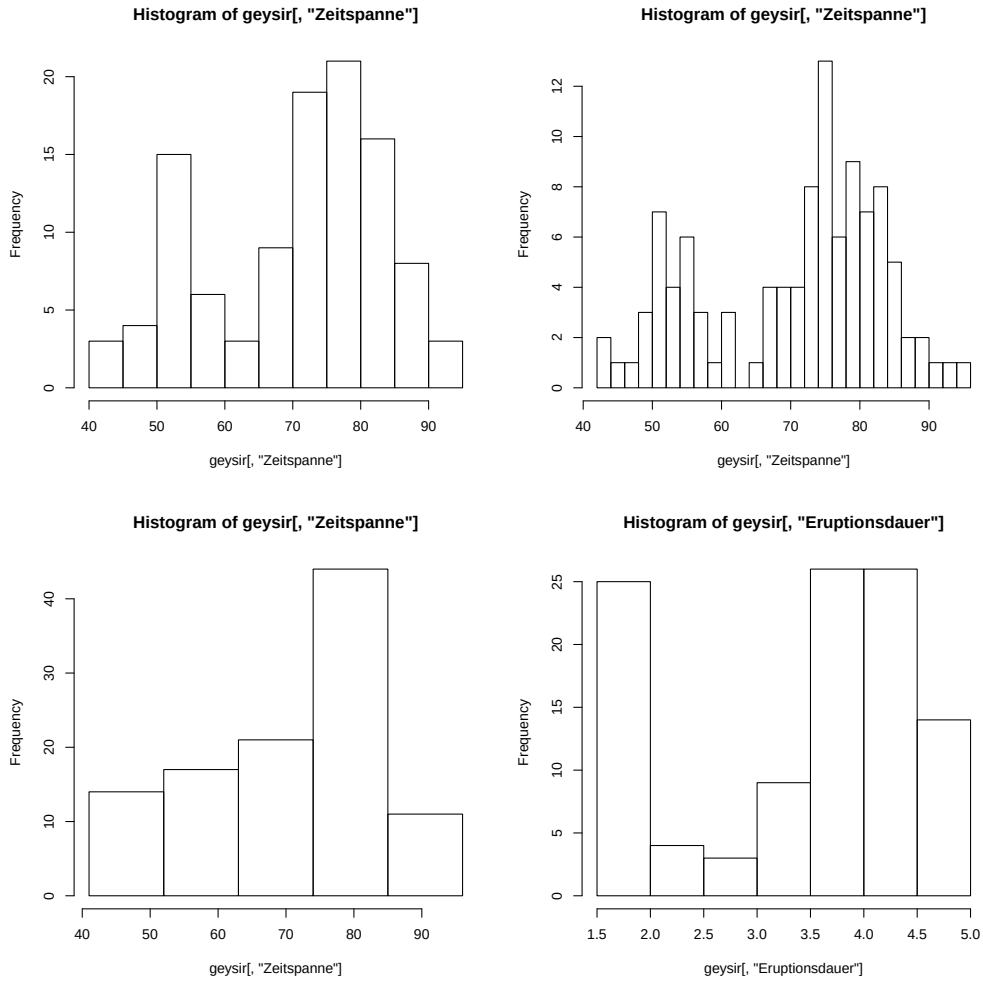
```
> unique(scale(verssummen.gesamt))

            [,1]
[1,] -1.1105021
[2,] -0.7411043
[3,] -0.2793571
[4,]  0.1823902
[5,]  1.3367582
[6,]  3.1837471
```

(**d**) $n - 1$

Aufgabe 5

```
> geysir <- read.table("http://stat.ethz.ch/Teaching/Datasets/geysir.dat",
+ header = TRUE) ## Datensatz einlesen
> par(mfrow = c(2,2)) ## 4 Grafiken im Grafikfenster
> ## Histogramme zeichnen
> hist(geysir[,"Zeitspanne"])
> hist(geysir[,"Zeitspanne"], breaks = 20)
> hist(geysir[,"Zeitspanne"], breaks = seq(41, 96, by = 11))
> hist(geysir[,"Eruptionsdauer"])
```

Die ersten drei Histogramme in der Abbildung unten zeigen die Intervalle zwischen zwei Ausbrüchen von Old Faithful. Aufallend ist, dass Zeitspannen um 55 Minuten aber auch zwischen 70 bis 85 Minuten häufiger vorkommen als andere Intervalle. So eine Verteilung mit zwei Gipfeln heisst auch *bimodal*.
Werden die Klassenbreiten ungeschickt gewählt, entdeckt man diese Besonderheit der Geysirdaten nicht. Das ist im dritten Histogramm passiert. Das Beispiel illustriert, dass die richtige Wahl der Klassenbreiten- bzw. grenzen wohlüberlegt sein mus.
Das vierte Histogramm schliesslich zeigt die Häufigkeiten verschiedener Eruptionsdauern. Hier sind die beiden Gipfel sehr deutlich erkennbar: "Entweder ist der Ausbruch sofort wieder vorbei, oder er dauert mindestens dreieinhalb Minuten". Ob die Dauer eines Ausbruchs aber etwas zu tun hat mit der Dauer des vorangegangenen Ruheintervalls (mit anderen Worten: ob die



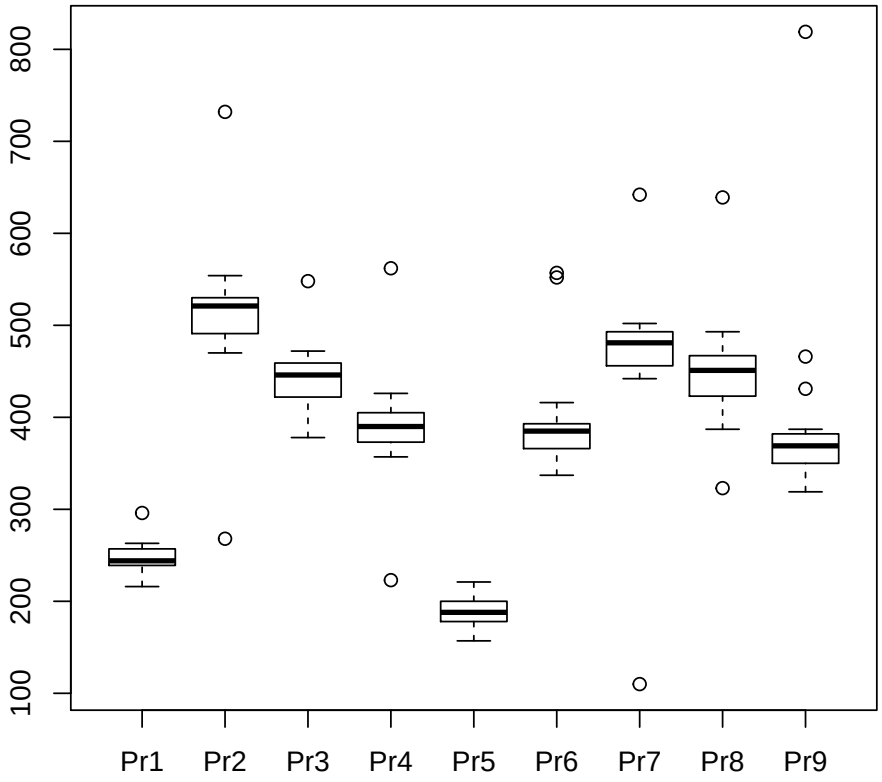
Gipfel des Histogramms aus Teilaufgabe b) den Gipfeln der Histogramme aus Teilaufgabe a) entsprechen), kann man aufgrund dieser Darstellungen nicht sagen.

Aufgabe 6

```
> url <- "http://stat.ethz.ch/Teaching/Datasets/klaerschlamm.dat"
> schlamm.all <- read.table(url, header = TRUE)
> schlamm <- schlamm.all[,-1] ## Labor-Spalte entfernen
```

(**a**) Aus den Boxplots erkennen wir, dass es vor allem bei den Proben 2, 4, 6, 7, 8 und 9 Ausreisser gibt. Das arithmetische Mittel und der Median unterscheiden wesentlich bei den Proben 2, 6, 7 und 9.

```
> boxplot(schlamm)
```



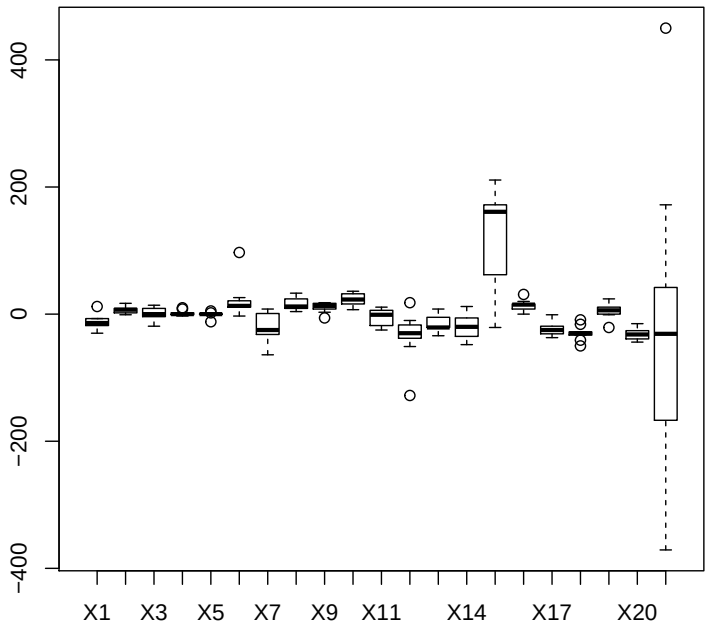

```
> summary(schlamm)
```

Pr1	Pr2	Pr3	Pr4	
Min. :216.0	Min. :268.0	Min. :378.0	Min. :223.0	
1st Qu.:239.0	1st Qu.:491.0	1st Qu.:422.0	1st Qu.:373.0	
Median :244.0	Median :521.0	Median :446.0	Median :390.0	
Mean :246.1	Mean :511.4	Mean :443.4	Mean :389.2	
3rd Qu.:257.0	3rd Qu.:530.0	3rd Qu.:459.0	3rd Qu.:405.0	
Max. :296.0	Max. :732.0	Max. :548.0	Max. :562.0	
Pr5	**Pr6**	**Pr7**	**Pr8**	**Pr9**
Min. :157.0	Min. :337.0	Min. :110.0	Min. :323	Min. :319.0
1st Qu.:178.0	1st Qu.:366.0	1st Qu.:456.0	1st Qu.:423	1st Qu.:350.0
Median :188.0	Median :385.0	Median :481.0	Median :451	Median :369.0
Mean :188.2	Mean :394.9	Mean :465.5	Mean :450	Mean :388.9
3rd Qu.:200.0	3rd Qu.:393.0	3rd Qu.:493.0	3rd Qu.:467	3rd Qu.:382.0
Max. :221.0	Max. :557.0	Max. :642.0	Max. :639	Max. :819.0

Bei den Proben 1 und 5 ist es plausibel, dass die Konzentration unter 400 mg/kg liegt, während wir bei Probe 2, 3, 7 und 8 dazu tendieren, den Grenzwert 400 mg/kg als überschritten zu betrachten. Die übrigen Proben, Probe 4, 6 und 9 sind eher Grenzfälle. Die Konzentrationen scheinen zwar unter 400 mg/kg zu liegen, die drei Proben weisen jedoch jeweils extreme Ausreisser über dem Grenzwert auf.

(**b**) Als erstes stechen die Messungen der Labors 15 und 21 ins Auge. Beide haben sowohl eine grosse Standardabweichung als auch systematische Fehler. Die Labors 6 und 12 haben beide Ausreisser zu verzeichnen. Die Labors 1, 7, 12, 13, 14, 17, 18, 20 und 21 geben systematisch zu kleine Werte an, während die Labors 6, 8, 10 und 15 zu grosse Werte erhalten. Die Labors 2, 3, 4, 5 und 19 scheinen zuverlässige Untersuchungen durchzuführen. Sowohl systematische wie auch Zufallsfehler scheinen sich hier in Grenzen zu halten.

```
> ## Fuer jede Spalte Median berechnen
> med <- apply(schlamm, 2, median)
> ## Median von jeder *Spalte* abziehen
> schlamm.centered <- scale(schlamm, scale = FALSE, center = med)
> ## Boxplot zeichnen. Dazu zuerst data-frame transponieren
> boxplot(data.frame(t(schlamm.centered)))
```



Serie 3

Vorbesprechung: 6/7. März 2013

Aufgabe 1

Eine Regressionsgerade hat die Gleichung $y = mx + 7.8$. Der Durchschnitt der x-Werte beträgt 7, derjenige der y-Werte ist 12. Die Standardabweichungen betragen $s_x = 2.5$ und $s_y = 1.8$.

(**a**) Berechnen Sie die Kovarianz zwischen den x- und den y-Werten.

(**b**) Berechnen Sie den Korrelationskoffizienten r.

Aufgabe 2

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

(**a**) Beide Ereignisse treten ein.

(**b**) Mindestens eines von beiden Ereignissen tritt ein.

(**c**) Höchstens eines von beiden Ereignissen tritt ein.

(**d**) Keines der beiden Ereignisse tritt ein.

(**e**) Genau eines der Ereignisse tritt ein .

Aufgabe 3

Die Rauchsensoren in einer Fabrik melden ein Feuer mit Wahrscheinlichkeit 0.95. An einem Tag ohne Brand geben sie mit Wahrscheinlichkeit 0.01 falschen Alarm. Pro Jahr rechnet man mit einem Brand.

(**a**) Die Alarmanlage meldet Feuer. Mit welcher Wahrscheinlichkeit brennt es tatsächlich?

(**b**) In einer Nacht ist es ruhig (kein Alarm). Mit welcher Wahrscheinlichkeit brennt es tatsächlich nicht?

Aufgabe 4

Bei einem Zufallsexperiment werden zwei Würfel gleichzeitig geworfen. Wir nehmen an, dass sie "fair" sind, d.h. die Augenzahlen 1 bis 6 eines Würfels treten mit gleicher Wahrscheinlichkeit auf.

(**a**) Beschreiben Sie den Ereignisraum in Form von Elementarereignissen.

(**b**) Wie gross ist die Wahrscheinlichkeit eines einzelnen Elementarereignisses?

(**c**) Berechnen Sie die Wahrscheinlichkeit, dass das Ereignis E_1 "Die Augensumme ist 7 " eintritt.

(**d**) Wie gross ist die Wahrscheinlichkeit, dass das Ereignis E_2 "Die Augensumme ist kleiner als 4" eintritt.

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

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

Aufgabe 5

Wo steckt in den folgenden Aussagen der Fehler? Begründen Sie!

(**a**) Bei einer gezinkten Münze wurde festgestellt, dass $P(\text{Kopf}) = 0.32$ und $P(\text{Zahl}) = 0.73$.

(**b**) Die Wahrscheinlichkeit für einen "Sechser" im Zahlenlotto ist $-3 \cdot 10^{-6}$.

(**c**) Bei einer Befragung wurden die Ereignisse

S: Befragte Person ist schwanger.
M: Befragte Person ist männlich.

untersucht. Man findet $P(S) = 0.1$, $P(M) = 0.5$ und $P(S \cup M) = 0.7$

Aufgabe 6

Im Wahrscheinlichkeitsbaum (Abbildung 1) wird für eine zufällig ausgewählte Person zuerst das Merkmal Geschlecht (w = weiblich, m = männlich) und danach das Merkmal Erwerbstätigkeit (E = erwerbstätig, N = nicht erwerbstätig) betrachtet. Aus dem Baum können nun zum Beispiel folgende Wahrscheinlichkeiten herausgelesen werden:

- Wahrscheinlichkeit, dass die Person weiblich ist; $P(w) = 0.514$.
- Wahrscheinlichkeit, dass eine Person erwerbstätig ist, wenn man schon weiss, dass sie männlich ist; $P(E|m) = 0.578$.

	E	N
w	$P(w \cap E) =$	
m		

(**a**) Füllen Sie die obenstehende Tabelle aus:

(**b**) Berechnen Sie die Wahrscheinlichkeit $P(w|E)$.

(**c**) Die Reihenfolge der Merkmale wird nun umgekehrt. Dies führt zum invertierten Wahrscheinlichkeitsbaum gemäss Abbildung 2. Berechnen Sie die gesuchten Wahrscheinlichkeiten.

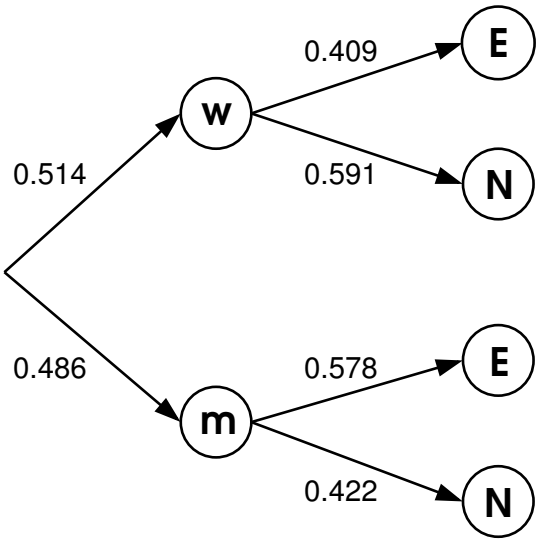


Figure 1: Wahrscheinlichkeitsbaum: Geschlecht vor Erwerbstätigkeit.

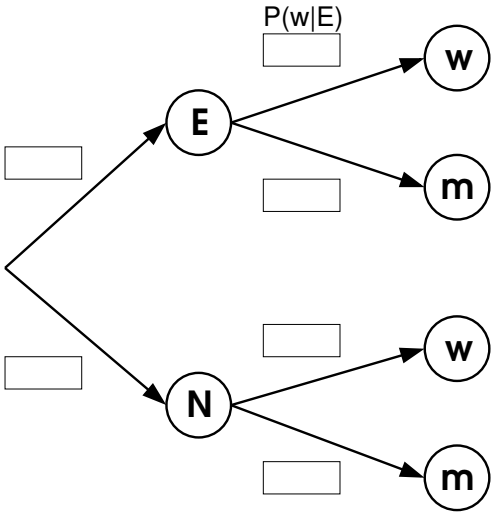


Figure 2: Wahrscheinlichkeitsbaum: Erwerbstätigkeit vor Geschlecht.

Lösungen zu Serie 3

Aufgabe 1

(**a**) Zwischen der empirischen Kovarianz und der Steigung einer Regressionsgeraden existiert folgender Zusammenhang

$$\hat{m} = \frac{s_{xy}}{s_x^2} .$$

Die Steigung der Regressionsgeraden berechnen wir aus folgender Beziehung

$$\hat{a} = \bar{y} - \hat{m}\bar{x} .$$

Daraus folgt für die Steigung der Regressionsgeraden $\hat{m} =$

```
> (12-7.8)/7
```

```
[1] 0.6
```

Somit ergibt sich $s_{xy} = \hat{m} \cdot s_x^2$ zu

```
> 0.6*2.5^2
```

```
[1] 3.75
```

(**b**) Korrelationskoeffizient $r = \frac{s_{xy}}{s_x s_y} =$

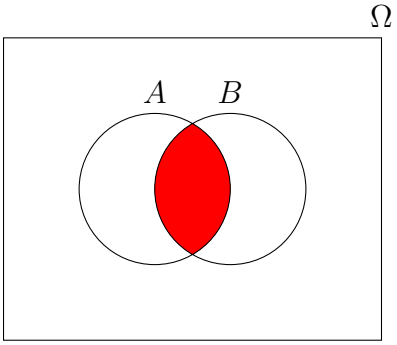
```
> round(3.75/(2.5*1.8),3)
```

```
[1] 0.833
```

Aufgabe 2

```
> A<-3/4
> B<-2/3
```

(**a**)

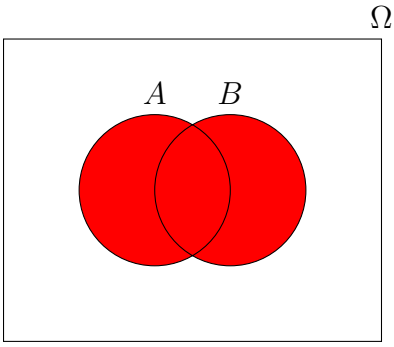


$P(\text{beide Ereignisse}) = P(A \cap B) = P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{2}{3} =$

```
> library(MASS)
> fractions(A*B)
```

```
[1] 1/2
```

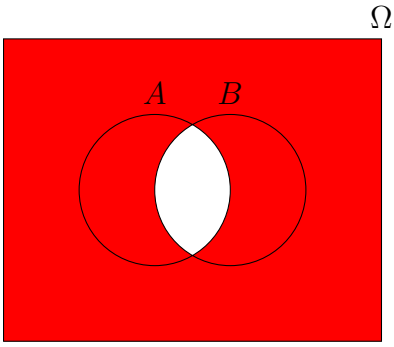
(**b**)



$P(\text{mindestens eines}) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A) * P(B)$
=

```
> library(MASS)
> fractions(A+B-A*B)

[1] 11/12
```

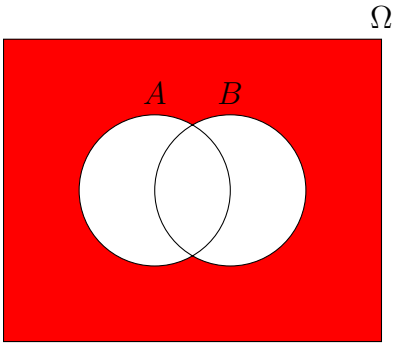
(**c**)



$P(\text{höchstens eines}) = 1 - P(A \cap B) = 1 - P(A) \cdot P(B)$

```
> library(MASS)
> fractions(1-A*B)

[1] 1/2
```

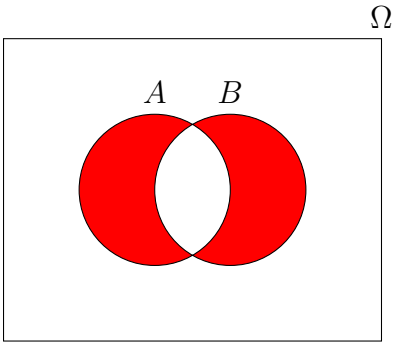
(**d**)



$P(\text{kein Ereignis}) = \overline{P(A \cup B)} = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A) \cdot P(B)) =$

```
> library(MASS)
> fractions(1-(A+B-A*B))
```

```
[1] 1/12
```



(**e**) $P(\text{genau ein Ereignis}) = P(A \cup B) - P(A \cap B) = P(A) + P(B) - 2P(A) \cdot P(B) =$

```
> library(MASS)
> fractions(A+B-2*A*B)

[1] 5/12
```

Aufgabe 3

Wir bezeichnen mit:

$F :=$ Ereignis, dass Feuer ausbricht
$A :=$ Ereignis, dass der Alarm losgeht .

Die Wahrscheinlichkeit, dass ein Feuer ausbricht, ergibt sich zu

$$P(F) = \frac{1}{365} .$$

Die Wahrscheinlichkeit, dass der Alarm losgeht, gegeben es bricht ein Feuer aus, ist

$$P(A|F) = 0.95 .$$

Die Wahrscheinlichkeit, dass es einen Alarm gibt, gegeben dass kein Feuer ausgebrochen ist, lautet

$$P(A|F^c) = 0.01 .$$

(**a**) Die Wahrscheinlichkeit, dass ein Feuer ausgebrochen ist, gegeben es gab einen Alarm, ist

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

Die Wahrscheinlichkeit, dass es einen Alarm gibt, lässt sich mit dem Gesetz der totalen Wahrscheinlichkeit ausdrücken:

$$P(A) = P(A|F) \cdot P(F) + P(A|F^c) \cdot P(F^c) .$$

Also ergibt sich für

$$P(F|A) = \frac{P(A|F) \cdot P(F)}{P(A|F) \cdot P(F) + P(A|F^c) \cdot P(F^c)} = \frac{0.95 \cdot \frac{1}{365}}{0.95 \cdot \frac{1}{365} + 0.01 \cdot (1 - \frac{1}{365})} = 0.207 \,.$$

(**b**) Die Wahrscheinlichkeit, dass kein Feuer ausgebrochen ist, gegeben es gab keinen Alarm, ist

$$P(F^c|A^c) = \frac{P(F^c \cap A^c)}{P(A^c)} = \frac{P(A^c|F^c) \cdot P(F^c)}{1 - P(A)} \,.$$

Die Wahrscheinlichkeit, dass kein Alarm losgeht, gegeben es bricht kein Feuer aus, ist

$$P(A^c|F^c) = 1 - P(A|F^c) \,.$$

Also finden wir

$$P(F^c|A^c) = \frac{(1 - P(A|F^c)) \cdot P(F^c)}{1 - P(A)} = \frac{(1 - 0.01) \cdot \frac{364}{365}}{1 - (0.95 \cdot \frac{1}{365} + 0.01 \cdot (1 - \frac{1}{365}))} = 0.999 \,.$$

Aufgabe 4

(**a**) $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,6)\}$, $|\Omega| = 36$

(**b**) $P(\text{Elementarereignis}) = \frac{1}{\Omega} = \frac{1}{36}$

(**c**) $E_1 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$
Anzahl günstige Fälle: $|E_1| = 6$
Anzahl mögliche Fälle: $\Omega = 36$
$P(E_1) = \frac{|E_1|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$

(**d**) $E_2 = \{(1,1),(2,1),(1,2)\}$;
$P(E_2) = \frac{|E_2|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$

(**e**) $E_3 = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$;
$P(E_3) = \frac{|E_3|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}$

(**f**) Mit dem Additionssatz:

$$\begin{aligned} P(E_2 \cup E_3) &= P(E_2) + P(E_3) - P(E_2 \cap E_3) \\ &= P(E_2) + P(E_3) - P(\{(1,1)\}) \\ &= \frac{3}{36} + \frac{9}{36} - \frac{1}{36} \\ &= \frac{11}{36} \,. \end{aligned}$$

Aufgabe 5

(**a**) Da Zahl und Kopf die möglichen Elementarereignisse sind, müsste die Summe deren Wahrscheinlichkeiten 1 sein. Dies ist hier aber nicht der Fall: $P(\Omega) = P(\text{Zahl}) + P(\text{Kopf}) = 1.05$.(Axiom 2 ist verletzt.)

(**b**) Die genannte Wahrscheinlichkeit ist negativ. (Axiom 1 ist verletzt.)

(**c**) Es gilt $S \cap M = \emptyset$ und darum müsste $P(S) + P(M) = P(S \cup M)$ wegen Axiom 3. Dies ist hieraber nicht erfüllt.

Aufgabe 6

(**a**)

	E	N
w	$P(w \cap E) = 0.514 \cdot 0.409 = 0.210266$	$P(w \cap N) = 0.514 \cdot 0.591 = 0.303774$
m	$P(m \cap E) = 0.486 \cdot 0.578 = 0.280908$	$P(m \cap N) = 0.486 \cdot 0.422 = 0.205092$

(**b**) $P(E) = P(w \cap E) + P(m \cap E) = 0.210266 + 0.280908 = 0.491134 \approx 0.491$
$P(w|E) = \frac{P(w \cap E)}{P(E)} = \frac{0.210266}{0.491134} \approx 0.428$

(**c**) $P(m|E) = \frac{P(m \cap E)}{P(E)} = \frac{0.280908}{0.491134} \approx 0.572$
$P(N) = P(w \cap N) + P(m \cap N) = 0.303774 + 0.205092 = 0.508866 \approx 0.509$
$P(w|N) = \frac{P(w \cap N)}{P(N)} = \frac{0.303774}{0.508866} \approx 0.597$
$P(m|N) = \frac{P(m \cap N)}{P(N)} = \frac{0.205092}{0.508866} \approx 0.403$

Serie 4

Vorbesprechung: 13/14. März 2013

Aufgabe 1

Ein Stand auf einem Volksfest bietet ein Würfelspiel an. Man wirft zwei sechsseitige Würfel. Je nach Ausgang des Wurfs muss man Geld bezahlen oder man erhält Geld. Hier sind die Regeln des Spiels:

(1) Bei einem Pasch (also $(1,1)$, $(2,2)$, etc.) gewinnt der Spieler 10 SFr (Gewinn 10 SFr).

(2) Bei $(1,2)$ oder $(2,1)$ gewinnt der Spieler 20 SFr (Gewinn 20 SFr).

(3) Bei allen anderen Ergebnissen verliert der Spieler 4 SFr (Gewinn -4 SFr).

(**a**) Sei X die Zufallsvariable, die den Gewinn des Spielers nach einem Wurf angibt. Bestimmen Sie die Wahrscheinlichkeitsverteilung.

(**b**) (Knobelaufgabe) Würden Sie dieses Spiel spielen? Überlegen Sie sich eine Möglichkeit, wie man mit einer Zahl angeben kann, ob sich das Spiel lohnt oder nicht.

Aufgabe 2

Bei einer Untersuchung werden Wasserproben (10 ml) auf Verunreinigungen untersucht. Da nur 2 Prozent aller Proben verunreinigt sind, wird vorgeschlagen, von 10 Einzelproben jeweils die Hälfte (5 ml) der Proben zu einer Sammelprobe (50 ml) zusammenzumischen und zunächst nur die Sammelprobe zu untersuchen. Wird in der Sammelprobe keine Verunreinigung festgestellt, so ist die Untersuchung für die 10 Einzeluntersuchungen beendet. Im anderen Fall werden alle 10 übriggebliebenen Hälften in 10 Einzeluntersuchungen geprüft.

(**a**) Wie gross ist die Wahrscheinlichkeit, in der Sammelprobe keine Verunreinigung zu finden (unter der Annahme, dass die Einzelproben unabhängig voneinander sind)?

(**b**) Sei die Zufallsvariable Y die Gesamtzahl benötigter Analysen. Welche Werte kann Y annehmen, und mit welchen Wahrscheinlichkeiten treten sie auf?

(**c**) Wieviele Analysen werden im Durchschnitt für die gesamte Untersuchung benötigt (d.h. wie gross ist $\mathrm{E}[Y]$)? Wieviele Analysen werden durch die Bildung von Sammelproben "im Durchschnitt" eingespart?

Aufgabe 3

Ein Hersteller von Reagenzgläsern möchte sicherstellen, dass eine grosse Lieferung weniger als 10% minderwertige Gläser enthält (Qualitätsstufe A). Zwecks Qualitätssicherung entnimmt er der Lieferung eine zufällige Stichprobe im Umfang von fünfzig Gläsern. Es stellt sich heraus, dass von diesen fünfzig Gläsern drei minderwertig sind.
Für den Hersteller stellt sich nun das Problem, aufgrund der gezogenen Stichprobe zu entscheiden, ob er tatsächlich beruhigt davon ausgehen kann, dass die ganze Lieferung einen Anteil minderwertiger Gläser $< 10\%$ enthält oder ob es als plausibel gelten kann, dass er in der Stichprobe "rein zufällig" einen Anteil minderwertiger Gläser unter 10% erwischt hat, obwohl die ganze Lieferung in Tat und Wahrheit einen Anteil minderwertiger Gläser von 10% oder mehr aufweist.

(**a**) Welches Modell bzw. welche Verteilung beschreibt die Anzahl minderwertiger Gläser in der Stichprobe unter der Annahme, dass die einzelnen Gläser voneinander unabhängig sind?

(**b**) Wie gross ist die Wahrscheinlichkeit,dass die gezogene Stichprobe genau drei minderwertige Gläser enthält, wenn der wahre Anteil minderwertiger Gläser in der Lieferung 10% beträgt?

(**c**) Wie gross ist die Wahrscheinlichkeit, dass die gezogene Stichprobe höchstens drei minderwertige Gläser enthält, wenn die Lieferung einen Anteil von 10% minderwertiger Gläser enthält?

(**d**) Formulieren Sie in wenigen Worten das "Problem" des Herstellers!

Aufgabe 4

Verwende **R** um folgende Grössen zu berechnen.
Es sei $X \sim \text{Bin}(50, 0.2)$.

(**a**) $P(X = 10)$

(**b**) $P(X \leq 5)$

(**c**) $P(X \geq 15)$

(**d**) Finde c sodass $P(X \leq c) \approx 0.99$

Aufgabe 5

Binomialkoeffizienten spielen in der abzählenden Kombinatorik eine zentrale Rolle, denn $\binom{n}{k}$ ist die Anzahl der Möglichkeiten, aus einer Menge mit n Elementen k Elemente auszuwählen, wobei die Reihenfolge der ausgewählten Elemente nicht berücksichtigt wird. Anschaulich lässt sich das so erklären: Man berechne mit $n!$ alle möglichen Vertauschungen, suche sich k "Felder" aus (beispielsweise 6 beim Lotto) und frage sich, wie viele Möglichkeiten es gibt, diese Felder zu besetzen (beim Lotto mit 49 Zahlen). Da es keine Rolle spielt, welches "Ereignis" sich auf welchem Feld ereignet hat, dividiert man alle unter diesen k Elementen möglichen Vertauschungen mit $k!$ heraus. Da es auch keine Rolle spielt, wie die Anordnung auf den uninteressanten Feldern aussieht, dividiert man mit $(n-k)!$ auch diese Vertauschungen heraus.

(**a**) Wie gross ist die Wahrscheinlichkeit, beim Lotto sechs richtige Zahlen zu ziehen?

(**b**) Das Programm eines Computers stellt für die Darstellung einer Zahl 15 Zeichenplätze (Bits) zur Verfügung, die mit 0 oder 1 belegt werden. Wieviele solcher Zahlen mit 7 Ziffern 1 gibt es? Wieviele Zahlen können insgesamt dargestellt werden?

Aufgabe 1

(**a**)

k	-4	10	20
$P(X=k)$	28/36=7/9	6/36=1/6	2/36=1/18

(**b**) Eine sinnvolle Kennzahl ist der sogenannte Erwartungswert, das heisst, man gewichtet die Gewinne mit ihrer Eintrittswahrscheinlichkeit und addiert sie:

$$E = -4 \cdot \frac{7}{9} + 10 \cdot \frac{1}{6} + 20 \cdot \frac{1}{18} = -\frac{1}{3}$$

Man erwartet im Mittel also einen negativen Gewinn; das Spiel ist nicht fair!

Aufgabe 2

(**a**) Sei X die Anzahl der verunreinigten Einzelproben in einer Sammelprobe. Die Erfolgswahrscheinlichkeit p, dass eine Einzelprobe verunreinigt ist, beträgt 0.02. Unter der Annahme, dass die Einzelproben voneinander unabhängig sind, gilt $X \sim \text{Bin}(n = 10,\ p = 0.02)$.

Die Wahrscheinlichkeit, in der Sammelprobe keine Verunreinigung zu finden, ist gegeben durch

$$P[X = 0] = \binom{10}{0} 0.02^0 \cdot 0.98^{10} = 0.98^{10} = 0.817$$

Anderer Lösungsweg: Jede einzelne Probe ist unabhängig von den anderen Proben mit 98% Wahrscheinlichkeit sauber. Also gilt (**Multiplikationssatz fuï unabhängige Ereignisse**)

$$P[\text{alle Proben sauber}] = \prod_{i=1}^{10} P[i - \text{te Probe sauber}] = 0.98^{10} = 0.817$$

(**b**) Die Zufallsvariable Y kann nur die Werte 1 oder 11 annehmen, denn:

- falls alle Proben sauber sind, ist man nach einer Untersuchung fertig: $Y = 1$
- sonst muss man jede Probe einzeln untersuchen (man darf nicht stoppen, wenn man die erste verunreinigte Probe gefunden hat, denn es könnte ja noch mehrere geben!), also $Y = 11$.

Folglich

$$\begin{cases} P[Y=1] = & P[\text{alle Proben sauber}] = 0.817 \\ P[Y=11] = & 1 - P[Y=1] = 0.183 \end{cases}$$

(**c**) Die durchschnittliche Anzahl Analysen pro Sammelprobe ist gegeben durch den Erwartungswert der Zufallsvariablen Y :

$$\begin{aligned} \mathbf{E}[Y] &= \sum_{k=0}^{\infty} kP[Y=k] = 1 \cdot P[Y=1] + 11 \cdot P[Y=11] \\ &= 1 \cdot 0.817 + 11 \cdot 0.183 \\ &= 2.83 \end{aligned}$$

"Im Durchschnitt" spart man also $10 - 2.83 = 7.17 \approx 7$ Untersuchungen ein.

Aufgabe 3

(**a**) Es sei X die Anzahl minderwertiger Gläser in der Stichprobe. Unter der Annahme der Unabhängigkeit und unter der Voraussetzung, dass alle Gläser gleich hergestellt wurden, so dass jedes Glas die gleiche Chance hat, von minderwertiger Qualität zu sein, ist X binomialverteilt.

(**b**) Unter den gegebenen Bedingungen ist $X \sim \text{Bin}(n, \pi)$ mit $n = 50$ und $\pi = 0.1$. Die gesuchte Wahrscheinlichkeit beträgt somit

$$P[X=3] = \binom{50}{3} 0.1^3 \cdot 0.9^{47} = 0.139\,.$$

(**c**) Wenn die Lieferung einen Anteil von 10% minderwertiger Gläser enthält, ist die Anzahl X minderwertiger Gläser in der Stichprobe $\text{Bin}(n, \pi)$-verteilt mit $n = 50$ und $\pi = 0.1$. Für die gesuchte Wahrscheinlichkeit ergibt sich

$$P[X \leq 3] = \sum_{k=0}^{3} \binom{50}{k} (0.1)^k \cdot (0.9)^{50-k} = 0.25\,.$$

(**d**) Für den Hersteller stellt sich das Problem, eine kritische Grenze des Anteils minderwertiger Gläser in einer Stichprobe festzulegen. Er will, falls die Lieferung qualitativ schlecht ist, dass die kritische Grenze mit hoher Wahrscheinlichkeit überschritten wird. Auf der anderen Seite will er, dass die Grenze möglichst nicht überschritten wird, falls die Qualität in Ordnung ist.

Kommentar:

Wie Teilaufgabe c) illustriert, kann ihm der Zufall dabei ziemlich in die Quere kommen: Mit einer Wahrscheinlichkeit von ca. 25% wird er aus einer Lieferung, die 10% minderwertige Gläser enthält, eine Stichprobe ($n = 50$) ziehen, die weniger als drei minderwertige

Gläser enthält. Drei minderwertige Gläser entsprechen in der Stichprobe einem Anteil von 6%.

Dies zeigt, dass er aufgrund einer Stichprobe vom Umfang $n = 50$, die 6% oder weniger minderwertige Gläser enthält, nicht unbedingt schliessen kann, dass die ganze Lieferung weniger als 10% minderwertige Gläser enthält. Im Mittel wird er ja jedes vierte Mal rein zufällig eine solche Stichprobe ziehen und eine falsche Entscheidung treffen! Die Frage ist nun, wie klein der Anteil minderwertiger Gläser in der gezogenen Stichprobe sein muss, damit der Hersteller davon ausgehen kann, dass die ganze Lieferung die Qualitätsstufe A erfüllt. Diese Frage führt auf ein statistisches Testproblem.

Aufgabe 4

(**a**) $P(X = 10) =$

```
> dbinom(x=10, size=50, prob=0.2)

[1] 0.139819
```

(**b**) $P(X \leq 5) =$

```
> pbinom(q=5, size=50, prob=0.2)

[1] 0.04802722
```

(**c**) $P(X \geq 15) =$

```
> 1-pbinom(q=14, size=50, prob=0.2)

[1] 0.06072208
```

(**d**)
```
> pbinom(q=0:20, size=50, prob=0.2)

 [1] 1.427248e-05 1.926784e-04 1.285415e-03 5.656361e-03 1.849602e-02
 [6] 4.802722e-02 1.033982e-01 1.904098e-01 3.073316e-01 4.437404e-01
[11] 5.835594e-01 7.106676e-01 8.139430e-01 8.894135e-01 9.392779e-01
[16] 9.691966e-01 9.855583e-01 9.937392e-01 9.974888e-01 9.990676e-01
[21] 9.996793e-01
```

Wähle die Position, die etwa 0.99 gibt $\Rightarrow$ 17

Aufgabe 5

(**a**) $P(\text{sechs richtige Zahlen beim Lotto}) = \frac{\binom{6}{6}\binom{49-6}{6-6}}{\binom{49}{6}} = \frac{1}{\binom{49}{6}} =$

```
> 1/choose(49,6)

[1] 7.151124e-08
```

(**b**) #Zahlen mit 7 Ziffern $1 = \binom{15}{7} =$

```
> choose(15,7)

[1] 6435
```

#Zahlen, die insgesamt dargestellt werden können $= 2^{15} =$

```
> 2^(15)

[1] 32768
```

Vorbesprechung: 20/21. März 2013

Aufgabe 1

Verwende **R** um folgende Grössen zu berechnen.
Es sei $X \sim \text{Poisson}(200)$ die Zufallsvariable, die die Anzahl Unfälle in einem Jahr beschreibt.

(**a**) Wie gross ist die Wahrscheinlichkeit, dass in einem Jahr genau 200 Unfälle passieren?

(**b**) Wie gross ist die Wahrscheinlichkeit, dass in einem Jahr höchstens 210 Unfälle passieren?

(**c**) Wie gross ist die Wahrscheinlichkeit, dass in einem Jahr zwischen 190 und 210 Unfälle passieren (beide Grenzen eingeschlossen)?

Aufgabe 2

Die Zufallsvariable, die die Anzahl eingehender Telefonanrufe in einer Telefonzentrale innerhalb von 10 Minuten beschreibt, nennen wir X. Sie folgt einer Poissonverteilung mit Erwartungswert $\lambda = 2$, d.h. $X \sim \text{Poisson}(\lambda)$.

(**a**) Wie gross ist die Wahrscheinlichkeit, dass es in einer bestimmten 10-Minuten-Periode keinen einzigen Anruf gibt?

(**b**) Wie gross ist die Wahrscheinlichkeit, dass es nicht mehr als drei Telefonanrufe in einer bestimmten 10-Minuten-Periode gibt?

(**c**) Wie gross ist die Wahrscheinlichkeit, dass es mehr als drei Telefonanrufe in einer bestimmten 10-Minuten-Periode gibt?

(**d**) Angenommen, die Anzahl Anrufe in einer 10-Minuten-Periode ist von der Anzahl Anrufe in einer anderen 10-Minuten-Periode unabhängig. Die Zufallsvariable, die die Anzahl Anrufe in einer Stunde beschreibt bezeichnen wir mit Y. Welcher Verteilung folgt Y?

Aufgabe 3

In der Vorlesung haben wir gesehen, wie man die Erfolgswahrscheinlichkeit π einer Binomialverteilung mit der Maximum-Likelihood-Methode schätzen kann, wenn man die Anzahl Versuche und die Anzahl Gewinne kennt. In dieser Aufgabe kombinieren wir mehrere solcher Beobachtungen zu einer Schätzung.
Angenommen Sie gehen über den Jahrmarkt und kaufen bei einer Losbude 30 Lose. Unter den 30 Losen sind 2 Gewinne. Am nächsten Tag erzählt Ihnen Ihr Studienkollege, dass er am Vorabend bei der gleichen Losbude 50 Lose gekauft hat und darunter 4 Gewinne hatte. Wie kombinieren Sie die beiden Ergebnisse, um mit der Maximum-Likelihood-Methode eine möglichst gute Schätzung der Erfolgswahrscheinlichkeit zu erhalten?

(**a**) Sei X_1 die Zufallsvariable, die die Anzahl Gewinne unter 30 Losen beschreibt ("Ihre" Gewinne). Wenn wir annehmen, dass jedes Los unabhängig von jedem anderen Los ein Gewinn oder eine Niete ist, dann folgt X einer Binomialverteilung mit $n_1 = 30$ und unbekanntem Erfolgsparameter π. Abgekürzt schreiben wir: $X_1 \sim \text{Bin}(n_1, \pi)$. Analog sei X_2 die Zufallsvariable, die die Gewinne Ihres Kollegen beschreibt: $X_2 \sim \text{Bin}(n_2, \pi)$ mit $n_2 = 50$ und dem gleichen Wert für die Erfolgswahrscheinlichkeit wie bei X_1. Angenommen, die Anzahl Gewinne, die Sie gezogen haben, ist unabhängig von der Anzahl Gewinne, die Ihr Kollege gezogen hat. Wie lässt sich dann $P(X_1 = x_1 \cap X_2 = x_2)$ schreiben?

(**b**) Wie lässt sich $\log(P(X_1 = x_1 \cap X_2 = x_2))$ schreiben? Versuchen Sie diesen Term in eine Summe mit mehreren Termen umzuschreiben. Welche Terme hängen von π ab und welche nicht?

(**c**) Der Maximum-Likelihood-Schätzer für π ist derjenige Zahlenwert, der, wenn man ihn anstelle von π einsetzt, den grösstmöglichen Wert für $\log(P(X_1 = x_1 \cap X_2 = x_2))$ (oder $P(X_1 = x_1 \cap X_2 = x_2)$; das ist egal, weil die Funktion log monoton ist) liefert. Finden Sie durch Ableiten und gleich Null setzen den Wert von π in Abhängigkeit von n_1, n_2, x_1 und x_2, der $\log(P(X_1 = x_1 \cap X_2 = x_2))$ maximiert.

Aufgabe 4

Das Pharmaunternehmen Life Co. hat ein neues Medikament zur Bekämpfung von ADHS entwickelt. Um die Wirksamkeit festzustellen wurde das Medikament mit $n = 10$ Patienten getestet. Die derzeitige Standardmethode zeigt bei 30% der behandelten Patienten eine Wirkung.

(**a**) Angenommen das neue Medikament ist genauso wirksam wie die Standardmethode, wie gross ist die Wahrscheinlichkeit, dass die Behandlung bei genau 2 Patienten eine Wirkung zeigt? Wie gross ist die Wahrscheinlichkeit, dass sie bei höchstens 2 Patienten eine Wirkung zeigt?

(**b**) Die Behandlung mit dem neuen Medikament war bei 4 Patienten erfolgreich. Führen Sie einen einseitigen Hypothesentest durch um festzustellen ob das neue Medikament wirksamer ist als die Standardmethode (bei einem Signifikanzniveau von 5%). Geben Sie explizit alle Schritte an.

(**c**) Wie ist die Macht eines Hypothesentests definiert? Geben Sie die Macht an für den Test H_0: $\pi = 0.3$ vs. $H_A : \pi = 0.6$ (π ist die Wirksamkeit).

Aufgabe 5

("Qualitätskontrolle von Reagenzgläsern") Ein Hersteller von Reagenzgläsern garantiert seinen Kunden, dass der Anteil minderwertiger Gläser kleiner als 10% ist. Zwecks Qualitätssicherung entnimmt er einer grossen Lieferung eine zufällige Stichprobe im Umfang von fünfzig Gläsern. Es stellt sich heraus, dass von diesen fünfzig Gläsern 3 minderwertig sind. Für den Hersteller ergibt sich nun das Problem: Kann er aufgrund der gezogenen Stichprobe tatsächlich beruhigt davon ausgehen, dass der Anteil minderwertiger Gläser in der ganze Lieferung wirklich kleiner als 10% ist. Führen Sie einen Hypothesentest mit dem Signifikanzniveau 5% durch. Lösen Sie damit das Problem des Herstellers.

Aufgabe 6

Wir betrachten nochmals das Beispiel "Qualitätskontrolle von Reagenzgläsern". Man nimmt jeweils eine Stichprobe von 50 Gläsern und zählt die Anzahl minderwertiger Exemplare (X). Das Testproblem bestand im Testen der Nullhypothese $H_0 : \pi = 0.1$ gegen die Alternative H_A: $\pi < 0.1$. Auf dem 5%-Niveau resultierte ein Verwerfungsbereich von $K = \{X \leq 1\}$.

(**a**) Bestimmen Sie die Wahrscheinlichkeit, dass der Test verwirft, wenn in Tat und Wahrheit $\pi = 0.075$ gilt. Dies heisst die Macht des Tests unter der Alternative $\pi = 0.075$. Verwenden Sie hierzu die R-Funktion **pbinom**.

Hinweise:

pbinom(q, size, prob) ist die kumulative Verteilungsfunktion einer Binomialverteilung an der Stelle **q**. Das Argument **size** ist die Stichprobengrösse und das Argument **prob** die Erfolgswahrscheinlichkeit.

(**b**) Angenommen man nimmt nun Stichproben der Grösse $n = 150$. Bestimmen Sie zuerst den Verwerfungsbereich und anschliessend die Macht wie oben.

Hinweise:

Mit **pbinom(0:50, size = ..., prob = ...)** können Sie die Verteilungsfunktion unter der Nullhypothse an den Stellen $0, \ldots, 50$ auswerten. Bestimmen Sie dann daraus den Verwerfungsbereich.

Lösungen zu Serie 5

Vorbesprechung:
2013

Aufgabe 1

(**a**) P(genau 200 Unfälle)=

```
> dpois(x=200, lambda=200)

[1] 0.02819773
```

(**b**) P(höchstens 210 Unfälle) =

```
> ppois(q=210, lambda=200)

[1] 0.772708
```

(**c**) P(zwischen 190 und 210 Unfälle) =

```
> ppois(q=210, lambda=200) - ppois(q=189, lambda=200)

[1] 0.5422097
```

Aufgabe 2

Da $X \sim \text{Poisson}(\lambda)$ mit $\lambda = 2$ gilt: $P(X = x) = \exp(-2)\frac{2^x}{x!}$

(**a**) $P(X = 0) = \exp(-2)\frac{2^0}{0!} = \exp(-2)\frac{1}{1} \approx 0.135$

(**b**)

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.135 + 0.271 + 0.271 + 0.180 \\ &\approx 0.857 \end{aligned}$$

(**c**) $P(X > 3) = 1 - P(X \leq 3) = 1 - 0.857 \approx 0.143$

(**d**) Nach Kapitel 3.7.2 folgt: $Y \sim \text{Poisson}(6 \cdot \lambda) = \text{Poisson}(12)$

Aufgabe 3

Es gilt: $X_1 \sim \text{Bin}(n_1, \pi)$ und $X_2 \sim \text{Bin}(n_2, \pi)$; X_1 und X_2 sind unabhängig.

(**a**) Da X_1 und X_2 unabhängig sind, gilt:

$$P(X_1 = x_1 \cap X_2 = x_2) = P(X_1 = x_1) \cdot P(X_2 = x_2)\,,$$

wobei $P(X_1 = x_1) = \binom{n_1}{x_1}\pi^{x_1}(1-\pi)^{n_1-x_1}$ und $P(X_2 = x_2) = \binom{n_2}{x_2}\pi^{x_2}(1-\pi)^{n_2-x_2}$.

(**b**)

$$\begin{aligned}
\log(P(X_1 = x_1 \cap X_2 = x_2)) &= \log(P(X_1 = x_1) \cdot P(X_2 = x_2)) \\
&= \log(P(X_1 = x_1)) + \log(P(X_2 = x_2)) \\
&= \log\left(\binom{n_1}{x_1}(1-\pi)^{n_1-x_1}\right) + \log\left(\binom{n_2}{x_2}\pi^{x_2}(1-\pi)^{n_2-x_2}\right) \\
&= \log\left(\binom{n_1}{x_1}\right) + x_1 \cdot \log(\pi) + (n_1 - x_1) \cdot \log(1-\pi) \\
&\quad + \log\left(\binom{n_2}{x_2}\right) + x_2 \cdot \log(\pi) + (n_2 - x_2) \cdot \log(1-\pi)\,.
\end{aligned}$$

(**c**)

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\pi}\left\{\log\left(\binom{n_1}{x_1}\right) + x_1 \cdot \log(\pi) + (n_1 - x_1) \cdot \log(1-\pi)\right. \\
&\left. + \log\left(\binom{n_2}{x_2}\right) + x_2 \cdot \log(\pi) + (n_2 - x_2) \cdot \log(1-\pi)\right\} \\
&= \frac{x_1}{\pi} - (n_1 - x_1) \cdot \frac{1}{1-\pi} + \frac{x_2}{\pi} - (n_2 - x_2) \cdot \frac{1}{1-\pi} \\
&= \frac{x_1 + x_2}{\pi} - \frac{((n_1 + n_2) - (x_1 + x_2))}{1-\pi}\,.
\end{aligned}$$

Wenn wir diesen Ausdruck gleich Null setzen und nach π auflösen, erhalten wir:

$$\pi = \frac{x_1 + x_2}{n_1 + n_2}\,.$$

Das Ergebnis ist also identisch mit dem Ergebnis, das wir erhalten hätten, wenn eine Person $30 + 50 = 80$ Lose gezogen hätte und dabei $2 + 4 = 6$ Gewinne gezogen hätte (da $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, \pi)$).
Das hier gesehene Prinzip, einen Parameter zu schätzen, indem man mehrere unabhängige Beobachtungen kombiniert, ist die mit Abstand häufigste Schätzmethode in der Statistik.

Aufgabe 4

(**a**)

$$P(X = 2) = \binom{10}{2} 0.3^2 0.7^8 = 0.23$$

$$P(X \leq 2) = P(X = 0) + P(x = 1) + P(X = 2) = 0.7^{10} + \binom{10}{1} 0.3^1 0.7^9 + \binom{10}{2} 0.3^2 0.7^8 = 0.38$$

(**b**) **1.** Modell: X ist die Anzahl erfolgreich behandelter Patienten, $X \sim \text{Bin}(10, \pi)$.

2. Die Nullhypothese ist $H_0 : \pi = 0.3$, die Alternative ist $H_A : \pi > 0.3$.

3. Die Teststatistik ist $T : P(T = t|H_0) = \binom{10}{t} 0.3^t 0.7^{10-t}$

4. Das Signifikanzniveau ist $\alpha = 0.05$.

5. Verwerfungsbereich:

	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$
$P(T \geq t)$	0.1503	0.0473	0.0106	0.0016	0.0001	5.9×10^{-6}

Daher ist der Verwerfungsbereich $K = \{6, 7, 8, 9, 10\}$.

6. Testentscheid: Da $4 \notin K$ wird H_0 nicht verworfen. Eine erhöhte Wirksamkeit des neuen Medikaments kann nicht nachgewiesen werden.

(**c**) Die Macht eines Tests ist die Wahrscheinlichkeit, dass die Nullhypothese verworfen wird, wenn die Alternative stimmt: $P(T \in K|H_A)$. (Alternativ: Macht $= 1 - P$(Fehler 2. Art) $= 1 - P(T \notin K|H_A)$)

Im konkreten Fall:

$$\text{Macht} = \binom{10}{6} 0.6^6 0.4^4 + \binom{10}{7} 0.6^7 0.4^3 + \binom{10}{8} 0.6^8 0.4^2 + \binom{10}{9} 0.6^9 0.4 + 0.6^1 0 = 0.6331\,.$$

Aufgabe 5

1. Modell: X: Anzahl defekter Reagenzgläser in einer Stichprobe aus 50 Reagenzgläsern. $X \sim \text{Bin}(50, \pi)$.

2. Nullhypothese: $H_0 : \pi = 0.1$
Alternative: $H_A : \pi < 0.1$

3. Teststatistik: T: Anzahl defekter Reagenzgläser in einer Stichprobe aus 50 Reagenzgläsern.
Verteilung der Teststatistik unter $H_0 : T \sim \text{Bin}(50, 0.1)$

4. Signifikanzniveau: $\alpha = 0.05$

5. **Verwerfungsbereich**: Falls H_0 stimmt, gilt:

$$\begin{aligned} P(T=0) &= 0.0052 & P(T \leq 0) &= 0.0052 \\ P(T=1) &= 0.0286 & P(T \leq 1) &= 0.0338 \\ P(T=2) &= 0.0779 & P(T \leq 2) &= 0.1117 \end{aligned}$$

   Der Verwerfungsbereich K für ein Signifikanzniveau von 5% ist also gegeben durch $K = \{0, 1\}$.

6. **Testentscheid**: Der beobachtete Wert der Teststatistik ist $t = 3$. Der beobachtete Wert der Teststatistik ($t = 3$) liegt nicht im Verwerfungsbereich der Teststatistik ($K = \{0, 1\}$). Die Nullhypothese kann daher auf dem 5% Signifikanzniveau nicht verworfen werden. Es kann also durchaus sein, dass der Anteil minderwertiger Gläser in der ganzen Lieferung 10% ist. Der Hersteller sollte also seine Lieferung nicht losschicken, sondern genauer untersuchen.

Aufgabe 6

(**a**) Wir müssen $P[X \leq 1]$ berechnen im Falle von $\pi = 0.075$.

```
> pbinom(1,50,0.075)

[1] 0.1025006
```

Wenn die Lieferung also nur 7.5% defekte Gläser enthält, so können wir dies mit unserem Test (mit 50 Proben) nur in ca. 10% der Fälle nachweisen!

(**b**) Mit **pbinom(0:50, 150, 0.1)** sehen wir, dass der Verwerfungsbereich $K = \{T \leq 8\}$ ist. Wir erhalten

```
> pbinom(8,150,0.075)

[1] 0.2000952
```

Dank der grösseren Stichprobe ist auch die Macht grösser geworden.

Vorbesprechung: 3/4. April 2013

Aufgabe 1

In einer medizinischen Pilotstudie sprachen 5 von 16 Patienten auf eine *neue* Behandlung an. Die Ansprechwahrscheinlichkeit auf die *Standardbehandlung* wird mit 15% angegeben. Ist die neue Behandlung der Standardbehandlung überlegen?

(**a**) Formulieren Sie die Null- und die Alternativhypothese und führen Sie den Test auf dem 5%-Niveau durch mit Hilfe der R-Funktionen **pbinom** oder **dbinom**.

(**b**) Bei welchem Wert des Signifikanzniveaus wechselt der Testentscheid von "Beibehalten" zu "Verwerfen"?

(**c**) Betrachten Sie den Test von Aufgabe a). Wie gross ist die Wahrscheinlichkeit, dass die Nullhypothese verworfen wird, wenn die Ansprechwahrscheinlichkeit auf die neue Behandlung 30% ist? Verwenden Sie wieder die Funktionen **pbinom** bzw. **dbinom**.

Aufgabe 2

Die low-cost Fluggesellschaft 'Air-Patatrack' verkauft (wie auch viele andere Fluggesellschaften) mehr Flugtickets pro Flug als Sitze im betreffenden Flugzeug vorhanden sind (so genanntes 'overbooking'). Grund für dieses Vorgehen ist, dass Kunden oft kurzfristig auf die Reise verzichten.
Air-Patatrack schätzt, dass 90% der gebuchten Tickets benutzt werden.

(a) Wie gross ist die Wahrscheinlichkeit, dass von vier Personen, genau eine die Reise nicht antritt?

Für den Flug Zürich-Agno benutzt die Gesellschaft ein 'Beechcraft C12-J' mit 26 Passagierplätzen. Für den nächsten Flug sind 28 Plätze gebucht.

(b) Welche Verteilung besitzt die Anzahl der Personen, die den Flug antreten möchten? Berechnen Sie den Erwartungswert und die Varianz dieser Verteilung.

(c) Wie gross ist die Wahrscheinlichkeit, dass nicht alle Passagiere Platz im Flugzeug finden?

Für die Flüge Zürich-Johannesburg benutzt die Air-Patatrack einen Airbus A380 mit 853 Sitzplätze. Für solche Flüge werden 890 Tickets verkauft. Von 890 Personen, die den Flug am 05.08 reserviert haben, sind nur 875 am Flughafen erschienen. Ist die Annahme, dass 90% der gebuchten Tickets benutzt werden angesichts dieser Daten plausibel?

(d) Führen Sie einen zweiseitigen Test auf dem 5% Niveau durch. Geben Sie die Null- und die Alternativhypothese, die Teststatistik sowie den Testentscheid an. (Benutzen Sie die Normalapproximation).

Aufgabe 3

An einer Losbude kaufen Sie 50 Lose. Unter den Losen sind 7 Gewinne.

(**a**) Was ist ein approximatives 95% Vertrauensintervall (verwenden Sie die Normalapproximation) für die Gewinnwahrscheinlichkeit?

(**b**) Lesen Sie die Hilfe der Funktion binom.test und berechnen Sie damit das 95%-Vertrauensintervall.

Aufgabe 4

In dieser Aufgabe wollen wir in einer Simulation die Überdeckungswahrscheinlichkeit von Vertrauensintervallen untersuchen.
Wir betrachten hierzu eine Binomialverteilung mit $n = 50$. Wählen Sie selber eine Erfolgswahrscheinlichkeit π.

(**a**) Simulieren Sie 20 Realisationen von obiger Binomialverteilung und bestimmen Sie für jede Realisation das 95%-Vertrauensintervall für die Erfolgswahrscheinlichkeit π. Wie oft erwarten Sie, dass der wahre Wert im Vertrauensintervall liegt? Wie oft liegt er tatsächlich drin?

R-Hinweise:

```
## 20 Werte simulieren
p <- ...
x <- rbinom(20, 50, p)
## Grenzen der Intervalle in Matrix speichern
## 1. Spalte ist untere Grenze, 2. Spalte obere
confint.bound  <- matrix(0, nrow = 20, ncol = 2)
contains.truth <- logical(20)
## Alle 20 Faelle untersuchen und Grenzen speichern
for(i in 1:20){
test <- binom.test(...) ## Setzen Sie die richtigen Argumente!
confint.bound[i,] <- test$conf.int
contains.truth[i] <-
  (p >= confint.bound[i,1]) & (p <= confint.bound[i,2])
}
sum(contains.truth)
```

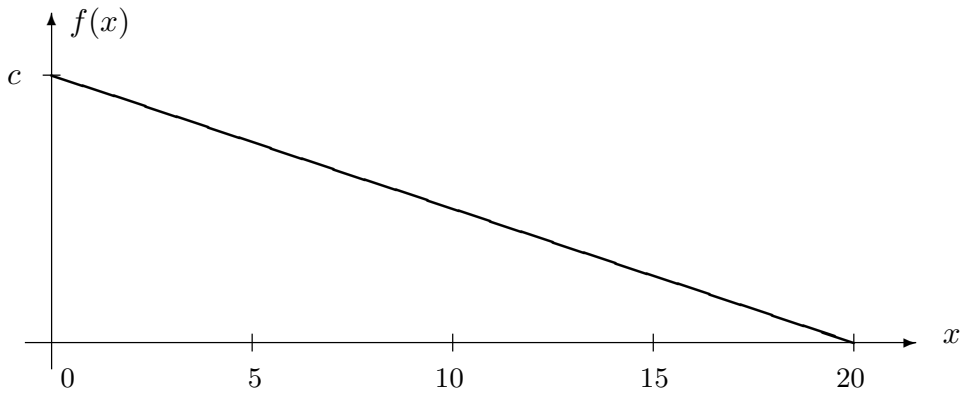
(**b**) Stellen Sie das Resultat aus (a) geeignet dar.

R-Hinweise:

```
## Relative Haeufigkeiten plotten
plot(x / 50, 1:20, xlim = c(0, 1), xlab = "Probability",
  ylab = "Simulation Number")
## Vertrauensintervalle als Liniensegemente plotten
for(i in 1:20){
  segments(confint.bound[i,1], i, confint.bound[i,2], i)
}
## Wahrer Wert als vertikale Linie einzeichnen
abline(v = ...)
```

Aufgabe 5

In der Stadt Zürich gibt es bekanntlich viele Baustellen. Die Dauer X der Arbeiten bei einer Baustelle liege zwischen 0 und 20 Wochen. Die Dichte $f(x)$ habe die folgende Form.



(**a**) Begründen Sie, warum $c = 0.1$ ist und schreibe die Dichte $f(x)$ explizit auf.

(**b**) Berechnen Sie die Wahrscheinlichkeit, dass die Bauzeit X weniger als

(i) 5

(ii) 10

Wochen beträgt.

(**c**) Skizzieren Sie die kumulative Verteilungsfunktion.

(**d**) Berechnen Sie den Erwartungswert, den Median und die Standardabweichung der Dauer X.

(**e**) $K = 40'000 \cdot \sqrt{X}$ entspreche dem Betrag in Franken, den die Arbeiten bei einer Baustelle kosten. Wie gross ist die Wahrscheinlichkeit, dass die Arbeiten bei einer Baustelle höchstens 120'000.- Fr. kosten?

Die vorgeschlagene Verteilung ist nur ein Modell. Man könnte die Dauer der Bauarbeiten zum Beispiel auch als exponential-verteilt annehmen.

(f) Für welchen Parameter λ hat die Exponentialverteilung denselben Erwartungswert wie die bisherig angenommene Verteilung?

(g) Berechnen Sie mit der gefundenen Exponentialverteilung nochmals Teilaufgabe (e).

Aufgabe 6

Ein technisches System hat eine exponentialverteilte Lebensdauer mit Parameter $c = 0.04$.

(**a**) Berechnen Sie den Median und den Erwartungswert. Mit welcher Wahrscheinlichkeit überlebt das System seine Lebenserwartung?

(**b**) Mit welcher Wahrscheinlichkeit liegt die Lebensdauer des Systems im Bereich $\mu \pm \sigma$?

(**c**) Beweisen Sie die Formeln für Erwartungswert und Varianz einer Exponentialverteilung mit Parameter $c > 0$.

Aufgabe 1

X sei die Anzahl Patienten, die auf die Behandlung ansprechen. Es gilt also $X \sim \text{Bin}(n, \pi)$ mit $n = 16$.

(**a**) Gemäss Fragestellung haben wir $H_0: \ \pi_0 = 0.15$ und $H_A: \ \pi > 0.15$.

Der Verwerfungsbereich hat also die Form $K = [c, n]$.

Wir bestimmen c indem, wir $P_{H_0}(X \in K)$ für verschiedene, grösser werdende c berechnen, so lange, bis die entsprechende Wahrscheinlichkeit kleiner oder gleich 5% wird. Dabei benutzen wir, dass

$$P_{H_0}(X \geq c) = 1 - P_{H_0}(X \leq c - 1)$$

gilt.

```
> 1-pbinom(1,16,0.15)
[1] 0.7160988
> 1-pbinom(2,16,0.15)
[1] 0.4386207
> 1-pbinom(3,16,0.15)
[1] 0.2101093
> 1-pbinom(4,16,0.15)
[1] 0.0790513
> 1-pbinom(5,16,0.15)
[1] 0.02354438
```

Somit ist unser $c = 5 + 1 = 6$, d.h. der Verwerfungsbereich ist $K = [6, 16]$. Da 5 nicht im Verwerfungsbereich liegt, wird die Nullhypothese beibehalten.

(**b**) Obigem R-Output entnehmen wir, dass bei der Beobachtung mit P-Wert $p = 0.0790513$ der Testentscheid von Beibehalten zu Verwerfen wechselt.

(**c**) Die Wahrscheinlichkeit, dass H_0 verworfen wird, wenn die wahre Ansprechwahrscheinlichkeit $\pi = 0.3$ ist, berechnet sich wie folgt:

$$P_{\pi=0.3}(T \in K) = P_{\pi=0.3}(T \geq 6) = 1 - P_{\pi=0.3}(T \leq 5) = 1 - \sum_{k=0}^{5} \binom{16}{k} 0.3^k 0.7^{16-k}$$

Wir berechnen dies mit R:

```
> 1-pbinom(5,16,0.3)

[1] 0.3402177
```

Die Wahrscheinlichkeit beträgt also 0.3402177. Dies ist die Macht dieses statistischen Tests.

Aufgabe 2

(a) Wir benutzen folgende Notation: $R = \textit{Reise}$; $A = \textit{Absage}$.

$$P[3R\ 1A] = \binom{4}{3} \cdot 0.9^3 \cdot 0.1^1 = 0.2916 = 29.16\%$$

(b) S_n sei der Anzahl Personen, die den Flug nehmen möchten. S_n ist binomialverteilt. Mit 28 Passagiere haben wir:

$$\begin{aligned} S_{28} &\sim \text{Bin}(28, 0.9) \\ \mathbf{E}[S_{28}] &= 28 \cdot 0.9 = 25.2 \\ \text{Var}(S_{28}) &= 28 \cdot 0.9 \cdot 0.1 = 2.52 \end{aligned}$$

(c)

$$\begin{aligned} P[\text{Zu viele Leute}] &= P[k = 27] + P[k = 28] \\ &= \binom{28}{27} \cdot 0.9^{27} \cdot 0.1^1 + \binom{28}{28} \cdot 0.9^{28} \cdot 0.1^0 \\ &= 0.1628 + 0.05233 \\ &= 0.215154 \\ &= 21.52\% \end{aligned}$$

(d)

- Nullhypothese und Alternative

$$\begin{aligned} H_0 &: \ \pi = \pi_0 = \frac{801}{890} \\ H_A &: \ \pi \neq \pi_0 \end{aligned}$$

- Das Signifikanzniveau ist $\alpha = 0.05$.
- Verwerfungsbereich (Normalapproximation):

$$K = [0, c_u] \cup [c_0, n] = [0, 783.45] \cup [818.54, 890]$$

wobei

$$\begin{aligned} c_u &= n\pi_0 - 1.96\sqrt{n\pi_0(1-\pi_0)} \quad \text{abrunden} \\ c_o &= n\pi_0 + 1.96\sqrt{n\pi_0(1-\pi_0)} \quad \text{aufrunden} \end{aligned}$$

- Testentscheidung: ist die beobachtete Anzahl Personen am Flughafen in K ? Ja, so wird die Nullhypohese deutlich verworfen.

Aufgabe 3

(**a**) Das 95% Vertrauensintervall ist gegeben durch $\frac{x}{n} \pm \frac{1.96}{\sqrt{n}}\sqrt{\frac{x}{n}(1-\frac{x}{n})}$. Mit x= Beobachtete Anzahl Gewinne und n= Anzahl Wiederholungen, finden wir [0.0438,0.2362].

(**b**)
```
?binom.test
binom.test(7,50)
```

Das 95% Vertrauensintervall ist gegeben durch

```
95 percent confidence interval:
```

0.0581917 0.2673960

Aufgabe 4

(**a**) Das Intervall sollte in 95% der Fälle den wahren Wert enthalten. Da wir 20 Realisationen betrachten, erwarten wir, dass der Wert im Schnitt 1 mal nicht im Intervall enthalten ist. Wir verwenden folgenden R-Code:

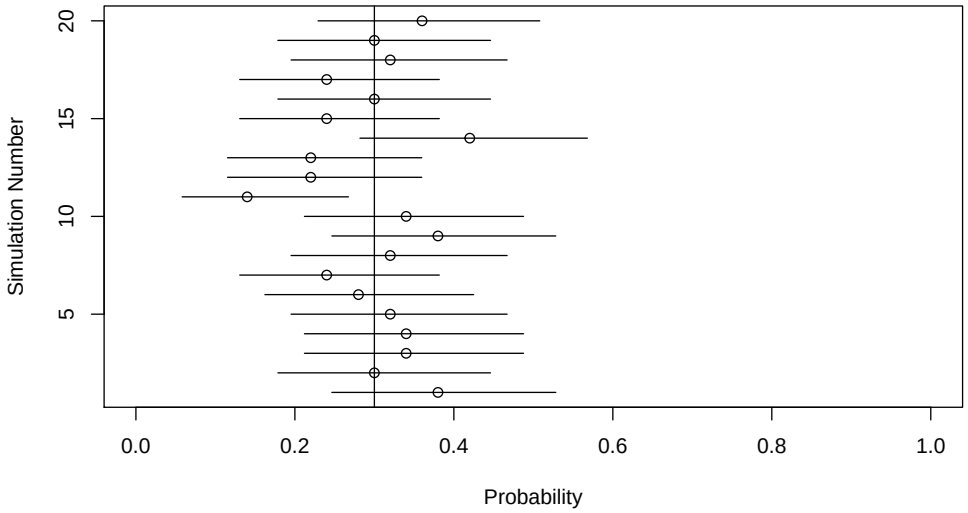
```
> set.seed(79) ## Macht Resultate reproduzierbar
> p <- 0.3
> x <- rbinom(20, 50, p) ## 20 Werte simulieren
> ## Grenzen der Intervalle in Matrix speichern
> ## 1. Spalte ist untere Grenze, 2. Spalte obere
> confint.bound  <- matrix(0, nrow = 20, ncol = 2)
> contains.truth <- logical(20)
> ## Alle 20 Faelle untersuchen und Grenzen speichern
> for(i in 1:20){
+ test <- binom.test(x[i], 50, p)
+ confint.bound[i,] <- test$conf.int
+ contains.truth[i] <-
+ (p >= confint.bound[i,1]) & (p <= confint.bound[i,2])
+ }
> sum(contains.truth)

[1] 19
```

Für unsere Simulationen ist der wahre Wert wie erwartet in 19 der Vertrauensintervalle enthalten. Je nach Simulation kann es natürlich sein, dass der Wert immer enthalten oder in weniger als 19 Fälle enthalten ist (die Anzahl Intervalle, die den wahren Wert enthalten, ist binomialverteilt mit $n = 20$ und Erfolgswahrscheinlichkeit 0.95)

(b) **R-Code**:

```
> ## Relative Haeufigkeiten plotten
> plot(x / 50, 1:20, xlim = c(0, 1), xlab = "Probability",
+ ylab = "Simulation Number")
> ## Vertrauensintervalle als Liniensegemente plotten
> for(i in 1:20){
+ segments(confint.bound[i,1], i, confint.bound[i,2], i)
+ }
> ## Wahrer Wert als vertikale Linie einzeichnen
> abline(v = p)
```

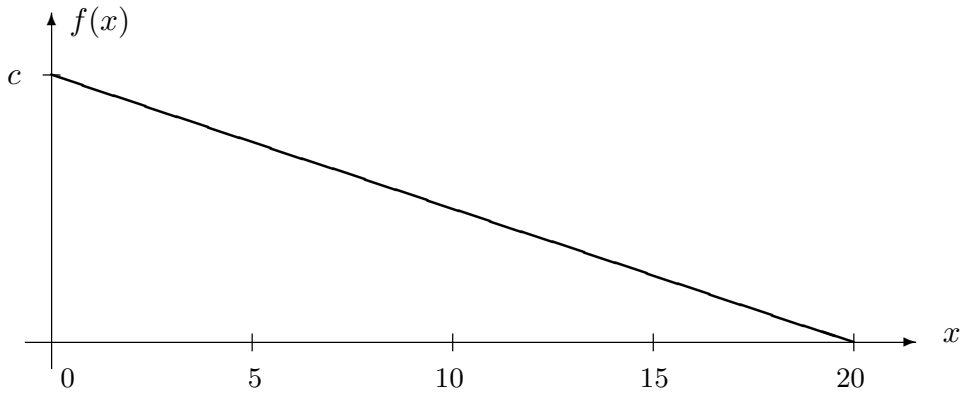
und wir erhalten so folgende Graphik



Vorbesprechung: 10/14. April 2013

Aufgabe 1

In der Stadt Zürich gibt es bekanntlich viele Baustellen. Die Dauer X der Arbeiten bei einer Baustelle liege zwischen 0 und 20 Wochen. Die Dichte $f(x)$ habe die folgende Form.



(**a**) Begründen Sie, warum $c = 0.1$ ist und schreibe die Dichte $f(x)$ explizit auf.

(**b**) Berechnen Sie die Wahrscheinlichkeit, dass die Bauzeit X weniger als

(i) 5

(ii) 10

Wochen beträgt.

(**c**) Skizzieren Sie die kumulative Verteilungsfunktion.

(**d**) Berechnen Sie den Erwartungswert, den Median und die Standardabweichung der Dauer X.

(**e**) $K = 40'000 \cdot \sqrt{X}$ entspreche dem Betrag in Franken, den die Arbeiten bei einer Baustelle kosten. Wie gross ist die Wahrscheinlichkeit, dass die Arbeiten bei einer Baustelle höchstens 120'000.- Fr. kosten?

Die vorgeschlagene Verteilung ist nur ein Modell. Man könnte die Dauer der Bauarbeiten zum Beispiel auch als exponential-verteilt annehmen.

(f) Für welchen Parameter λ hat die Exponentialverteilung denselben Erwartungswert wie die bisherig angenommene Verteilung?

(g) Berechnen Sie mit der gefundenen Exponentialverteilung nochmals Teilaufgabe (e).

Aufgabe 2

Monte Carlo Algorithmen sind randomisierte Algorithmen und stellen ein gutes Werkzeug für Simulationen von stochastischen Prozessen dar. Auch die Zahl π lässt sich mit Hilfe von Monte Carlo Simulationen bestimmen. Im folgenden möchten wir ein Computerprogramm erstellen, mit welchem man die Zahl π aufgrund von Monte Carlo Methoden simulieren kann. Man generiert hierzu zufällige Punkte $P \in \{(x,y)|x \in [-1,1] \text{ und } y \in [-1,1]\}$ und überprüft, ob diese innerhalb des Einheitskreises mit Kreismittelpunkt $M_K = (0,0)$ und Radius $r = 1$ liegen. Die sich ergebende Wahrscheinlichkeitsverteilung $P[(x,y) \in \text{Kreis}]$ stellt die Fläche eines Viertels des Einheitskreises dar. π kann nun mit folgender Formel berechnet werden

$$\frac{\text{Kreisfläche}}{\text{Quadratfläche}} = \frac{r^2 \cdot \pi}{(2 \cdot r)^2} \overset{r=1}{=} \frac{\pi}{4} = \frac{\text{Treffer in Kreisfläche}}{\text{generierte Punkte im Rechteck}} = P[(x,y) \in \text{Kreis}] \,.$$

Bestimmen Sie mit Hilfe dieser Überlegung die Zahl π.

R-Hinweise:

```
## Generieren von 100 gleichmaessig verteilten Zufallszahlen im Intervall [-1,1]
runif(100,min=-1,max=1)
## Bestimmen der Anzahl Zahlen kleiner als eins; Beispiel: Anzahl von 100
## zufaellig im Intervall [0,10] generierten Zahlen, die kleiner als 1 sind:
sum(runif(100,min=0,max=10)<1)
```

Aufgabe 3

Ein technisches System hat eine exponentialverteilte Lebensdauer mit Parameter $c = 0.04$.

(**a**) Berechnen Sie den Median und den Erwartungswert. Mit welcher Wahrscheinlichkeit überlebt das System seine Lebenserwartung?

(**b**) Mit welcher Wahrscheinlichkeit liegt die Lebensdauer des Systems im Bereich $\mu \pm \sigma$?

(**c**) Beweisen Sie die Formeln für Erwartungswert und Varianz einer Exponentialverteilung mit Parameter $c > 0$.

Aufgabe 4

Aufgrund langjähriger Untersuchungen ist bekannt, dass der Bleigehalt X in einer Bodenprobe annähernd normalverteilt ist. Ausserdem weiss man, dass der Erwartungswert 32 ppb beträgt und dass die Standardabweichung 6 ppb beträgt.

(**a**) Machen Sie eine Skizze der Dichte von X und zeichnen Sie die Wahrscheinlichkeit, dass eine Bodenprobe zwischen 26 und 38 ppb Blei enthält, in die Skizze ein.

(**b**) Wie gross ist die Wahrscheinlichkeit, dass eine Bodenprobe höchstens 40 ppb Schwermetall enthält?

Hinweis: Gehen Sie zur standardisierten Zufallsvariablen Z über und benutzen Sie die R-Funktion **pnorm**.

(**c**) Wie gross ist die Wahrscheinlichkeit, dass eine Bodenprobe höchstens 27 ppb Schwermetall enthält?

(**d**) Welcher Bleigehalt wird mit einer Wahrscheinlichkeit von 97.5% unterschritten? Das heisst, bestimmen Sie dasjenige c, so dass die Wahrscheinlichkeit, dass der Bleigehalt kleiner oder gleich c ist, genau 97.5% beträgt.

(**e**) Welcher Bleigehalt wird mit einer Wahrscheinlichkeit von 10% unterschritten?

(**f**) Wie gross ist die Wahrscheinlichkeit, die in Aufgabe a) eingezeichnet wurde?

Aufgabe 5

In einer Studie wurde untersucht, wie bei Mäusen die Aufnahme von Eisen (Fe^{3+}) von der Dosis abhängt. Dazu wurden 54 Mäuse zufällig in 3 Gruppen zu je 18 Mäusen eingeteilt und jeweils mit Dosis hoch, mittel und tief gefüttert (hoch = 10.2 millimolar, mittel=1.2 millimolar, tief=0.3 millimolar). Mittels radioaktiver Markierung wurde der Anteil des zurückgehaltenen Eisens in Prozent nach einer gewissen Zeit bestimmt.
Die Daten können Sie einlesen mit dem Befehl

```
> iron <- read.table("http://stat.ethz.ch/Teaching/Datasets/ironF3.dat",header = TRUE)
```

(**a**) Erstellen Sie für jede der 3 Versuchsbedingungen einen Boxplot, am Besten gerade nebeneinander. Wie unterscheiden sich die Daten der verschiedenen Versuchsbedingungen?

(**b**) Transformieren Sie alle Werte mit dem Logarithmus und erstellen Sie wieder die 3 Boxplots wie bei Aufgabe a). Was hat sich durch die Transformation geändert?

(**c**) Erstellen Sie einen Normalplot der Daten bei mittlerer Dosis vor und nach dem Logarithmieren. Wann passt die Normalverteilung besser? Verwenden Sie die R-Funktion **qqnorm**.

Vorbesprechung: 17/18. April 2013

Aufgabe 1

Für grossangelegte Simulationen müssen im allgemeinen **pseudozufällige** Zahlen generiert werden; diese Zahlen heissen pseudozufällig, da sie mit Hilfe eines Algorithmus erzeugt werden und daher nicht "wirklich" zufällig sind. Angenommen wir möchten die Performance von Warteschlangennetzwerken beurteilen mit Hilfe einer Simulation, dann müssen wir zufällige Zeitintervalle zwischen den Ankünften von Kunden generieren. Wir nehmen an, diese Zeitintervalle folgen einer Exponentialverteilung.
Verfügen wir über keinen Zufallsgenerator für exponentialverteilte Zufallszahlen, dann können wir exponentialverteilte Zufallszahlen mit Hilfe von gleichmässig im Intervall $[0, 1]$ verteilten Zufallszahlen erzeugen, und zwar mit folgender Überlegung: Sei U eine uniform auf dem Intervall $[0, 1]$ verteilte Zufallsvariable, und sei $X = F_X^{-1}(U)$, wobei $F_X(x) = 1 - e^{-\lambda\, x}$ die kumulative Verteilungsfunktion der Exponentialverteilung ist. Dann gilt

$$P(X \leq x) = P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_U(F_X(x)) = F_X(x)\,,$$

wobei wir im letzten Schritt benutzt haben, dass die kumulative Verteilungsfunktion der uniformen Verteilung gegeben ist durch $F_U(u) = u$, falls $u \in [0, 1]$. Die Zufallsvariable, die durch $X = F_X^{-1}(U)$ definiert wurde, folgt also einer Exponentialverteilung.

(**a**) Lösen Sie $F_X(x) = 1 - e^{-\lambda\, x} = u$ nach x auf, d.h., bestimmen Sie die Funktion $F_X^{-1}(\,\cdot\,)$. Generieren Sie nun Zufallszahlen $X \sim \text{Exp}(\lambda = 2)$ mit $F_X^{-1}(U)$, wobei U uniform auf $[0, 1]$ verteilt ist. *Hinweis:* Konsultieren Sie Serie 2 Aufgabe 2 für den **R**−Befehl `runif`.

(**b**) Überprüfen Sie nun mit einem qq-Plot, ob die in Teilaufgabe (a) generierten Zahlen exponentialverteilt sind.
R-Hinweis:

```
## Bestimmen der theoretischen Quantilen der Exponentialverteilung mit
## lambda=2: wenn z.B. n=100 Datenpunkte vorliegen, dann werden die
## theoretischen Quantile generiert mit
n <- 100
qexp((seq(1,n,by=1)-0.5)/n,rate=2)
## die empirischen Quantilen fuer den qq-plot erhalten Sie, indem Sie die
## Datenpunkte (dargestellt als Kompoenenten eines Vektors x) der Groesse
## nach ordnen, und zwar mit der Funktion
sort(x)
```

Aufgabe 2

Ein Statistiker beobachtet, dass ein Angler innerhalb von 2 Stunden 15 Fische fängt. Er nimmt an, dass es sich um einen Poissonprozess handelt und überlegt sich:

(**a**) Mit welcher Wahrscheinlichkeit dauert es länger als 12 Minuten, bis der nächste Fisch anbeisst?

(**b**) Mit welcher Wahrscheinlichkeit beissen innerhalb der nächsten 12 Minuten genau 2 Fische an?

Aufgabe 3

(**a**) Gegeben sind zwei unabhängige Zufallsvariable X und Y mit den Kennwerten $\mu_X = 40$, $\sigma_X = 15$, $\mu_Y = 85$ und $\sigma_Y = 18$. Berechnen Sie $\mathrm{E}(X + 2Y)$, $\mathrm{Var}(X + 2Y)$ und $\mathrm{E}(X^2)$.

(**b**) Ein Werk produziert rechteckige Glasscheiben, deren Länge X und Breite Y (in mm gemessen) voneinander unabhängig produktionsbedingten Schwankungen unterliegen. Es gilt $\mu_X = 1000$, $\sigma_X = 0.02$, $\mu_Y = 500$, $\sigma_Y = 0.01$. Wie gross sind Erwartungswert und Standardabweichung des Umfangs U ?

(**c**) (*Zusatzaufgabe*) Bestimmen Sie die Wahrscheinlichkeitsdichte $f_X(x)$ der Zufallsvariablen $X = Z^2$, wobei $Z \sim \mathcal{N}(0, 1)$.

Aufgabe 4

In dieser Aufgabe untersuchen Sie die Wirkung des Zentralen Grenzwertsatzes mittels Simulation. Gehen Sie von einer Zufallsvariablen X aus, die folgendermassen verteilt ist: die Werte 0, 10 und 11 werden je mit einer Wahrscheinlichkeit $\frac{1}{3}$ angenommen. Simulieren Sie nun die Verteilung von X sowie die Verteilung des Mittelwerts $\overline{X}$ von mehreren X.

(**a**) Simulieren Sie X. Stellen Sie die Verteilung von X mittels eines Histogramms von 1000 Realisierungen von X dar, und vergleichen Sie sie mittels des Normalplots mit der Normalverteilung.

```
> par(mfrow=c(4,2))    # Mehrere Grafiken neben- und untereinander
> werte <- c(0,10,11)  # moegliche Werte von X
> sim <- sample(werte,1000, replace = TRUE)  # X simulieren
> hist(sim, main=paste("Original"))           # Histogramm erstellen
> qqnorm(sim)          # Normalplot erstellen
```

(**b**) Simulieren Sie nun $\overline{X} = \frac{X_1+X_2+X_3+X_4+X_5}{5}$, wobei die X_i die gleiche Verteilung haben wie X und unabhängig sind. Stellen Sie die Verteilung von $\overline{X}$ anhand von 1000 Realisierung von $\overline{X}$ dar, und vergleichen Sie mit der Normalverteilung.

```
> n<-5
> sim<-matrix(sample(werte,n*1000,replace=TRUE),ncol=n)
>      # X_1,...,X_n simulieren und in einer n-spaltigen Matrix
>      # (mit 1000 Zeilen) anordnen
> sim.mean<- apply(sim,1,"mean")   #In jeder Matrixzeile Mittelwert berechnen
> hist(sim.mean)
> title(paste("Mittelwerte von",n,"Beobachtungen"))
> qqnorm(sim.mean)
```

(**c**) Simulieren Sie nun die Verteilung von $\overline{X}$ auch für die Fälle, wo $\overline{X}$ das Mittel von 10 resp. 200 X_i ist.

Aufgabe 1

Die Zufallsvariable X ist definiert als $X = F_X^{-1}(U)$, wobei $U \sim \text{Uniform}([0,1])$ und $F_X(x) = 1 - \lambda e^{-\lambda x}$ die kumulative Verteilungsfunktion der Exponentialverteilung ist. Dann ist

$$X = F_X^{-1}(U) = \frac{-\log(1-U)}{\lambda}\,,$$

Definieren wir die Zufallsvariable

$$V = 1 - U\,,$$

so gilt $V \sim \text{Uniform}([0,1])$. Die Zufallsvariable X kann äquivalent folglich auch definiert werden als

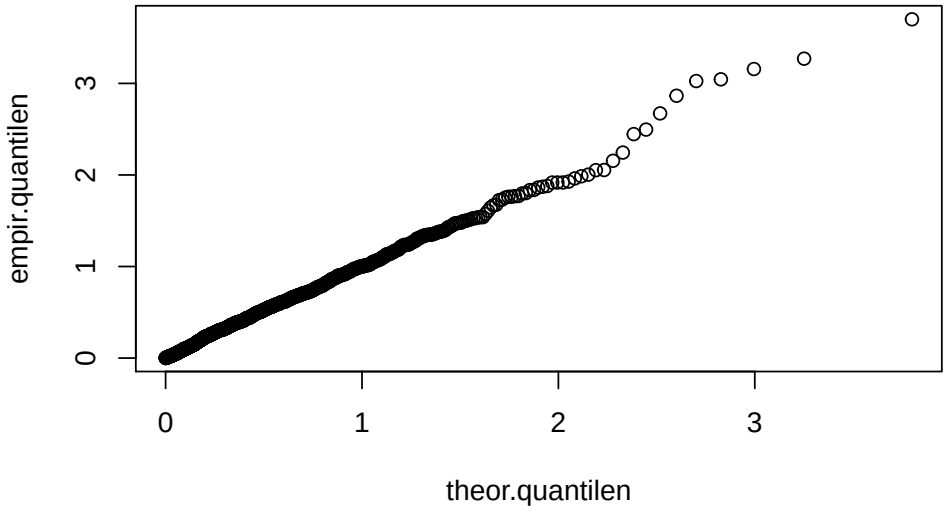
$$X = \frac{-\log(V)}{\lambda}\,.$$

Wir erzeugen nun aufgrund der obigen Vorschrift mit `R` $n = 1000$ exponentialverteilte Zufallszahlen $X_1, X_2, ..., X_{1000}$.

```
> ## Generiere 1000 gleichmÃ¤ssig verteilte Zufallszahlen im Intervall [0,1]
> v <- runif(1000,min=0,max=1)
> ## Generiere 1000 exponentialverteilte Zufalsszahlen mit Hilfe der gleichmÃ¤ssig
> ## verteilten Zufallszahlen
> x <- -log(v)/2
```

Um zu überprüfen, ob die generierten Zahlen X_i $(i = 1, ..., 1000)$ exponentialverteilt sind, fertigen wir einen qq-Plot an.

```
> ## qq-Plot
> ## Berechne die theoretischen Quantilen
> n <- 1000
> theor.quantilen <- qexp((seq(1,n,by=1)-0.5)/n,rate=2)
> ## Berechne die empirischen Quantilen
> empir.quantilen <- sort(x)
> qqplot(theor.quantilen,empir.quantilen)
```



Die empirischen Quantilen können als eine lineare Funktion der theoretischen Quantilen mit Ordinatenabschnitt 0 betrachtet werden, woraus wir schliessen, dass die generierten Zufallzahlen tatsächlich exponentialverteilt sind.

Aufgabe 2

(**a**) Die Wartezeit T ist exponentialverteilt, d.h. hat die Wahrscheinlichkeitsdichte ist $f(t) = \lambda \cdot \mathrm{e}^{-\lambda t}$ für $t > 0$ mit Parameter $\lambda = \frac{1}{8}$. Also

$$P(T > 12) = 1 - P(T \leq 12) = 1 - (1 - \mathrm{e}^{-12/8}) = \mathrm{e}^{-1.5} = 0.223$$

(**b**) X sei die Anzahl Fische, die in den nächsten 12 Minuten anbeissen; X ist poissonverteilt mit $\lambda = 1.5$ (in 12 Minuten beissen durchschnittlich 1.5 Fische an), also

$$P(X = 2) = \mathrm{e}^{-1.5} \cdot \frac{1.5^2}{2!} = 0.251$$

Aufgabe 3

(**a**) $\mathrm{E}(X + 2Y) = \mu_X + 2\mu_Y = 210\,, \quad \mathrm{Var}(X + 2Y) = \sigma_X^2 + 4\sigma_Y^2 = 1521\,,$
$\mathrm{E}(X^2) = \mathrm{Var}(X) + (\mathrm{E}(X))^2 = \sigma_X^2 + \mu_X^2 = 1825\,.$

(**b**) $U = 2X + 2Y; \quad \mathrm{E}(U) = 2\mathrm{E}(X) + 2\mathrm{E}(Y) = 2\mu_X + 2\mu_Y = 3000\,,$
$\sigma_U = \sqrt{4\sigma_X^2 + 4\sigma_Y^2} = 0.0447\,.$

(**c**) Wir haben $X = Z^2$, wobei $Z \sim \mathcal{N}(0,1)$. Dann ist

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(Z^2 \leq x) \\ &= P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= \Phi(\sqrt{x}) - \Phi(-\sqrt{x})\,. \end{aligned}$$

Wir erhalten die Wahrscheinlichkeitsdichte von X, indem wir die kumulative Verteilungsfunktion $F_X(x)$ nach x ableiten. Da $\Phi'(x) = \phi(x)$, ergibt sich mit der Kettenregel

$$\begin{aligned} fX(x) &= \frac{1}{2}x^{-1/2}\phi(\sqrt{x}) + \frac{1}{2}x^{-1/2}\phi(-\sqrt{x}) \\ &= x^{-1/2}\phi(\sqrt{x})\,, \end{aligned}$$

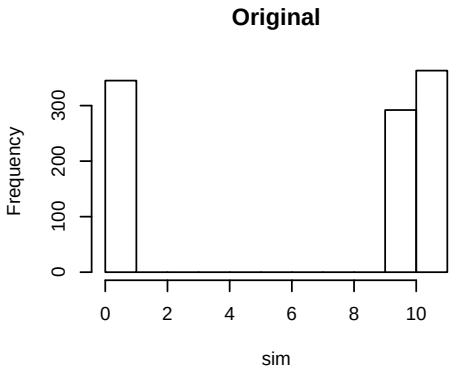
wobei wir im letzten Schritt die Symmetrie von Φ benützt haben. Werten wir den letzten Ausdruck aus, so finden wir

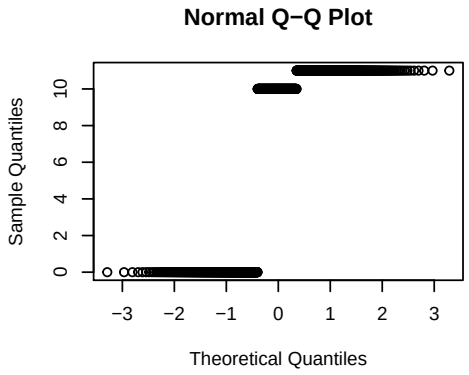
$$f_X(x) = \frac{x^{-1/2}}{\sqrt{2\pi}}e^{-x^2/2}\,, \quad x \geq 0\,.$$

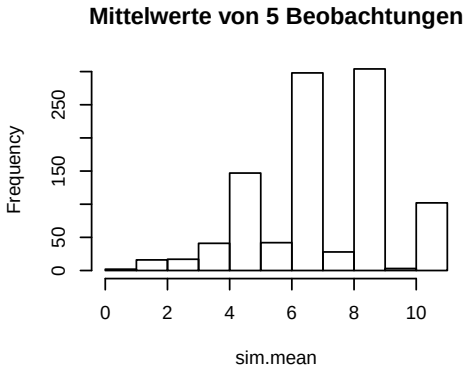
Diese Wahrscheinlichkeitsdichte wird **Chi-Quadrat** Wahrscheinlichkeitsdichte mit einem Freiheitsgrad genannt.

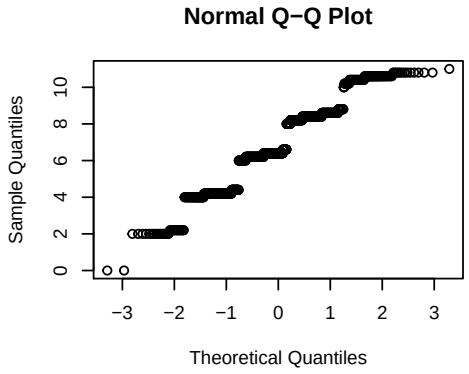
Aufgabe 4

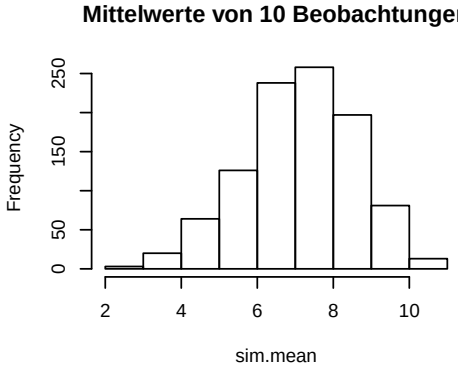
Die untenstehenden Graphiken zeigen, dass die Form der Verteilung des Mittelwerts von unabhängigen Zufallsvariablen auch dann der Normalverteilung immer ähnlicher wird, wenn die Variablen selber überhaupt nicht normalverteilt sind. An der x-Achse sieht man auch, dass die Varianz immer kleiner wird.

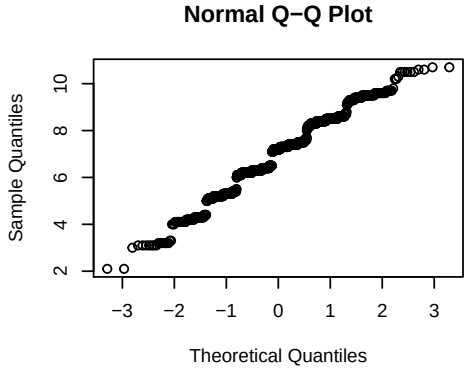

Original
Frequency
0
100
200
300
0
2
4
6
8
10
sim

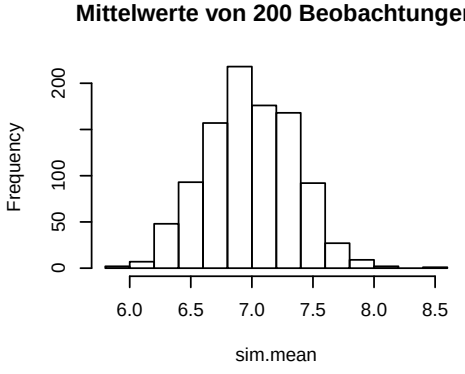

Normal Q−Q Plot
Sample Quantiles
0
2
4
6
8
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles

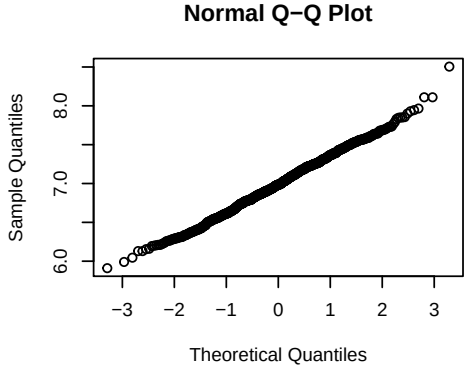

Mittelwerte von 5 Beobachtungen
Frequency
0
50
150
250
0
2
4
6
8
10
sim.mean


Normal Q−Q Plot
Sample Quantiles
0
2
4
6
8
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Mittelwerte von 10 Beobachtungen
Frequency
0
50
150
250
2
4
6
8
10
sim.mean


Normal Q−Q Plot
Sample Quantiles
2
4
6
8
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Mittelwerte von 200 Beobachtungen
Frequency
0
50
100
200
6.0
6.5
7.0
7.5
8.0
8.5
sim.mean


Normal Q−Q Plot
Sample Quantiles
6.0
7.0
8.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Vorbesprechung: 24/25. April 2013

Aufgabe 1

Die Auswertung eines Integrals

$$I(f) = \int_0^1 f(x)dx$$

kann sehr oft nicht analytisch erfolgen. Der gebräuchlichste Ansatz in diesem Fall besteht darin, das Integral numerisch zu berechnen. Dazu existieren verschiedene Computerprogramme. Eine andere geläufige Methode, um ein solches Integral zu berechnen, ist die sogenannte **Monte Carlo Methode**. Man generiert dabei uniform verteilte Zufallsvariablen auf dem Intervall $[0, 1]$, d.h., $X_1, X_2, ..., X_n$ und berechnet

$$\hat{I}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\,.$$

Aufgrund des *Gesetzes der grossen Zahlen* ist für grosse n

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \mathrm{E}[f(X)]\,.$$

Somit ist

$$E[f(X)] = \int_0^1 f(x)dx = I(f)\,.$$

Dieses einfache Schema kann beliebig angepasst werden, zum Beispiel an unterschiedliche Integrationsgrenzen.
Berechnen Sie folgendes Integral

$$I(f) = \frac{1}{\sqrt{2\pi}}\int_0^1 e^{-x^2/2}dx\,.$$

Berechnen Sie das Integral, indem Sie 1000 uniform über das Intervall $[0, 1]$ verteilte Punkte, $X_1, ..., X_{1000}$ generieren. Berechnen Sie den genauen numerischen Wert des Integrals mit der R-Funktion `pnorm`.

Aufgabe 2

Aus der uniformen Verteilung $X \sim \mathrm{Uniform}([0, 10])$ soll eine Stichprobe vom Umfang n gezogen werden.

(**a**) Es sei $n = 60$. Bestimmen Sie ein symmetrisches Intervall $I = [\mu - e, \mu + e]$ um den Erwartungswert μ so, dass sich das arithmetische Mittel der Stichprobe mit der Wahrscheinlichkeit von 95% in I befindet. Ein solches Intervall heisst **Prognoseintervall**.

Hinweis: Standardisieren Sie das arithmetische Mittel $\overline{X}_n$ und benützen Sie den Zentralen Grenzwertsatz.

(**b**) Umgekehrt: Wie gross muss n gewählt werden, damit $e = 0.2$ wird?

(**c**) Überprüfen Sie (a) experimentell, d.h. mit `R`: ziehen Sie viele Stichproben (z.B. 200) und zählen Sie, wie viele ausserhalb von I liegen.
R-Hinweise:

```
> n<-60  # Anzahl Stichproben
> sim<-matrix(runif(n*200,min=0,max=10),ncol=n)
>        # X_1,...,X_n simulieren und in einer n-spaltigen Matrix
>        # (mit 200 Zeilen) anordnen
> sim.mean<- apply(sim,1,"mean")   #In jeder Matrixzeile Mittelwert berechnen
> plot(sim.mean)
```

Zeichnen Sie mit `abline(h=...)` die Intervallgrenzen des Prognoseintervalls in der obigen Graphik ein.

Aufgabe 3

Das Gastroberatungsunternehmen Lecker und Co. kreiert eine neue Speisekarte für ein Schnellrestaurant. Lecker und Co. nimmt auf Grund langjähriger Erfahrung an, dass etwa (unabhängig von der Anzahl der Kunden) 80% der Kunden des Schnellrestaurants die neue Speisekarte bevorzugen werden.

(**a**) Am Einführungstag speisen 356 Kunden im Restaurant. Wie gross ist unter obiger Annahme der Erwartungswert für die Anzahl der Kunden, welche die neue Speisekarte bevorzugen?

Lecker und Co. führt bei den 356 Kunden eine kurze Befragung durch. 261 Kunden geben dabei an, dass sie die neue Karte bevorzugen. Der Rest findet die alte Karte mindestens genauso gut wie die neue Karte.

(**b**) Wie gross ist die Wahrscheinlichkeit unter der Annahme von Lecker und Co., dass keiner der ersten vier befragten Kunde die neue Karte bevorzugt, der fünfte Kunde jedoch die neue Karte bevorzugt? Wie gross ist die Wahrscheinlichkeit, dass drei der ersten vier befragten Kunden die neue Karte bevorzugen?

(**c**) Wie gross ist unter der Annahme von Lecker und Co. die Wahrscheinlichkeit, dass höchstens 261 Kunden die neue Karte bevorzugen? Benutzen Sie die Normalapproximation.

(**d**) Lecker und Co. hat nun starke Zweifel, dass wirklich 80% der Kunden die neue Karte bevorzugen. Lecker und Co. will deshalb die Annahme mit Hilfe der Befragung kritisch überprüfen. Führen Sie den entsprechenden (zweiseitigen) Test durch. Verwenden Sie die Normalapproximation.

Aufgabe 4

Ein Weinhändler behauptet, dass die von ihm gefüllten Weinflaschen mindestens 70 Zentiliter enthalten. Ein skeptischer Konsument vermutet aber, dass der Weinhändler zu wenig Wein abfüllt und will diese Behauptung überprüfen. Deshalb kauft er 12 Weinflaschen und misst ihren Inhalt. Er findet:

$$71,\ 69,\ 67,\ 68,\ 73,\ 72,\ 71,\ 71,\ 68,\ 72,\ 69,\ 72 \text{ (in Zentiliter).}$$

(**a**) Nehmen Sie zunächst an, dass die Standardabweichung der Abfüllung im voraus bekannt ist. Sie beträgt $\sigma = 1.5$ Zentiliter. Da die Standardabweichung der Messungen bekannt ist, können wir einen z-Test durchführen. Führen Sie den (einseitigen; in welche Richtung?) Test auf dem 5%- Signifikanzniveau durch und formulieren Sie in einem Satz die Schlussfolgerung für den kritischen Konsumenten.

(**b**) Tatsächlich ist die Standardabweichung der Abfüllungen aber nicht bekannt. Man muss sie also aus den gemachten Stichproben schätzen:

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx 1.96^2$$

Da nun die Standardabweichung geschätzt wurde und nicht mehr exakt bekannt ist, kann der z- Test nicht mehr durchgeführt werden. Verwenden Sie nun den t-Test auf dem 5%-Signifikanzniveau. Was ändert sich an obigem Test? Wie lautet das Ergebnis?

Aufgabe 5

Im National Bureau of Standards (USA) wurden regelmässig Wägungen des 10-Gramm-Standardgewichtstücks durchgeführt. Bei 9 Wägungen erhielt man als durchschnittliche Differenz -403 Mikrogramm vom 10 Gramm-Sollgewicht und eine Standardabweichung von 3.127 Mikrogramm für eine einzelne Wägung.

(**a**) Geben Sie das exakte, zweiseitige 95%-Vertrauensintervall für die wahre Differenz an, unter der Annahme, dass die Messfehler normalverteilt sind.

(**b**) Könnte die wahre Differenz -400.0μg betragen? Entscheiden Sie aufgrund des Resultats in Aufgabe (a). (Kurze Begründung)

Aufgabe 1

Wir berechnen das Integral

$$I(f) = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-x^2/2} dx$$

mit der Monte-Carlo Methode, da analytisch keine Lösung in geschlossener Form existiert. Zuerst generieren wir 1000 uniform im Intervall $[0, 1]$ verteilte Zufallszahlen $X_1, ..., X_{1000}$

```
> n <- 1000
> x <- runif(n,min=0,max=1)
```

Danach werten wir den Integranden für alle 1000 Zufallszahlen aus

```
> integrand <- exp(-x^2/2)
```

Das mit Monte-Carlo berechnete Integral ergibt dann den Wert

```
> 1/(sqrt(2*pi))*sum(integrand)/n
```

```
[1] 0.344725
```

Das Integral, das wir eben mit der Mone-Carlo Methode berechnet haben, ist natürlich die Differenz der kumulativen Verteilungsfunktion der Standardnormalverteilung ausgewertet an den Stellen 0 und 1.

```
> pnorm(1)-pnorm(0)
```

```
[1] 0.3413447
```

Aufgabe 2

(**a**) Für die uniforme Verteilung $X \sim \text{Uniform}([0, 10])$ gilt $\text{E}(X) = \mu = 5$, $\sigma_X = \frac{5}{\sqrt{3}}$. Nach dem Zentralen Grenwertsatz gilt für grosse n: wird der arithmetische Mittelwert $\overline{X}_n$ der Stichprobe vom Umfang n standardisiert, so gilt für die standardisierte Zufallsvariable

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma_X/\sqrt{n}} = \frac{\overline{X}_n - 5}{5/\sqrt{3n}} \sim \mathcal{N}(0, 1) \, .$$

Es gilt dann für das gesuchte Intervall $[\mu - e, \mu + e]$, dass

$$\begin{aligned} P(\mu - e \leq \overline{X}_n \leq \mu + e) &= P\left(-\frac{e}{5/\sqrt{3n}} \leq \frac{\overline{X}_n - 5}{5/\sqrt{3n}} \leq \frac{e}{5/\sqrt{3n}}\right) \\ &= P\left(-\frac{e}{5/\sqrt{3n}} \leq Z_n \leq \frac{e}{5/\sqrt{3n}}\right) \\ &= \Phi\left(\frac{e}{5/\sqrt{3n}}\right) - \Phi\left(-\frac{e}{5/\sqrt{3n}}\right) \\ &= 0.95 \end{aligned}$$

Aufgrund der Symmetrie der Standardnormalverteilung genügt es, eine Seite der Verteilung zu betrachten: $\Phi\left(\frac{e}{5/\sqrt{3n}}\right)$ soll also 97.5% der Fläche unter der Gesamtkurve entsprechen. Das 97.5%-Quantil $q(0.975)$ der Standardnormalverteilung ist

$$\frac{e}{5/\sqrt{3n}} = q(0.975) = \Phi^{-1}(0.975) = 1.96\,.$$

Also ist

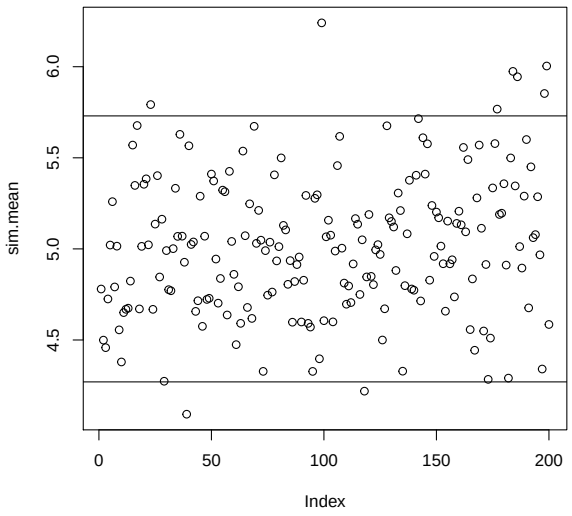
$$e = \frac{5}{\sqrt{3n}} \cdot q(0.975) = \frac{5}{\sqrt{3 \cdot 60}} \cdot 1.96 = 0.73$$

(**b**) Auflösen der Gleichung $e = 0.2 = \frac{5}{\sqrt{3n}} \cdot 1.96$ nach n liefert $n = 800$.

(**c**) $I = [5 - 0.73, 5 + 0.73]$, $n = 200$. Man erwartet ca. $0.05 \cdot 200 = 10$;

```
> n <- 60
> sim <- matrix(runif(n*200,min=0,max=10),ncol=n)
> sim.mean <- apply(sim,1,"mean")
> plot(sim.mean)
> abline(h=5.73)
> abline(h=4.27)


> d<-sum(sim.mean>5.73)+sum(sim.mean<4.27)
```

In unserem Beispiel sind es 9.



Aufgabe 3

(**a**) X_i = Inhalt (in Zentiliter) der i-ten Weinflasche, $i = 1, \cdots, n = 12$.

1. **Modell**: $X_1, \cdots, X_{12}$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, $\sigma^2 = 1.5^2$ bekannt.
2. **Nullhypothese**: $H_0: \ \mu = \mu_0 = 70$
   Alternative: $H_A: \ \mu < \mu_0$
3. **Teststatistik**:
$$Z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}$$
   Verteilung der Teststatistik unter $H_0: \ Z \sim \mathcal{N}(0, 1)$
4. **Signifikanzniveau**: $\alpha = 5\%$
5. **Verwerfungsbereich für die Teststatistik**:
$$\Phi^{-1}(0.95) = 1.645 \Rightarrow K = (-\infty, -1.645]$$
6. **Testentscheid**:
$$z = \sqrt{12}\frac{70.25 - 70}{1.5} = 0.5774$$
   $z \notin K \rightarrow H_0$ beibehalten. Es ist also durchaus plausibel, dass der Weinhändler den Wein korrekt abfüllt.

(**b**) 1. **Modell**: $X_1, \cdots, X_{12}$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, σ^2 unbekannt; geschätzter Wert: $\hat{\sigma}_x^2 = 1.96^2$
2. **Nullhypothese**: $H_0: \ \mu = \mu_0 = 70$
   Alternative: $H_A: \ \mu < \mu_0$
3. **Teststatistik**:
$$T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\hat{\sigma}_X}$$
   Verteilung der Teststatistik unter $H_0: \ T \sim t_{n-1}$

4. **Signifikanzniveau**: $\alpha = 5\%$
5. **Verwerfungsbereich für die Teststatistik**:

$$t_{11;0.95} = 1.796 \Rightarrow K = (-\infty, -1.796]$$

6. **Testentscheid**:

$$t = \sqrt{12}\frac{70.25 - 70}{1.96} = 0.441$$

$t \notin K \rightarrow H_0$ beibehalten. Wir kommen also zum selben Ergebnis wie in Teilaufgabe (a).

Aufgabe 4

(**a**) $\left[-403 \pm t_{9-1;97.5\%} \cdot \frac{3.127}{\sqrt{9}}\right] = [-403 \pm 2.31 \cdot 1.042] = [-405.4, -400.6]$

(**b**)) Da -400.0 nicht im 95%-Vertrauensintervall liegt, würde die Nullhypothese $H_0: \ \mu = -400.0$ zu Gunsten der Alternative $H_A: \ \mu \neq -400.0$ auf dem 5%-Signifikanzniveau verworfen werden. Die Beobachtungen und die Hypothese $H_0: \ \mu = -400.0$ passen also nicht gut zusammen und daher ist die wahre Differenz wohl nicht -400.0.

Vorbesprechung: 1/2. Mai 2013

Aufgabe 1

Ein radioaktives Nuklid, das beim Zerfall Helium-4-Atomkerne aussendet, wird als *Alphastrahler* bezeichnet. Der vom zerfallenden Atomkern emittierte Helium-4-Atomkern wird *Alphateilchen* genannt. Alphateilchen können zum Beispiel mit einem Zink Sulfid Schirm detektiert werden, wobei Lichtblitze bei der Kollision von Alphateilchen mit dem Schirm entstehen. Die Emission von Alphateilchen von einer radioaktiven Quelle pro Zeiteinheit ist nicht konstant, sondern fluktuiert auf zufällige Art und Weise. Ein typischer Alphastrahler ist americium 241, ein Zerfallsprodukt von Plutonium-241, das in radioaktiven Abfällen häufig anzutreffen ist. Rauchmelder enthalten kleine Mengen von americium 241.
In einem Experiment wurde nun die Anzahl Zerfälle von americium 241 in einem Intervall von 10 Sekunden gemessen. Das Experiment wurde 1207 Mal wiederholt, jedes Mal wurde die Anzahl Zerfälle in 10 Sekunden gemessen. In untenstehender Tabelle ist in der ersten Spalte die Anzahl Zerfälle aufgeführt, in der zweiten Spalte, wie oft diese Anzahl Zerfälle in den 1207 Experimenten beobachtet worden ist. So sind in 18 der 1207 Experimente 0, 1 oder 2 Alphateilchen detektiert worden. In 28 der 1207 Experimente wurden 3 Alphateilchen detektiert etc.

Anzahl Zerfälle	Beobachtet in Anzahl Experimente
0-2	18
3	28
4	56
5	105
6	126
7	146
8	164
9	161
10	123
11	101
12	74
13	53
14	23
15	15
16	9
17+	5

Tabelle 1: Anzahl Zerfälle in 10 Sekunden und Anzahl Experimente (von insgesamt 1207), in denen diese Anzahl Zerfälle beobachtet wurde.

(**a**) Wieviele Zerfälle wurden im gesamten beobachtet? Nehmen Sie an, dass 18 mal 2 Zerfälle beobachtet wurden und 5 mal 17 Zerfälle.
R-Hinweis: Benützen Sie `rep()` und `sum()`.

(**b**) Stellen Sie mit einer geeigneten graphischen Darstellungsmethode die Häufigkeiten für jede Anzahl Zerfälle dar. Beachten Sie in der Darstellung das Intervall für $0-2$ und $17+$ Zerfälle.

(**c**) Zeichnen Sie die relativen Häufigkeiten für jede Anzahl Zerfälle pro 10 Sekunden auf. Welche graphische Darstellung wählen Sie, um relative Häufigkeiten anzugeben. Wie interpretieren Sie diese relativen Häufigkeiten?

(**d**) Stellen Sie die geschätzten Wahrscheinlichkeiten für jede Anzahl Zerfälle in einem Stabdiagramm dar.

(**e**) Berechnen Sie das arithmetische Mittel und die empirische Standardabweichung der Daten. Zeichnen Sie diese in die Graphik von Teilaufgabe (d) ein und erklären Sie deren Bedeutung.

(**f**) Zeichnen Sie die empirische kumulative Verteilungsfunktion. Erklären Sie die Bedeutung der kumulativen Verteilungsfunktion.

(**g**) Welches (theoretische) Modell für die beobachtete Wahrscheinlichkeitsverteilung würden Sie wählen? Berücksichtigen Sie dabei, dass die Gesamtzahl Atome extrem gross ist und ein Zerfall ein seltenes Ereignis darstellt. Warum ist man überhaupt an einem theoretischen Modell interessiert? Bestimmen Sie die theoretische Wahrscheinlichkeitsverteilung und zeichnen Sie diese als Stabdiagramm auf mit Mittelwert und Standardabweichung.

(**h**) Sie machen ein weiteres Experiment und beobachten 20 Zerfällle in 10 Sekunden. Wenn Sie eine Million Experimente durchführen, wie oft werden Sie ein solches oder extremeres Ereignis beobachten? Wie nennt man diese Wahrscheinlichkeit?

(**i**) Angenommen, 9 weitere Labors in der gesamten Welt repetieren dieses Experiment und messen 1207 mal die Anzahl Zerfälle in 10 Sekunden. Wie sind die in allen Labors geschätzten $\hat{\lambda}$ verteilt? Wie gross ist die Streuung (Standardabweichung) der zehn geschätzten $\hat{\lambda}$, wenn nun 12070 mal die Anzahl Zerfälle in 10 Sekunden gemessen wird?

(**j**) Angenommen, zum Zeitpunkt $t_0 = 0$ gab es einen Zerfall. Wie gross ist die Wahrscheinlichkeit, dass erst nach dem Zeitpunkt t erneut ein Zerfall eintreten kann? Welche Wahrscheinlichkeitsverteilung erhalten Sie für die Zerfallszeit?

(**k**) Skizzieren Sie die Wahrscheinlichkeitsdichte für die Zerfallszeit von der americium 241 Probe. Tragen Sie ebenfalls den Erwartungswert und Standardabweichung ein. Verdeutlichen Sie sich an diesem Beispiel die Definitionen von Wahrscheinlichkeitsdichte, Erwartungswert und Standardabweichung.

(**l**) (Zusatzaufgabe) Nach einer Reaktorschmelze in einem Kernkraftwerk ist eine erhebliche Menge americium 241 freigesetzt worden. Wieviele Jahre dauert es, bis 50 Prozent davon zerfallen ist? In der von uns gemessenen Probe befanden sich $1.64 \cdot 10^{10}$ Atomkerne.

Aufgabe 2

Aufgrund langjähriger Untersuchungen ist bekannt, dass der Bleigehalt X von Kopfsalaten annähernd normalverteilt ist

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

Ausserdem weiss man, dass der Erwartungswert 32 ppb beträgt und dass die Standardabweichung 6 ppb beträgt.

(**a**) Machen Sie eine Skizze der Dichte von X und zeichnen Sie die Wahrscheinlichkeit, dass ein Kopfsalat zwischen 26 und 38 ppb Blei enthält, in die Skizze ein.

(**b**) Wie gross ist die Wahrscheinlichkeit, dass ein Kopfsalat höchstens 40 ppb Schwermetall enthält? *Hinweis*: Gehen Sie zur standardisierten Zufallsvariablen Z über und benutzen Sie die die `R`-Funktionen `pnorm()`.

(**c**) Wie gross ist die Wahrscheinlichkeit, dass ein Kopfsalat höchstens 27 ppb Schwermetall enthält?

(**d**) Welcher Bleigehalt wird mit einer Wahrscheinlichkeit von 97.5% unterschritten? Das heisst, bestimmen Sie dasjenige c, so dass die Wahrscheinlichkeit, dass der Bleigehalt kleiner oder gleich c ist, genau 97.5% beträgt.

(**e**) Welcher Bleigehalt wird mit einer Wahrscheinlichkeit von 10% unterschritten?

(**f**) Wie gross ist die Wahrscheinlichkeit, die in Aufgabenteil (a) eingezeichnet wurde?

(**g**) Es wurden in 10 Salatköpfen der Bleigehalt X gemessen. Dabei wurde ein Mittelwert von $\overline{X}_{10} = 31$ ppb erhalten. Die Standardabweichung sei bekannt und beträgt 6 ppb. Geben Sie ein 99% Vertrauensintervall für den Mittelwert an.

(**h**) Wieviele Beobachtungen sind nötig, um die Breite des in Teilaufgabe (g) bestimmten Vertrauensintervalles auf die Hälfte zu reduzieren? Wieviele (unabhängige) Bestimmungen des Bleigehalts müssen geplant werden, wenn der Bleigehalt mit einer Stichprobe "auf 1 ppb genau" bestimmt werden soll, d.h., wenn die Breite des 99% des Konfidenzintervalls nicht grösser als 1 ppb sein soll?

(**i**) Normalerweise ist die Standardabweichung σ unbekannt. Um welchen Faktor verändert sich die Breite des Vertrauensintervalls in Teilaufgabe (a), wenn man die Standardabweichung aus den Daten geschätzt hat?

Aufgabe 3

Unterhalb einer Kläranlage wurden 16 unabhängige Wasserproben aus einem Fluss entnommen und jeweils deren Ammoniumkonzentration X_i (in $\mu gNH_4 - N/l$) mit einem Messgerät bestimmt. Der Mittelwert der Proben ergab $\bar{x} = 204.2$.
Wir wollen nun wissen, ob mit einem Experiment eine Überschreitung des Grenzwerts von $200 \mu gNH_4 - N/l$ nachgewiesen werden kann (auf dem 5% Niveau).

(**a**) Nehmen Sie an, die Standardabweichung der Messungen sei im voraus aufgrund früherer Studien bekannt. Sie betrage $10\mu gNH_4 - N/l$.

Führen Sie unter dieser Annahme einen z-Test durch, um zu prüfen, ob eine Grenzwertüberschreitung nachgewiesen werden kann.

Geben Sie die Modellannahmen, H_0, H_A, den Verwerfungsbereich, den Wert der Teststatistik und das Testergebnis explizit an.

(**b**) Wie wahrscheinlich ist es, dass man mit 16 unabhängigen Wasserproben eine Grenzwertüberschreitung nachweisen kann, wenn die wahre Ammoniumkonzentration tatsächlich über dem Grenzwert und zwar bei $205\mu gNH_4 - N/l$ liegt?

(**c**) Wie wahrscheinlich ist es, dass man mit 16 unabhängigen Wasserproben fälschlicherweise eine Grenzwertüberschreitung nachweist, obwohl die wahre Ammoniumkonzentration bei $200\mu gNH_4 - N/l$ liegt und den Grenzwert somit genau einhält?

(**d**) Nehmen Sie an, dass die Standardabweichung von $10\mu g/l$ aus den 16 Proben geschätzt worden ist. Deshalb ist nun ein t-Test (zur Nullhypothese $\mu_o = 200\mu g/l$) und nicht ein z-Test angebracht. Führen Sie den t-Test durch.

Aufgabe 4

Ein Weinhändler behauptet, dass die von ihm gefüllten Weinflaschen mindestens 70 Zentiliter enthalten. Ein skeptischer Konsument vermutet aber, dass der Weinhändler zu wenig Wein abfüllt und will diese Behauptung überprüfen. Deshalb kauft er 12 Weinflaschen und misst ihren Inhalt. Er findet:

71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72 (in Zentiliter).

(**a**) Nun zweifeln wir daran, ob die Daten wirklich gut durch eine Normalverteilung beschrieben werden können (diese Annahme haben wir sowohl beim z- als auch beim t-Test gemacht). Wenn die Normalverteilungsannahme nicht gemacht werden kann, können wir den Vorzeichen-Test durchführen. Führen Sie also den Vorzeichen-Test auf dem 5%-Signifikanzniveau durch. Wie lautet nun das Ergebnis?

(**b**) Wie lautet das Ergebnis mit dem Wilcoxon-Test?
R-Hinweis:
`wilcox.test(x,mu=...)`

pbinom(0:12, 12, p=0.7)

Aufgabe 5

Untenstehend finden Sie mehrere Beispiele für Vergleiche von 2 Stichproben. Beantworten Sie für jedes Beispiel **kurz** die folgenden Fragen:

- Handelt es sich um gepaarte oder um ungepaarte Stichproben? Begründen Sie!
- Ist der Test einseitig oder zweiseitig durchzuführen? Begründen Sie!
- Wie lautet die Nullhypothese in Worten?
- Wie lautet die Alternativhypothese in Worten?

(**a**) In einem Experiment sollte der Effekt von Zigarettenrauchen auf Blutplättchenanhäufungen untersucht werden. Dazu wurden 11 Probanden vor und nach dem Rauchen einer Zigarette Blutproben entnommen, und es wurde gemessen, wie stark sich die Blutplättchen anhäuften. Es interessiert, ob sich Blutplättchen durch das Rauchen vermehrt anhäufen.

(**b**) Die nächsten Daten sind aus einer Studie von Charles Darwin über die Fremd- und Selbstbefruchtung. 15 Paare von Setzlingen mit demselben Alter, je einer durch Selbst- und einer durch Fremdbefruchtung produziert, wurden gezüchtet. Beide Teile je eines Paares hatten nahezu gleiche Bedingungen. Das Ziel bestand darin zu sehen, ob die fremdbefruchteten Pflanzen mehr Lebenskraft besitzen als die selbstbefruchteten (d.h., ob sie grösser werden). Es wurden die Höhen jeder Pflanze nach einer fixen Zeitspanne gemessen.

(**c**) Beeinflusst der Kalziumgehalt in der Nahrung den systolischen Blutdruck? Zur Überprüfung dieser Frage wurde einer Versuchsgruppe von 10 Männern während 12 Wochen ein Kalziumzusatz verabreicht. Einer Kontrollgruppe von 11 Männern gab man ein Placebopräparat.

(**d**) In einem Experiment wurde untersucht, ob Mäuse zwei Formen von Eisen (Fe^{2+} und Fe^{3+}) unterschiedlich gut aufnehmen. Dazu wurden 36 Mäuse in zwei Gruppen zu je 18 unterteilt und die eine Gruppe mit Fe^{2+} und die andere mit Fe^{3+} "gefüttert". Da das Eisen radioaktiv markiert war, konnte sowohl die Anfangskonzentration wie auch die Konzentration einige Zeit später gemessen werden. Daraus wurde für jede Maus der Anteil des aufgenommenen Eisens berechnet.

Aufgabe 6

Edwin Hubble untersuchte seit 1920 am Mount Wilson Observatory die Eigenschaften von galaktischen Nebeln ausserhalb der Milchstrasse. Mit Überraschung bemerkte er einen Zusammenhang zwischen der Distanz eines Nebels zur Erde und dessen Geschwindigkeit, sich von der Erde fortzubewegen (Fluchtgeschwindigkeit). Hubbles ursprüngliche Daten von 24 galaktischen Nebeln (E. Hubble, "Proceedings of the National Academy of Science 15 (1929): 168-73.) sind in Tabelle 2 gezeigt. Die Fluchtgeschwindigkeit ist in Kilometer pro Sekunde

angegeben und konnte aufgrund der Rotverschiebung im Lichtspektrum der Nebel mit grosser Genauigkeit bestimmt werden. Die Distanz eines Nebels zur Erde wird in megaparsec gemessen: ein Megaparsec entspricht etwa $3.09 \cdot 10^{10}$ m. Die Distanzen werden durch Vergleich der mittleren Luminosität von Nebeln mit der Luminosität von bestimmten bekannten Sternen bestimmt, wobei diese Methode relativ ungenau ist.

Nebel	Geschwindigkeit (km/sec)	Distanz (Mparsec)
S. Mag.	170	0.032
L. Mag. 2	290	0.034
NGC 6822	-130	0.214
NGC 598	-70	0.263
NGC 221	-185	0.275
NGC 224	-220	0.275
NGC 5457	200	0.450
NGC 4736	290	0.500
NGC 5194	270	0.500
NGC 4449	200	0.630
NGC 4214	300	0.800
NGC 3031	-30	0.900
NGC 3627	650	0.900
NGC 4626	150	0.900
NGC 5236	500	0.900
NGC 1068	920	1.000
NGC 5055	450	1.100
NGC 7331	500	1.100
NGC 4258	500	1.400
NGC 4151	960	1.700
NGC 4382	500	2.000
NGC 4472	850	2.000
NGC 4486	800	2.000
NGC 4649	1090	2.000

Tabelle 2: Zusammenhang zwischen Distanz und Fluchtgeschwindigkeit von galaktischen Nebeln.

(**a**) Erstellen Sie von den Daten in Tabelle 2 ein Streudiagramm.

(**b**) Schätzen Sie aus den Daten die Parameter β_0 und β_1 für die Regressionsgerade

$$y = \beta_0 + \beta_1 x \, .$$

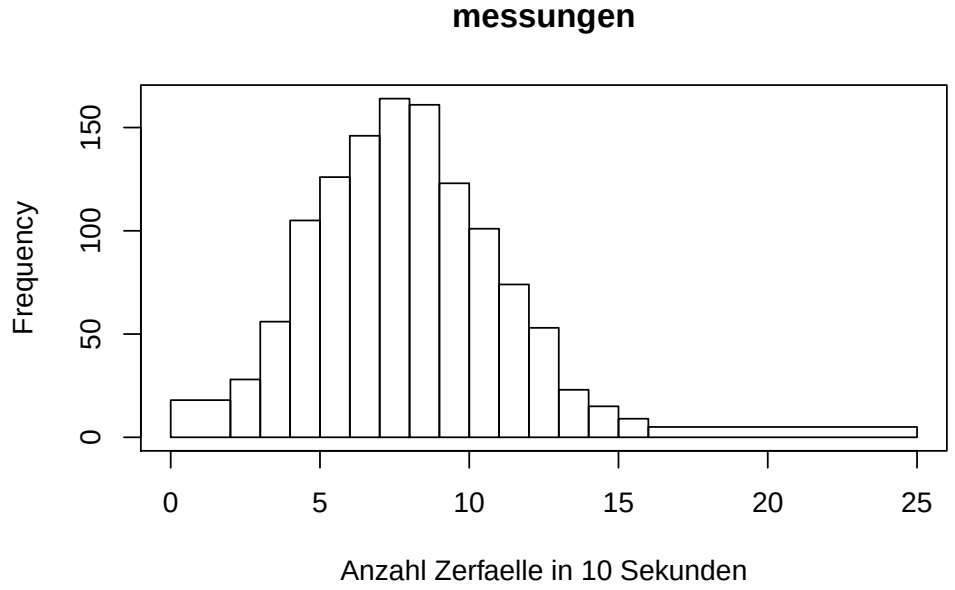
Aufgabe 1

(**a**) Die Gesamtzahl Zerfälle berechnen wir in `R` mit

```
> anzahl.zerfaelle <- seq(2,17,by=1)
> anzahl.beobachtungen <- c(18,28,56,105,126,146,164,161,123,101,74,53,23,15,9,5)
> messungen <- rep(anzahl.zerfaelle,anzahl.beobachtungen)
> sum(messungen)

[1] 10102
```

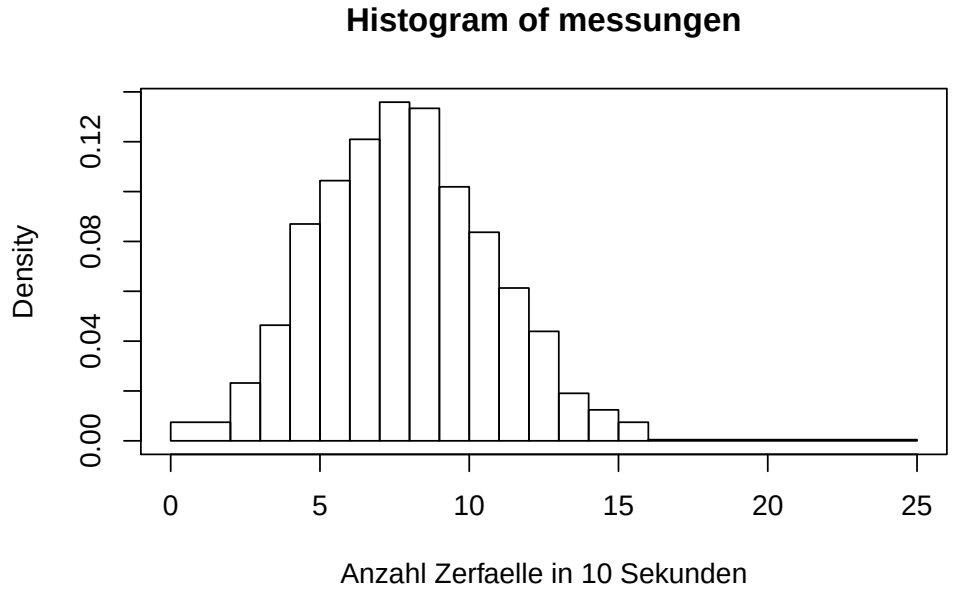
(**b**) Um eine Häufigkeitsverteilung darzustellen, wählen wir als graphische Darstellungsmethode das Histogramm.

```
> anzahl.zerfaelle <- seq(2,17,by=1)
> anzahl.beobachtungen <- c(18,28,56,105,126,146,164,161,123,
+         101,74,53,23,15,9,5)
> messungen <- rep(anzahl.zerfaelle,anzahl.beobachtungen)
> hist(messungen,breaks=c(0,anzahl.zerfaelle[-16],25),
+         xlab="Anzahl Zerfaelle in 10 Sekunden", main="messungen", freq=TRUE)
> box()
```



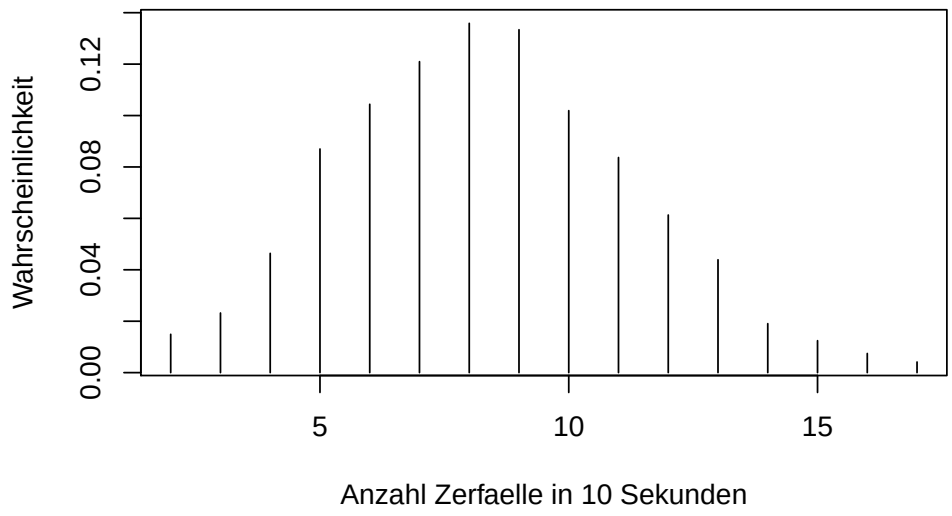
(**c**) Um auf der vertikalen Achse die relative Häufigkeit aufzuzeichnen, ändern wir die Einstellung in der Histogrammfunktion zu `hist(...,freq=FALSE)`.

```
> anzahl.zerfaelle <- seq(2,17,by=1)
> anzahl.beobachtungen <- c(18,28,56,105,126,146,164,161,123,
+         101,74,53,23,15,9,5)
> messungen <- rep(anzahl.zerfaelle,anzahl.beobachtungen)
> hist(messungen,breaks=c(0,anzahl.zerfaelle[-16],25),
+         freq=FALSE,xlab="Anzahl Zerfaelle in 10 Sekunden")
> box()
```



Wir interpretieren die relativen Häufigkeiten der beobachteten Anzahl Zerfälle pro 10 Sekunden als (geschätzte) *Wahrscheinlichkeiten*: die Wahrscheinlichkeit, dass in 10 Sekunden eine entsprechende Anzahl Zerfälle beobachtet wird.

(**d**)
```
> anzahl.zerfaelle <- seq(2,17,by=1)
> anzahl.beobachtungen <- c(18,28,56,105,126,146,164,161,
+         123,101,74,53,23,15,9,5)
> wahrscheinlichkeiten <- anzahl.beobachtungen/sum(anzahl.beobachtungen)
> plot(anzahl.zerfaelle,wahrscheinlichkeiten,type="h",
+         xlab="Anzahl Zerfaelle in 10 Sekunden", ylab="Wahrscheinlichkeit")
```



(**e**) Das arithmetischen Mittel für die Anzahl Zerfälle berechnen wir mit

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

```
> mean(messungen)

[1] 8.369511
```

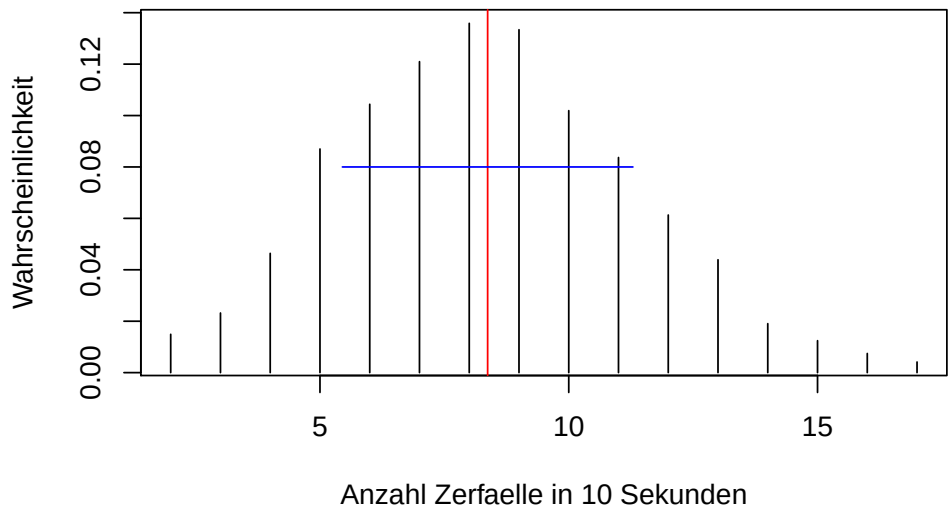
Das arithmetische Mittel beschreibt die mittlere Lage der Häufigkeitsverteilung der Anzahl Zerfäelle in 10 Sekunden. Die (empirische) Standardabweichung für die Anzahl Zerfälle berechnen wir mit

$$s_x = \sqrt{var_x} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

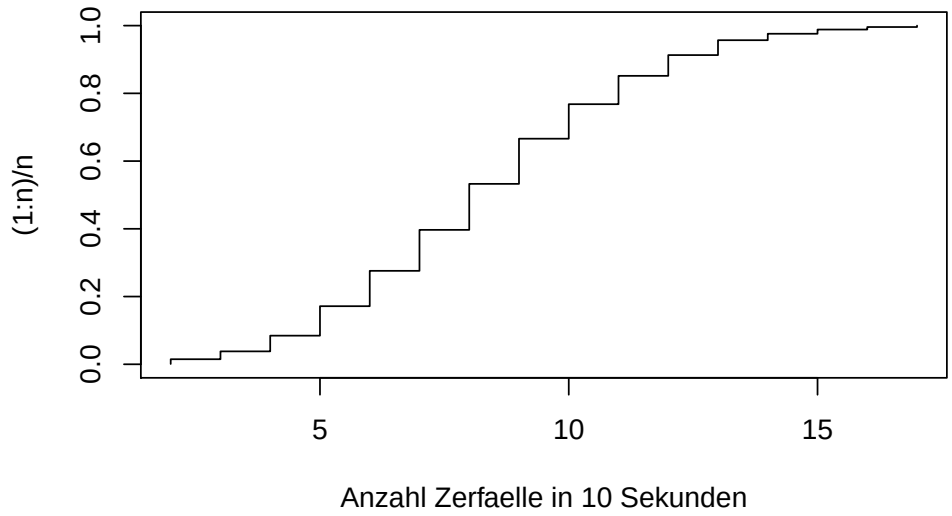
```
> sd(messungen)

[1] 2.920903
```

Die Standardabweichung (ebenso die Varianz) ist ein Mass fuer die Breite der Streuung in der Häufigkeitsverteilung der Anzahl Zerfaelle.

```
> anzahl.zerfaelle <- seq(2,17,by=1)
> anzahl.beobachtungen <- c(18,28,56,105,126,146,164,161,
+         123,101,74,53,23,15,9,5)
> wahrscheinlichkeiten <- anzahl.beobachtungen/sum(anzahl.beobachtungen)
> plot(anzahl.zerfaelle,wahrscheinlichkeiten,type="h",
+         xlab="Anzahl Zerfaelle in 10 Sekunden", ylab="Wahrscheinlichkeit")
> abline(v=8.37,col="red")
> lines(x=c(8.37-2.92,8.37+2.92),y=c(0.08,0.08),col="blue")
```



(**f**)
```
> n <- length(messungen)
> plot(sort(messungen),(1:n)/n, type="s",ylim=c(0,1),
+         xlab="Anzahl Zerfaelle in 10 Sekunden")
```



Die (empirische) kumulative Verteilungsfunktion gibt an, welcher Prozentsatz aller Beobachtungen kleiner als der betrachtete Messwert ist.

(**g**) Die Poisson-Verteilung bietet sich an als Verteilung für seltene Ereignisse bei vielen unabhängigen Versuchen mit zwei möglichen Ausgängen. Wir sind an einem theoretischen Modell interessiert, da wir das beobachtete Phänomen mathematisch formulieren möchten, um dann ein physikalisches Gesetz daraus herzuleiten. Zudem erlaubt uns ein

theoretisches Modell, für alle auch nicht beobachtete Ereignisse eine Wahrscheinlichkeit anzugeben. Wir betrachten also die in den 1207 Experimenten bestimmte Anzahl Zerfälle in 10 Sekunden als Realisierungen von Poisson verteilte Zufallsvariablen. Jede Realsierung hat die Wahrscheinlichkeit:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \,.$$

Wir müssen nun den Parameter λ schätzen. Wir tun dies mit der *Momentenmethode*, wobei wir uns daran erinnern, dass wir den Erwartungswert der Wahrscheinlichkeitsverteilung gleich dem tatsächlich beobachteten Wert setzen. Der Erwartungswert der Poission-Verteilung ist

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \lambda \,.$$

Die mittlere Anzahl Zerfälle in 10 Sekunden beträgt 8.37. Somit

$$\widehat{\mathrm{E}(X)} = 8.37 = \widehat{\lambda} \,.$$

Die Varianz ist definiert als

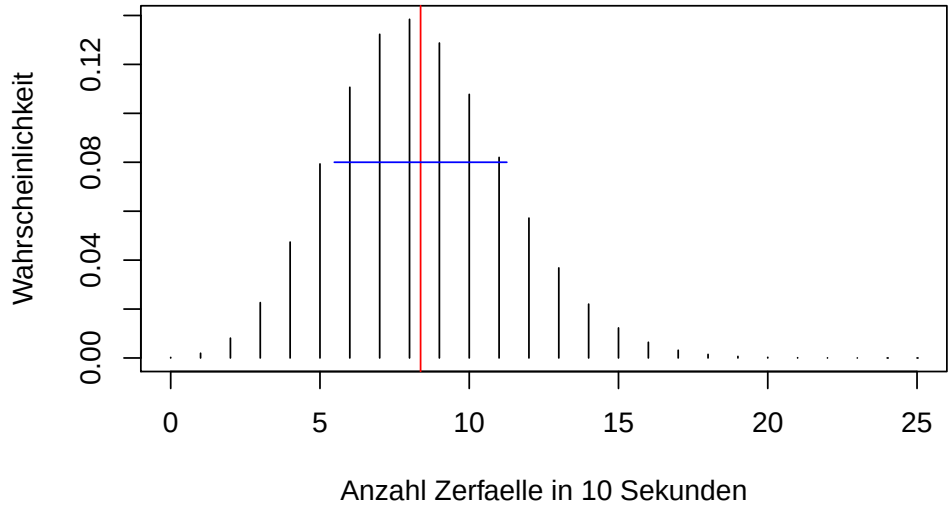
$$\mathrm{Var}(X) = \sum_{k=0}^{\infty} (k - \mathrm{E}(X))^2 P(X = k) = \lambda \,.$$

Die geschätzte Standardabweichung der Poisson-Verteilung ist gegeben durch

$$\widehat{\sigma(X)} = \sqrt{\widehat{\lambda}}$$

Somit ist $\widehat{\sigma(X)} = 2.89$.

```
> anzahl.zerfaelle <- seq(0,25,by=1)
> plot(anzahl.zerfaelle,dpois(anzahl.zerfaelle,lambda=8.37),
+         type="h",xlab="Anzahl Zerfaelle in 10 Sekunden", ylab="Wahrscheinlichkeit")
> abline(v=8.37,col="red")
> lines(x=c(8.37-2.89,8.37+2.89),y=c(0.08,0.08),col="blue")
```



(**h**) Dies ist der sogenannte P-Wert, den wir mit `R` berechnen

```
> 1-ppois(19,lambda=8.37)

[1] 0.0004432101
```

In einer Million Experimente würde man also 20 oder mehr Zerfaelle in 10 Sekunden bloss

```
> (1-ppois(19,lambda=8.37))*10^6

[1] 443.2101
```

Mal beobachten.

(**i**) $\widehat{\lambda}$ wurde durch den Mittelwert der Anzahl Zerfälle in 10 Sekunden geschätzt. Da die gemessene Anzahl Zerfälle in jedem Experiment unabhängig von den anderen Experimenten ist, und jede gemessene Anzahl Zerfälle als Realisierung einer Poisson-verteilten Zufallsvariablen aufgefasst werden kann, dürfen wir den Zentralen Grenzwertsatz anwenden. Gemäss dem Zentralen Grenzwertsatz ist $\widehat{\lambda}$ folgendermassen verteilt:

$$\widehat{\lambda} = \overline{\lambda}_n \sim \mathcal{N}(\mu, \sigma^2/n) \,,$$

wobei μ und σ den Erwartungswert, resp. die Standardabweichung der Poissonverteilung bezeichnen. Diese sind natürlich nicht bekannt und müssen geschätzt werden. Wenn nun in allen 10 Labors die Zahl der Messungen verzehnfacht wird, dann nimmt die Streuung der geschätzten $\widehat{\lambda}$ um den Faktor $1/\sqrt{10}$ ab.

(**j**) Die Wahrscheinlichkeit, dass sich erst nach der Zeit t wieder ein Zerfall ereignet, ist

$$P(T > t) = P(\text{kein Zerfall in } [0, t]) \,.$$

Die Anzahl Zerfälle im Zeitintervall $[0,t]$ folgt einer Poisson-Verteilung mit Parameter λt. Folglich ist

$$P(T > t) = P(\text{kein Zerfall in } [0,t]) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}\,.$$

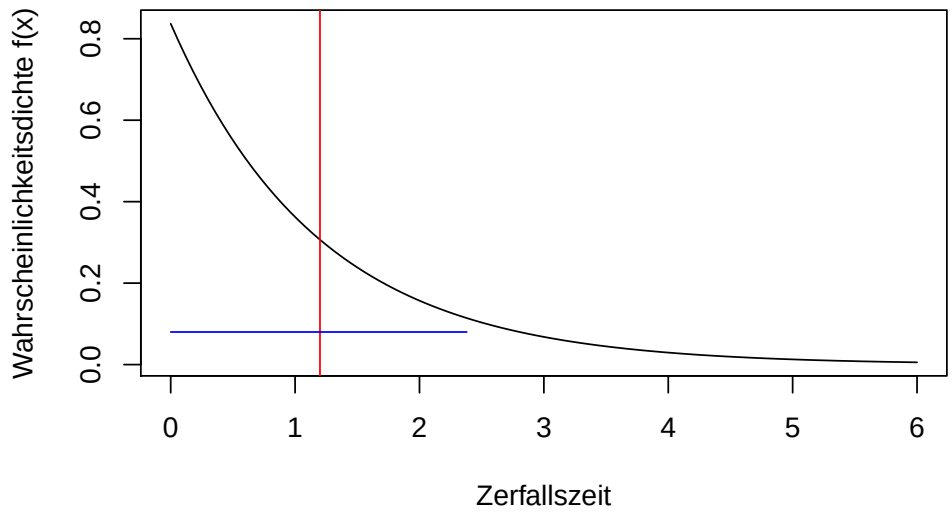
Also folgt die Lebenszeit T eines Atomes einer Exponentialverteilung mit Parameter λ. Die kumulative Verteilungsfunktion ist gegeben durch

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t} \quad \text{für } t \geq 0\,.$$

(**k**) Wir benützen nun den Parameter $\lambda = 0.837$, da wir in Einheiten von Sekunden und nicht von 10 Sekunden arbeiten möchten. Der Erwartungswert und die Standardabweichung der Exponentialverteilung sind gegeben durch

$$\mathrm{E}(T) = \sigma_T = \frac{1}{\lambda} = 1.19$$

```
> curve(dexp(x,rate=0.837),from=0,to=6,xlab="Zerfallszeit",
+         ylab="Wahrscheinlichkeitsdichte f(x)")
> abline(v=1.2,col="red")
> lines(x=c(1.19-1.19,1.19+1.19),y=c(0.08,0.08),col="blue")
```



Der numerische Wert von $f(x)$ ist **keine** Wahrscheinlichkeit. Die Wahrscheinlichkeitsdichte $f(x)$ kann wie folgt verstanden werden: $f(x)dx$ entspricht der Wahrscheinlichkeit $P(x < X < x + dx)$. Der Erwartungswert einer stetigen Zufallsvariaben ist definiert als

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx\,.$$

Die Definition der Varianz einer stetigen Zufallsvariablen lautet

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x) dx\,.$$

Die Wahrscheinlichkeit, dass in 3 Sekunden nach dem letzten Zerfall noch kein weiterer Zerfall beobachtet wurde, ist

```
> 1-pexp(3,rate=0.837)

[1] 0.08118701
```

und entspricht der Fläche unter der Wahrscheinlichkeitsdichtekurve rechts vom Wert $t = 3$.

(**l**) Um zu bestimmen, nach welcher Zeit unser americium 241 Sample mit 50% Wahrscheinlichkeit zerfallen ist, müssen wir die Wahrscheinlichkeitsverteilung für die Zerfallszeit eines einzelnen americium 241 Atomkerns kennen. Wir erinnern uns, dass die Summe von zwei Poisson-verteilten Zufallszahlen $X \sim \text{Poisson}(\lambda_X)$ und $Y \sim \text{Poisson}(\lambda_Y)$ wiederum Poisson-verteilt ist: $X + Y \sim \text{Poisson}(\lambda_X + \lambda_Y)$. Also ist die Wahrscheinlichkeitsverteilung für die Zerfallszeit T eines einzelnen americium 241 Atomkerns

$$T \sim \text{Exp}\left(\frac{\widehat{\lambda}}{n}\right) ,$$

wobei $\widehat{\lambda}$ die aufgrund unseres Experimentes geschätzte mittlere Anzahl Zerfälle pro Sekunde ist und n die Anzahl Atomkerne in unserem Sample. Gesucht ist also die Halbwertszeit $t_{1/2}$, nach der die Hälfte aller Atome zerfallen ist

$$F_T(t_{1/2}) = 0.5 .$$

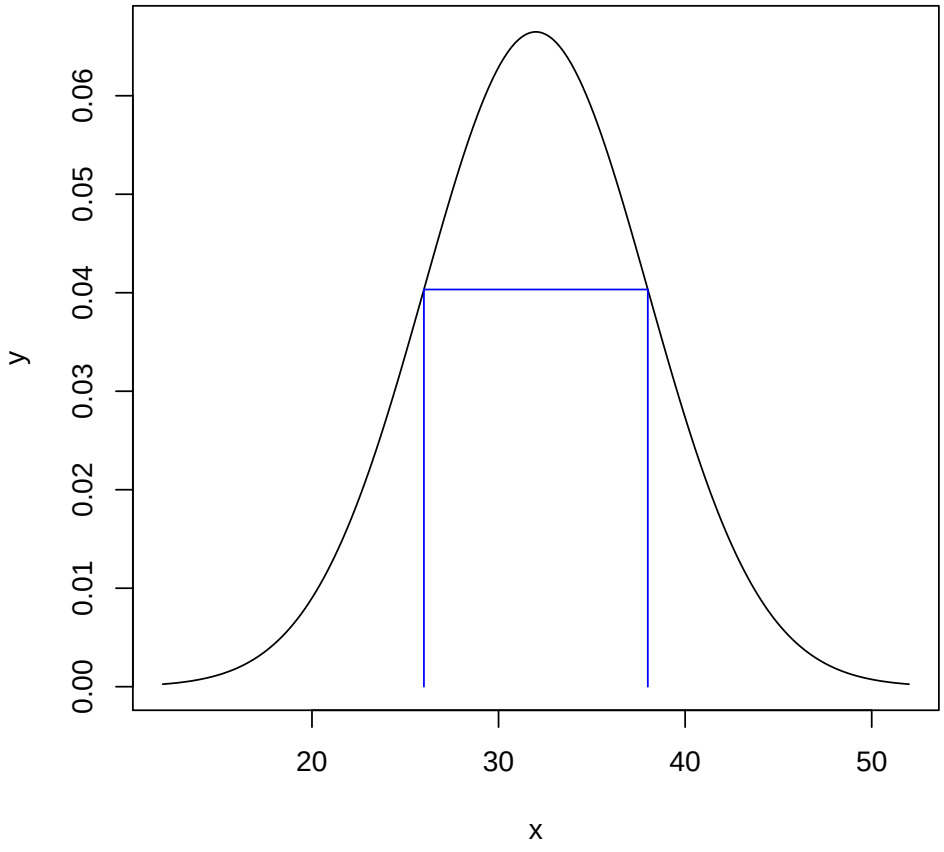
Mit `R` ergibt sich für die Halbwertszeit $t_{1/2}$

```
> qexp(0.5,rate=0.837/(1.64*10^10))

[1] 13581378448
```

Also nach rund 430 Jahren.

Aufgabe 2

(**a**) Skizze:



(**b**) X bezeichne den Schwermetallgehalt in Kopfsalaten. Es gilt:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{mit } \mu = 32 \text{ und } \sigma^2 = 6^2$$

Wir gehen zur standardisierten Zufallsvariablen $Z = (X - \mu)/\sigma$ über. Es gilt: $Z \sim \mathcal{N}(0, 1)$

$$P[X \leq 40] = P\left[Z \leq \frac{40 - 32}{6}\right] = P[Z \leq 1.33] = \Phi(1.33) =$$

```
> pnorm(1.33)

[1] 0.9082409
```

(**c**)

$$P[X \leq 27] = P[Z \leq -0.83] = \Phi(-0.83) = 1 - \Phi(0.83) =$$

```
> 1-pnorm(0.83)
```

```
[1] 0.2032694
```

(**d**)

$$P[X \leq c] = 0.975 = P\left[Z \leq \frac{c-32}{6}\right] = \Phi\left(\frac{c-32}{6}\right)$$

Wir finden $\Phi(1.96) = 0.975$. Also muss gelten:

$$\frac{c-32}{6} = 1.96 \text{ und deshalb } c = 32 + 1.96 \cdot 6 = 43.76$$

(**e**) Wir haben $\Phi(1.28) = 0.9$ und $\Phi(-1.28) = 1 - 0.9 = 0.1$. Somit

$$c = 32 - 1.28 \cdot 6 = 24.31$$

(**f**)

$$\Phi(1) - \Phi(-1) = 2 \cdot \Phi(1) - 1 = 2 \cdot 0.8413 - 1 = 0.6826$$

(**g**) Wir bezeichnen mit X den Bleigehalt in Kopfsalaten. Es gilt

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

Der Mittelwert von $\overline{X}_n$ von $n = 10$ Stichproben ist auch normalverteilt mit Standardabweichung $\sigma/\sqrt{n}$,

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)\,,$$

mit $\sigma^2/n = \frac{36}{10} = 3.6$ und μ unbekannt. Da $\Phi(2.58) = 0.995$, liegen 99% aller Beobachtungen von der standardisierten Zufallsvariable

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

in dem Intervall $[-2.58, 2.58]$. Also liegen 99% aller Beobachtungen von $\overline{X} - \mu$ im Intervall $[-2.58 \cdot \sigma/\sqrt{n}, 2.58 \cdot \sigma/\sqrt{n}]$. Ein 99% Vertrauensintervall für μ ist demnach gegeben durch

$$\left[\widehat{\overline{X}}_{10} - 2.58 \cdot \frac{\sigma}{\sqrt{n}}, \widehat{\overline{X}}_{10} + 2.58 \cdot \frac{\sigma}{\sqrt{n}}\right]\,,$$

wobei $\widehat{\overline{X}}_{10}$ der beobachtete Mittelwert ist, hier $\widehat{\overline{X}}_{10} = 31$. Mit dem bekannten Wert von $\sigma = 6$ und $n = 10$ erhält man also ein 99% Vertrauensintervall

$$[26.1, 35.9]\,.$$

(**h**) Aus Teilaufgabe (g) sieht man, dass die Breite des Vertrauensintervalles wie $1/\sqrt{n}$ abfällt mit der Anzahl n von Beobachtungen. Also sind viermal so viele Beobachtungen, $4 \cdot 10$ nötig, um die Breite des Vertrauensintervalls zu halbieren. Wie aus Teilaufgabe (g) ersichtlich ist die Breite des 99% Vertrauensintervalles

$$2 \cdot 2.58 \cdot \frac{\sigma}{\sqrt{n}}\,.$$

Um die Breite des Vertrauensintervalles kleiner als 1 ppb zu erhalten, muss die Anzahl n der Beobachtungen entsprechend gross werden:

$$\begin{aligned} 2 \cdot 2.58 \cdot \frac{\sigma}{\sqrt{n}} &\leq 1 \\ 51.6 &\leq \sqrt{n} \\ n &\geq 2663\,. \end{aligned}$$

Es müssen mindestens 2663 Beobachtungen vorliegen, um ein 99% Vertrauensintervall von weniger als 1 ppb Breite zu erhalten.

(**i**) Die standardisierte Zufallsvariable

$$\frac{\overline{X}_n - \mu}{\widehat{\sigma}/\sqrt{n}}$$

mit Schätzwert $\widehat{\sigma}$ für σ ist nicht mehr normalverteilt (wie für ein bekanntes, festes σ), sondern folgt einer t-Verteilung mit 9 Freiheitsgraden. Das 99.5% Quantil dieser Verteilung ist bei

```
> qt(0.995,9)

[1] 3.249836
```

Somit fallen 99% der Beobachtungen von

$$\frac{\overline{X}_n - \mu}{\widehat{\sigma}/\sqrt{n}}$$

in das Intervall $[-3.25, 3.25]$. Ein Vertrauensintervall für μ ist daher gegeben durch

$$\left[\widehat{\overline{X}}_{10} - 3.25 \cdot \frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\overline{X}}_{10} + 3.25 \cdot \frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

Für $\widehat{\sigma} = 6$ und $n = 10$ ergibt sich das Vertrauensintervall

$$[24.8, 37.2]\ .$$

Durch Vergleich mit (g) findet man, dass das Vertrauensintervall einen Faktor 3.25/2.58, also um 26% grösser geworden ist.

Aufgabe 3

(**a**) **1. Modell**: X_i: i-te Ammoniumbestimmung, X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$ mit $\sigma = 10$.

2. Nullhypothese: $H_0 : X_i$ i.i.d. $\mathcal{N}(\mu_0, \sigma^2)$ mit $\mu_0 = 200$ und $\sigma = 10$
Alternative: $H_A : X_i$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ mit $\mu > 200$ und $\sigma = 10$ (einseitig)

3. Teststatistik: $Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$
Verteilung der Teststatistik unter H_0: $Z \sim \mathcal{N}(0, 1)$

4. Signifikanzniveau: $\alpha = 0.05$

5. Verwerfungsbereich für die Teststatistik: $K = \{z : \ \Phi(z) > 0.95\} =]1.64, \infty[$
(Dies entspricht dem Verwerfungsbereich $]204.1, \infty[$ für $\overline{X}_n$)

6. Testentscheid: Der Wert der Teststatistik ist

$$z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{204.2 - 200}{\sigma/\sqrt{16}} = 1.68 \, .$$

$1.68 \in K$, also wird die Nullhypothese verworfen. Eine Grenzwertüberschreitung ist statistisch gesichert.

(**b**) Aus der Teilaufgabe (a) folgt, dass die Nullhypothese verworfen werden kann, falls der Mittelwert aller Messungen grösser als 204.1 ist,

$$\overline{X}_n > 204.1 \, .$$

Um die Wahrscheinlichkeit zu berechnen, dass eine Grenzwertüberschreitung nachgewiesen werden kann (H_0 verworfen werden kann), geht man wieder zu einer standardisierten Zufallsvariablen über. Mit $\mu_A = 205$ und $\sigma = 10$ erhält man

$$\begin{aligned} P[\overline{X}_n > 204.1] &= P\left[\frac{\overline{X}_n - \mu_A}{\sigma/\sqrt{n}} > \frac{204.1 - \mu_A}{\sigma/\sqrt{n}}\right] \\ &= P\left[\frac{\overline{X}_n - \mu_A}{\sigma/\sqrt{n}} > -0.36\right] \\ &= P[Z > -0.36] \, . \end{aligned}$$

Dies entspricht also der Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable Z mit Varianz 1,

$$Z \sim \mathcal{N}(0, 1) \, ,$$

einen Wert grösser als -0.36 annimmt. Diese Wahrscheinlichkeit ist wegen der Symmetrie der Normalverteilung gleich zu

$$P[Z \leq 0.36] = 0.6406 \, ,$$

wie man leicht mit `R` bestimmt:

```
> pnorm(0.36)

[1] 0.6405764
```

Die Macht des Tests ist also rund 64%.

(**c**) Dies ist genau das Niveau des Tests und war als 5% vorgegeben.

(**d**) Es ist schwieriger, eine Grenzüberschreitung nachzuweisen, wenn die Standardabweichung aus den Daten geschätzt wird. Die Verteilung der Teststatistik

$$T = \frac{\overline{X}_n - 200}{\widehat{\sigma}/\sqrt{n}}$$

folgt einer t-Verteilung mit $n-1$ Freiheitsgraden. Der t-Test wird folgendermassen formal durchgeführt:

1. **Modell**: X_i: i-te Ammoniumbestimmung, X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$ mit σ unbekannt .
2. **Nullhypothese**: $H_0 : X_i$ i.i.d. $\mathcal{N}(\mu_0, \sigma^2)$ mit $\mu_0 = 200$
   Alternative: $H_A : X_i$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ mit $\mu > 200$ (einseitig)
3. **Teststatistik**: $T = \frac{\overline{X}_n - \mu_0}{\widehat{\sigma}/\sqrt{n}}$
   Verteilung der Teststatistik unter H_0: $Z \sim \mathcal{N}(0, 1)$
4. **Signifikanzniveau**: $\alpha = 0.05$
5. **Verwerfungsbereich für die Teststatistik**: $K = \{t : \ t_{15,0.95} > 0.95\} =]1.753, \infty[$
   (Dies entspricht dem Verwerfungsbereich $]204.1, \infty[$ für $\overline{X}_n$)
6. **Testentscheid**: Der Wert der Teststatistik ist

$$t = \frac{\overline{X}_n - \mu_0}{\widehat{\sigma}/\sqrt{n}} = \frac{204.2 - 200}{\widehat{\sigma}/\sqrt{16}} = 1.68 \, .$$

   $1.68 \notin K$, also kann die Nullhypothese nicht verworfen werden. Eine Grenzwertüberschreitung ist statistisch nicht gesichert.

Der Unterschied zum z-Test ist nicht sehr gross, führt hier aber gerade dazu, dass die Nullhypothese nicht mehr verworfen werden kann.

Aufgabe 4

(**a**) In dieser Teilaufgabe bezeichnet μ den Median der stetigen Zufallsvariblen X, also $P(X \leq \mu) = 0.5$. Um Hypothesen über den Median der Verteilung von X_i zu testen, verwendet man den Vorzeichentest.

1. **Modell**: $X_1, \cdots, X_{12}$ i.i.d. , wobei X_i eine beliebige Verteilung hat.
2. **Nullhypothese**: $H_0 : \ \mu = \mu_0 = 70$
   Alternative: $H_A : \ \mu < \mu_0$
3. **Teststatistik**: V: Anzahl X_i's mit $(X_i > \mu_0)$
   Verteilung der Teststatistik unter $H_0 : \ V \sim \text{Bin}(12, 0.5)$
4. **Signifikanzniveau**: $\alpha = 0.05$

5. **Verwerfungsbereich für die Teststatistik**: $K = [0, c]$ mit $c = \max\{v : \ P(V \leq v) \leq \alpha\}$.
   Es gilt $P(V \leq 2) = 0.019$ und $P(V \leq 3) = 0.073$. Also $c = 2$. Diese Werte ermitteln sich mit Hilfe von `R`:

   ```
   > pbinom(3,12,p=0.5)
   [1] 0.07299805
   ```

6. **Testentscheid**: $v = 7$; $7 \notin K$; $\rightarrow H_0$ beibehalten.

(**b**) Wir führen einen einseitigen Wilcoxon-Test mit Nullhypothese $\mu_0 = 70$ cl mit `R` durch

```
> wilcox.test(c(71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72),mu=70,
+         alternative = "less")

        Wilcoxon signed rank test with continuity correction

data:  c(71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72)
V = 44.5, p-value = 0.6838
alternative hypothesis: true location is less than 70
```

Der p-Wert beträgt also 0.6838, woraus wir schliessen, dass die Nullhypothese belassen werden kann.

Aufgabe 5

(**a**) **Gepaarte Stichprobe**: Zu jeder Blutplättchenmenge vor dem Rauchen gehört die Blutplättchenmenge derselben Person nach dem Rauchen.

Einseitiger Test: Wir wollen nicht wissen, ob sich die Bluplättchenmenge *verändert* hat, sondern ob sie sich *erhöht* hat.

H_0: Rauchen hat keinen Einfluss auf die Anhäufung der Blutplättchen. ($\mu_R = \mu_{NR}$)

H_A: Durch Rauchen erhöht sich die Anhäufung der Blutplättchen. ($\mu_R > \mu_{NR}$)

(**b**) **Gepaarte Stichprobe**: Zu jeder Höhe eines selbstbefruchteten Setzlings gehört die Höhe des fremdbefruchteten "Partners".

Einseitiger Test: Wir wollen nicht wissen, ob sich die Höhen *unterscheiden*, sondern ob die fremdbefruchteten Setzlinge *grösser* werden als die selbstbefruchteten.

H_0: Die Höhen unterscheiden sich nicht.($\mu_f = \mu_s$)

H_A: Fremdbefruchtete Setzlinge werden grösser als selbstbefruchtete. ($\mu_f > \mu_s$)

(**c**) **Ungepaarte Stichprobe**: Ungleiche Anzahl in den Gruppen. Zu einem Blutdruck aus der Versuchsgruppe gehört nicht ein spezifischer aus der Kontrollgruppe.

Zweiseitiger Test: Wir wollen nur wissen, ob das Kalzium einen Einfluss hat auf den Blutdruck, *egal* ob nach oben oder unten.

H_0: Kalzium hat keinen Einfluss auf den Blutdruck. ($\mu_{\text{Kalz}} = \mu_{\text{Kontr}}$)

H_A: Kalzium hat einen Einfluss auf den Blutdruck. ($\mu_{\text{Kalz}} \neq \mu_{\text{Kontr}}$)

(**d**) **Ungepaarte Stichprobe**: Die Anzahlen in den beiden Gruppen brauchen nicht gleich zu sein. Zur Eisenmessung einer "Fe^{2+}-Maus" gehört nicht eine bestimmte Messung einer "Fe^{3+}-Maus".

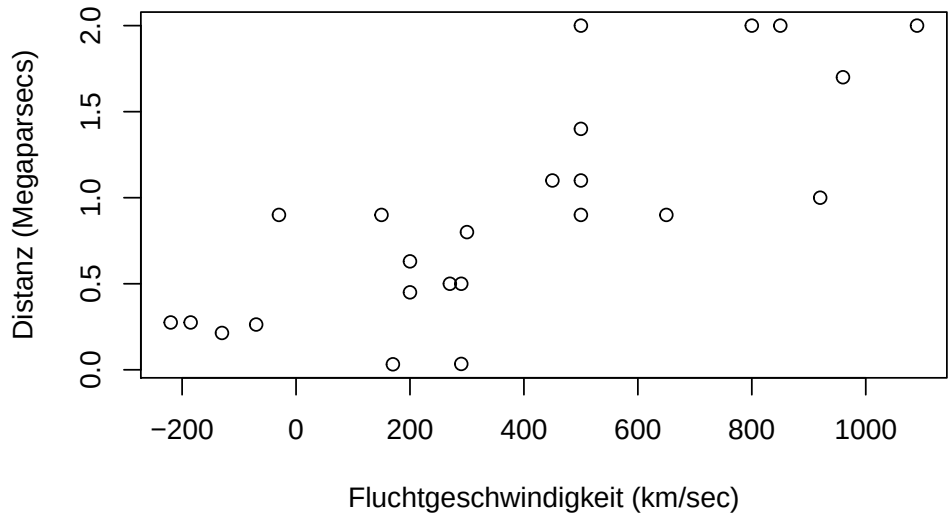
Zweiseitiger Test: Wir wollen nur wissen, ob die Mäuse die verschiedenen Eisenformen *unterschiedlich* gut aufnehmen.

H_0: Die Eisenaufnahme ist von der Form unabhängig. ($\mu_2 = \mu_3$)

H_A: Die Eisenaufnahme ist von der Form abhängig. ($\mu_2 \neq \mu_3$)

Aufgabe 6

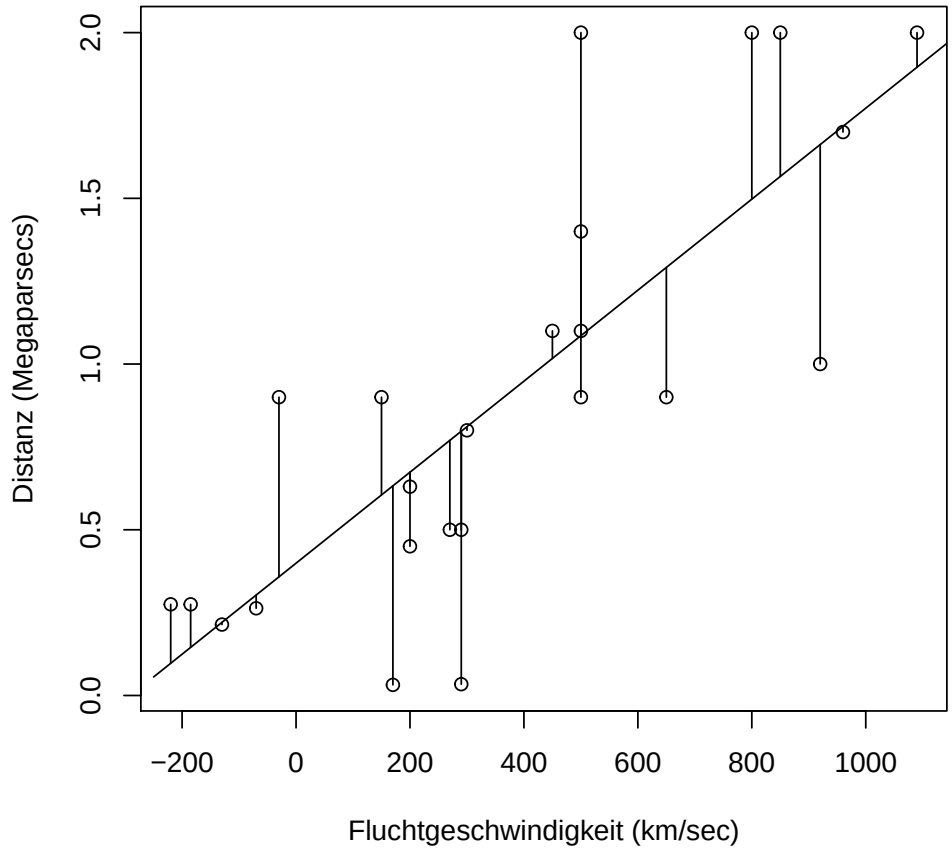
(**a**)
```
> recession.velocity <- c(170,290,-130,-70,-185,-220,200,290,270,200,300,-30,
+         650,150,500,920,450,500,500,960,500,850,800,1090)
> distance <- c(0.032,0.034,0.214,0.263,0.275,0.275,0.450,0.500,
+         0.500,0.630,0.800,0.900,0.900,0.900,0.900,1.000,1.100,1.100,
+         1.400,1.700,2.000,2.000,2.000,2.000)
> plot(recession.velocity, distance, ylab="Distanz (Megaparsecs)",
+         xlab="Fluchtgeschwindigkeit (km/sec)")
```



(**b**) Die Parameter β_0 und β_1 lassen sich mit der Methode der kleinsten Quadrate schätzen, wobei wir folgende Formeln für die Parameterschätzungen erhalten hatten:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\,. \end{aligned}$$

```
> beta_1 <- sum((distance-mean(distance))*
+                 (recession.velocity-mean(recession.velocity)))/
+         (sum((recession.velocity-mean(recession.velocity))^2))
> beta_1

[1] 0.001372936

> beta_0 <- mean(distance)-beta_1*mean(recession.velocity)
> beta_0

[1] 0.3990982

> plot(recession.velocity, distance, ylab="Distanz (Megaparsecs)",
+         xlab="Fluchtgeschwindigkeit (km/sec)")
> lines(-250:1200,beta_0+beta_1*(-250:1200),type="l",new=TRUE)
> segments(recession.velocity,beta_0+beta_1*(recession.velocity),
+         recession.velocity,distance)
```



β_1 kann auch bestimmt werden mit der folgenden Beziehung

$$\hat{b} = \frac{s_{xy}}{s_x^2} \, .$$

```
> beta_1 <- cov(recession.velocity,distance)/var(recession.velocity)
> beta_1

[1] 0.001372936
```

Vorbesprechung: 15/16. Mai 2013

Aufgabe 1

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

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

Kennzahlen für die Differenz: $\bar{x}$ beträgt -5.78, die Standardabweichung $s = 6.2$.

Es wird vermutet, dass Gerät B systematisch grössere Werte misst. Bestätigen die Messwerte diese Vermutung oder ist eine zufällige Schwankung als Erklärung plausibel?

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

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

(**c**) Sei Z die Zufallsvariable, die zählt, bei wie vielen der 9 Messungen Gerät A einen grösseren Wert misst, als Gerät B. Wie ist Z verteilt, wenn die Geräte bis auf Zufallsschwankungen das Gleiche messen?

Aufgabe 2

In der folgenden Tabelle sind die Kieferlängen von 10 männlichen und 10 weiblichen Goldschakalen eingetragen:

männlich x_i	120	107	110	116	114	111	113	117	114	112
weiblich y_j	110	111	107	108	110	105	107	106	111	111

Einige Kennzahlen: $\bar{x} = 113.4$, $\bar{y} = 108.6$, $s_x^2 = 13.82$, $s_y^2 = 5.16$

(**a**) Handelt es sich um gepaarte oder ungepaarte Stichproben? Begründe!

(**b**) Unterscheiden sich die Kieferlängen von Männchen und Weibchen signifikant? Führen Sie von Hand einen vollständigen Zwei-Stichproben t-Test durch. Geben Sie Null- und Alternativhypothese, Teststatistik, kritischen Wert resp. Verwerfungsbereich sowie Testentscheid explizit an! Muss der Test ein- oder zweiseitig durchgeführt werden?

(**c**) Führen Sie den t-Test nun noch mit Hilfe von **R** durch. Geben Sie den resultierenden p-Wert sowie den daraus folgenden Testentscheid an.

```
>jackals <- read.table("http://stat.ethz.ch/Teaching/Datasets/jackals.dat",
            header=TRUE) # Datensatz einlesen
>jackals # Datensatz anschauen
>t.test(jackals[,"M"], jackals[,"W"])    # t-Test durchfuehren
```

Bemerkung: Die Anzahl Freiheitsgrade, die im **R**-Output angegeben werden, ist anders, als die von Ihnen in b) verwendete. Dies liegt daran, dass **R** nicht davon ausgeht, dass die Varianzen in beiden Gruppen gleich gross sind und deshalb der t-Test leicht anders gerechnet wird als in der Vorlesung gelernt. Wenn **R** dasselbe machen soll wie Sie von Hand, dann muss noch ein Argument gesetzt werden:

```
>t.test(jackals[,"M"], jackals[,"W"], var.equal = TRUE)
```

(**d**) Führe mit Hilfe von **R** einen Wilcoxon-Test durch. Gib wiederum p-Wert und Testentscheid an.

```
> wilcox.test(jackals[,"M"], jackals[,"W"],) # Wilcoxon-Test durchfuehren
```

(**e**) Falls die Resultate der beiden Tests unterschiedlich ausgefallen wären, welchem würden Sie eher vertrauen? Weshalb?

Aufgabe 3

Der Sportschuh-Hersteller Hypatia AG, offizieller Ausrüster des syldawischen olympischen Teams, will das beste Material für die bevorstehenden olympischen Spiele zur Verfügung stellen. Dazu lässt der Hersteller die neuesten beiden Schuh-Kreationen, "SpeedShoe" und "Lightning", von zehn syldawischen Athleten auf einer 400m-Bahnrunde testen, wobei jeder Athlet zuerst mit dem einen (zufällig gewählten) Modell läuft und dann mit dem anderen. Die Kennzahlen der gemessenen Zeiten sind wie folgt:

SpeedShoe	$\bar{x} = 46.02$	$\hat{\sigma}_x = 1.56$
Lightning	$\bar{y} = 46.24$	$\hat{\sigma}_y = 1.52$
Differenz	$\bar{x} - \bar{y} = -0.22$	$\hat{\sigma}_{x-y} = 0.26$

Sie dürfen davon ausgehen, dass die Rundenzeiten durch unabhängige $\mathcal{N}(\mu_x, \sigma_x^2)$- resp. $\mathcal{N}(\mu_y, \sigma_y^2)$-verteilte Zufallsvariablen beschrieben werden können.

Es soll nun getestet werden, ob das Modell "SpeedShoe" zu besseren Leistungen verhilft als das Modell "Lightning".

(**a**) Handelt es sich hier um einen gepaarten oder einen ungepaarten Test? Begründen Sie kurz.

(**b**) Führen Sie einen einseitigen t-Test auf dem 5%-Niveau durch.

(**c**) Geben Sie ein einseitiges 95%-Vertrauensintervall an für die Differenz $\mu_x - \mu_y$.

Aufgabe 4

Wir bohren ein Loch in einen Permafrostboden. In den Tiefen von 0, 0.2, 0.5, 0.6, 0.8, 0.9, 1.2 und 6m messen wir die Bodentemperatur. Diese könnte im Sommer folgende Werte haben: 6, 4.2, 0.6, -2.1, -5.2, -7.3, -8.9 und 15°C.
Als Hilfe sind bei den Unteraufgaben die **R** Befehle angegeben.

R-*Hinweise*: jeweils .. durch korrekte Werte ersetzen!

(**a**) Zeichnen Sie ein Streudiagramm der Temperatur in Abhängigkeit der Tiefe. Was fällt Ihnen bei diesen Daten auf? Geben Sie (zwei) mögliche Interpretationen der Daten.

```
> tiefe <- c(..); temp <- c(..)
> plot(tiefe, temp, main="Streudiagramm")
```

(**b**) Berechnen Sie die empirische Korrelation der Daten ohne den Ausreisser und vergleichen Sie sie mit derjenigen aller Daten ($\rho = 0.6$).

```
> tiefeoA <- tiefe[-..];   tempoA <- temp[-..] # 1 Beobachtung weglassen
> cor(..,..)
```

(**c**) Passen Sie eine Gerade an die Daten an. (Schätzen Sie den Achsenabschnitt β_0 und die Steigung β_1 nach der Methode der kleinsten Quadrate und zeichnen Sie die Gerade in das Streudiagramm ein.) Lassen Sie dann den Ausreisser weg und schätzen Sie die Gerade mit den übrigen Daten.

```
> fit1 <- lm(temp ~ tiefe) # Lineares Modell anpassen
> fit2 <- lm(tempoA ~ tiefeoA) # Lineares Modell anpassen ohne Ausreisser
> ## Regressionsgerade fuer fit1 zeichnen mit gestrichelter Linie
> abline(fit1,lty=2)
> abline(fit2) # Regressionsgerade fuer fit2 zeichnen
```

Aufgabe 5

In dieser Aufgabe betrachten wir 4 Datensätze, die von Anscombe konstruiert wurden. In jedem der Datensätze gibt es eine Zielvariable Y und eine erklärende Variable X.

(**a**) Stellen Sie jeden der 4 Datensätze als Streudiagramm dar, zeichnen Sie die Regressionsgerade ein und kommentieren Sie die Ergebnisse.

(**b**) Vergleichen Sie die Schätzungen von β_0, β_1 und σ^2, sowie das sogenannte "Gütemass" R^2, das später genauer besprochen wird.

R-Hinweise:

```
data(anscombe) ## Einlesen des Datensatzes
```

Die Regression kann man mit

```
reg <- lm(y1~x1, data = anscombe) oder
reg <- lm(anscombe$y1 ~ anscombe$x1)
summary(reg)
```

berechnen und numerisch auswerten. Mit `par(mfrow=c(2,2))` wird das Grafikfenster so eingeteilt, dass alle 4 Bilder nebeneinander passen. Den Scatterplot und die Regressionsgerade erhält man mit

```
plot(anscombe$x1, anscombe$y1)
abline(reg)
```

Die Schätzungen für die Koeffizienten β_0, β_1 und σ, sowie das Gütemass R^2 erhält man mit

```
summary(reg)
```

Aufgabe 1

(**a**) Es handelt sich um **gepaarte** Stichproben. Am gleichen Ort wird mit beiden Geräten gemessen.

(**b**)
- **Modell:** $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ wird durch S geschätzt.
- **Nullhypothese:** $H_0: \ \mu = \mu_0 = 0$.
  Alternative: $H_A: \ \mu < \mu_0$.
- **Teststatistik:**
$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S}$$
  Verteilung der Teststatistik unter H_0: $T \sim t_{n-1}$
- **Signifikanzniveau:** $\alpha = 5\%$
- **Verwerfungsbereich:**
$$K = (-\infty, -t_{9-1,0.95}] = (-\infty, -1.86]$$
- **Testentscheid:**
$$t = \frac{\bar{x} - 0}{S/\sqrt{9}} = -2.8$$

Der Wert t der Teststatistik liegt im Verwefungsbereich, d.h. eine neue Eichung der Geräte ist angezeigt.

(**c**) Z wäre binomialverteilt mit Parametern $n = 9$ und $p = 1$. Darauf aufbauend kann man auch einen Test durchführen (man spricht vom sogennanten Vorzeichentest). Der Vorteil ist, dass man keine Normalverteilung mehr annehmen muss.

Aufgabe 2

(**a**) Es handelt sich um ungepaarte Stichproben, da zu den einzelnen Männchen nicht jeweils ein bestimmtes Weibchen gehört. Die Anzahlen in den beiden Stichproben brauchen auch gar nicht gleich gross zu sein.

(**b**) **Zweiseitiger t-Test:**

X_i: i-ter Wert der Kieferlänge der männlichen Tiere, $i = 1, \ldots, n = 10$

Y_j: j-ter Wert der Kieferlänge der weiblichen Tiere, $j = 1, \ldots, m = 10$

Nullhypothese H_0:	X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, Y_j i.i.d. $\mathcal{N}(\mu, \sigma^2)$ unabhängig
Alternative H_A:	$X_i \sim (\mu_1, \sigma^2)$, $Y_j \sim \mathcal{N}(\mu_2, \sigma^2)$ mit $\mu_1 \neq \mu_2$
Teststatistik:	$T = (\bar{X} - \bar{Y})/s_{\bar{X}-\bar{Y}}$, wobei
	$s_{\bar{X}-\bar{Y}} = \sqrt{(\frac{1}{n} + \frac{1}{m})\frac{1}{n+m-2} \cdot ((n-1)s_x^2 + (m-1)s_y^2)}$
	Unter H_0 gilt: $T \sim t_{n+m-2}$, also hier $T \sim t_{18}$
Verwerfungsbereich:	Tabelle: $t_{18,0.975} = 2.1$ (Test zweiseitig auf 5%-Niveau)
	somit: $\mathcal{K} = \{\|T\| > t_{18,0.975}\} = \{\|T\| > 2.1\}$
Wert der Teststatistik:	$s_{\bar{X}-\bar{Y}} = \sqrt{\frac{2}{10}\frac{1}{18} \cdot (9 \cdot 13.82 + 9 \cdot 5.19)} = 1.38$
	$T = \frac{113.4-108.6}{1.38} = 3.48$

Entscheidung: Da $T \in \mathcal{K}$ ("T Element des Verwerfungsbereichs"), wird die Nullhypothese H_0 auf dem 5%-Niveau durch den t-Test **verworfen**.

(**c**) Der **R**-Output für den t-Test sieht folgendermassen aus:

```
> jackals<-read.table("http://stat.ethz.ch/Teaching/Datasets/jackals.dat",header=TRUE)
> t.test(jackals[,"M"],jackals[,"W"])

        Welch Two Sample t-test

data:  jackals[, "M"] and jackals[, "W"]
t = 3.4843, df = 14.894, p-value = 0.00336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.861895 7.738105
sample estimates:
mean of x mean of y
    113.4     108.6
```

Der p-Wert ist $0.0034 < 0.05$, also wird die Nullhypothese verworfen.

(**d**) Der **R**-Output für den Wilcoxon-Test sieht folgendermassen aus:

```
> wilcox.test(jackals[,"M"],  jackals[,"W"])

        Wilcoxon rank sum test with continuity correction

data:  jackals[, "M"] and jackals[, "W"]
W = 87.5, p-value = 0.004845
alternative hypothesis: true location shift is not equal to 0
```

Der p-Wert ist $0.0048 < 0.05$, also wird auch bei diesem Test die Nullhypothese verworfen.

(**e**) Das Resultat des Wilcoxon-Tests ist vertrauenswürdiger, da er im Gegensatz zum t-Test nicht annimmt, dass die Daten normalverteilt sind und wir diese Voraussetzung in keiner Weise überprüft haben. Allerdings ist die stark unterschiedliche Standardabweichung in den zwei Gruppen problematisch für beide Tests.

Aufgabe 3

(**a**) Der Test ist gepaart, da zu jedem Läufer genau ein Wert mit jedem Schuhtyp vorliegt und die beiden Werte eines Läufers von dessen Lauffähigkeiten abhängen.

(**b**) **1. Modell:** $X_1 - Y_1, \ldots, X_n - Y_n$ i.i.d. $\mathcal{N}(\mu_{X-Y}, \sigma^2_{X-Y})$, σ_{X-Y} wird durch $\widehat{\sigma_{x-y}}$ geschätzt.

2. Nullhypothese: $H_0: \ \mu_{X-Y} = \mu_0 = 0$.
Alternative: $H_A: \ \mu_{X-Y} < 0$

3. Teststatistik:

$$T = \frac{\sqrt{n}(\bar{X}_n - \bar{Y}_n - \mu_0)}{\widehat{\sigma_{X-Y}}}$$

Verteilung der Teststatistik unter H_0: $T \sim t_{n-1}$.

4. Signifikanzniveau: $\alpha = 5\%$.

5. Verwerfungsbereich für die Teststatistik:

$$K = (-\infty, -t_{9;0.95}) = (-\infty, -1.833)$$

6. Testentscheid:

$$t = \frac{\sqrt{n}(\bar{x}_n - \bar{y}_n)}{\hat{\sigma}_{x-y}} = \frac{\sqrt{10}(46.02 - 46.24)}{0.26} = -2.68$$

Der Wert t der Teststatistik liegt im Verwefungsbereich, d.h. "SpeedShoe" verhilft zu besseren Leistungen.

(**c**) Das einseitige 95%-Vertrauensintervall beinhaltet alle Werte von $\mu_x - \mu_y$, bei denen der entsprechende Test die Nullhypothese auf dem Signifikanzniveau 5% aufgrund des beobachteten Wertes $\overline{x} - \overline{y}$ nicht verwirft. Beim oben betrachteten einseitigen t-Test hat der Verwerfungsbereich die Form $K = (-\infty, -t_{9;0.95})$. Der t-Test verwirft H_0 nicht, wenn der Wert der Teststatistik nicht im Verwerfungsbereich der Teststatistik ist. Wenn H_0 auf dem Signifikanzniveau $\alpha = 5\%$ nicht verworfen wird, muss also gelten

$$-t_{9;0.95} \leq \frac{\sqrt{n}(\bar{x}_n - \bar{y}_n - (\mu_x - \mu_y))}{\widehat{\sigma_{x-y}}} \, .$$

Somit ist

$$\mu_x - \mu_y \leq \bar{x}_n - \bar{y}_n + \frac{\widehat{\sigma_{x-y}} t_{9;0.95}}{\sqrt{n}}$$

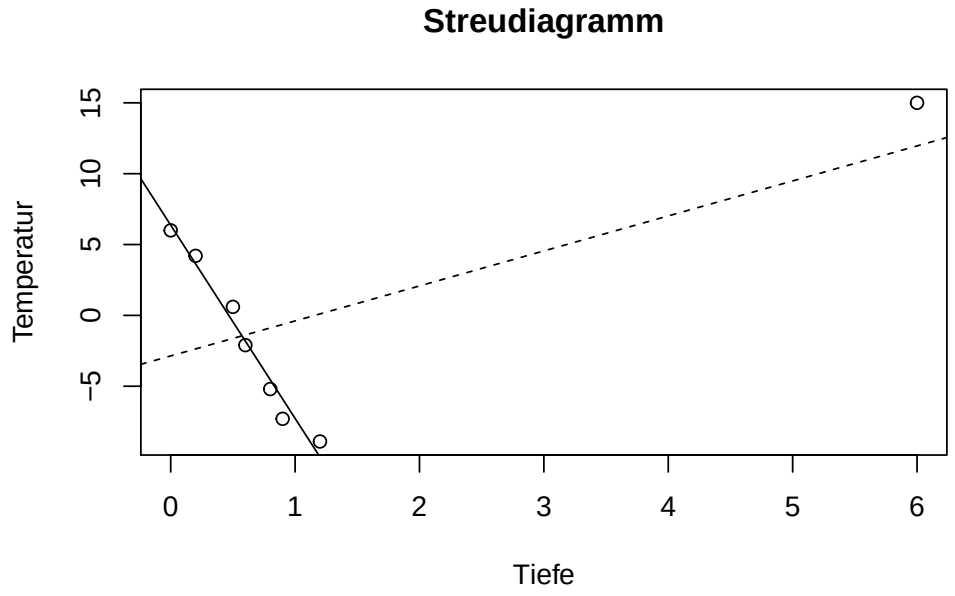
Das einseitige 95%-Vertrauensintervall für $\mu_x - \mu_y$ ist also

$$\begin{aligned} &(-\infty, \bar{x}_n - \bar{y}_n + \hat{\sigma}_{x-y} \cdot t_{9;0.95}/\sqrt{10}) \\ =&(-\infty, -0.22 + 0.26 \cdot 1.833/\sqrt{10}) = (-\infty, -0.069) \end{aligned}$$

Aufgabe 4

(**a**) Aus dem Streudiagrammm sieht man, dass die ersten 7 Punkte sehr gut durch eine Gerade beschrieben werden. Der letzte Punkt liegt hingegen völlig ausserhalb der Geraden.

```
> tiefe <-  c(0,0.2,0.5,0.6,0.8,0.9,1.2,6);
> temp  <-  c(6, 4.2, 0.6,-2.1, -5.2, -7.3,-8.9,15)
> tiefe2 <-  c(0,0.2,0.5,0.6,0.8,0.9,1.2);
> temp2  <-  c(6, 4.2, 0.6,-2.1, -5.2, -7.3,-8.9)
> plot(tiefe,  temp,  main="Streudiagramm",ylab="Temperatur",xlab="Tiefe")
> abline(lm(temp~tiefe),lty=2)
> abline(lm(temp2~tiefe2),lty=1)
```



Es kann sich um einen groben Fehler handeln; es ist aber auch möglich, dass das lineare Modell zur Beschreibung des Zusammenhangs zwischen Tiefe und Temperatur nicht geeignet ist. (Zum Beispiel könnte der Zusammenhang quadratisch sein, oder stückweise linear. Weitere Möglichkeiten sind denkbar, z.B. Hinweis auf heisse Quelle, etc.)

(**b**) Sei X die Tiefe und Y die Temperatur. Empirische Korrelation:

$$\rho_{XY} = \frac{\sum_{i=1}^{7}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{7}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{7}(y_i - \bar{y})^2}} = -0.99$$

wobei $\bar{x}$ und $\bar{y}$ auch ohne den Ausreisser zu berechnen sind.
Ohne Ausreisser sind Tiefe und Temperatur sehr stark negativ korreliert, mit dem Ausreisser aber positiv (0.6)!

(**c**) In der obigen Figur sind die beiden Regressionsgeraden exakt eingetragen: der Ausreisser bewirkt, dass sich die Gerade fast um 90 dreht, d.h. die kleinste Quadrate Regressionsgerade ist sehr anfällig auf Ausreisser!

Die Schätzungen der Koeffizienten lauten ohne Ausreisser:

$$\begin{aligned}\widehat{\beta_1} &= \frac{\sum_{i=1}^{7}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ &= \frac{(6+1.81)(0-0.6)+\ldots+(-8.9+1.81)(1.2-0.6)}{(0-0.6)^2+\ldots+(1.2-0.6)^2} \\ &= -13.64\end{aligned}$$

und

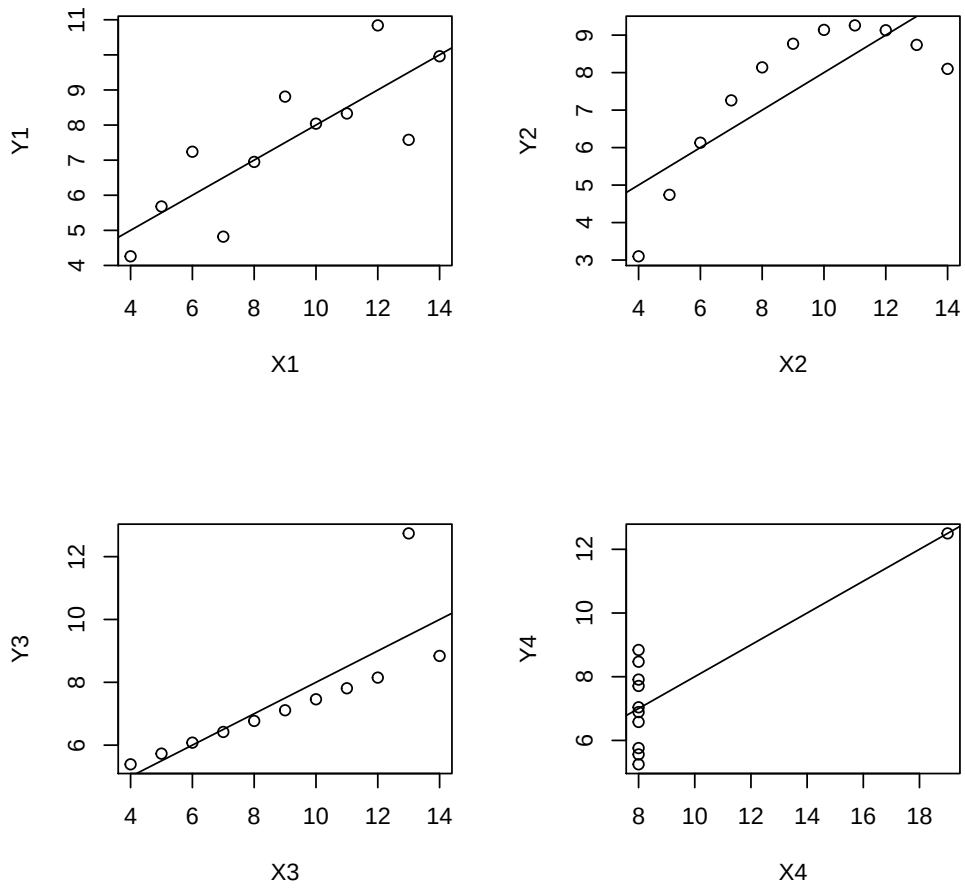
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = -1.81 - (-13.64)\cdot 0.6 = 6.37$$

Mit Ausreisser ergibt sich: $\widehat{\beta_1} = 2.47$ und $\widehat{\beta_0} = -2.86$.
Diese Schätzungen können auch im **R**-output abgelesen werden.

Aufgabe 5

(**a**) Betrachtet man die vier Streudiagramme, so sieht man, dass nur im ersten Fall eine lineare Regression korrekt ist. Im zweiten Fall ist die Beziehung zwischen X und Y nicht linear, sondern quadratisch. Im dritten Fall gibt es einen Ausreisser, welcher die geschätzten Parameter stark beeinflusst. Im vierten Fall wird die Regressionsgerade durch einen einzigen Punkt bestimmt.

```
> data(anscombe)
> reg <- lm(anscombe$y1 ~ anscombe$x1)
> reg2 <- lm(anscombe$y2 ~ anscombe$x2)
> reg3 <- lm(anscombe$y3 ~ anscombe$x3)
> reg4 <- lm(anscombe$y4 ~ anscombe$x4)
> par(mfrow=c(2,2))
> plot(anscombe$x1, anscombe$y1,ylab="Y1",xlab="X1")
> abline(reg)
> plot(anscombe$x2, anscombe$y2,ylab="Y2",xlab="X2")
> abline(reg2)
> plot(anscombe$x3, anscombe$y3,ylab="Y3",xlab="X3")
> abline(reg3)
> plot(anscombe$x4, anscombe$y4,ylab="Y4",xlab="X4")
> abline(reg4)
```



(**b**) Bei allen vier Modellen sind die Schätzungen des Achsenabschnitts β_0, der Steigung β_1 und der Fehlervarianz σ^2, sowie das Gütemass R^2 fast identisch:

	Modell 1	Modell 2	Modell 3	Modell 4
Achsenabschnitt ($\widehat{\beta_0}$)	3.000	3.001	3.002	3.002
Steigung ($\widehat{\beta_1}$)	0.500	0.500	0.500	0.500
$\hat{\sigma}^2$	1.529	1.531	1.528	1.527
R^2	0.667	0.666	0.666	0.667

Fazit: Es genügt **nicht**, nur $\widehat{\beta}_0, \widehat{\beta}_1, \sigma$ und R^2 anzuschauen. In allen Modellen sind diese Schätzungen fast gleich, aber die Datensätze sehen ganz unterschiedlich aus. Eine (graphische) Überprüfung der Modellannahmen ist also unumgänglich.

Vorbesprechung: 22/23. Mai 2013

Aufgabe 1

Der Datensatz von Forbes zeigt Messungen von Siedepunkt (in $°F$) und Luftdruck (in inches of mercury) an verschiedenen Orten in den Alpen. Die Daten stehen als Datensatz `forbes.dat` mit den Variablen `Temp` und `Press` zur Verfügung.

(**a**) Tragen Sie in einem Streudiagramm den Druck gegen die Temperatur auf. Macht es Sinn, diese Daten mit einer Regressionsgeraden zu modellieren?
R-Anleitung:

```
> forbes <- read.table("http://stat.ethz.ch/Teaching/Datasets/forbes.dat",header=TRUE)
> par(mfrow = c(3,1)) # Ermoeglicht 3 Graphiken untereinander zu platzieren.
> plot(forbes[,"Temp"], forbes[,"Press"])
```

(**b**) Berechnen Sie die Koeffizienten der Regressionsgeraden und tragen Sie die Regressionsgerade ins Streudiagramm ein.

```
> forbes.fit <- lm(Press ~ Temp, data = forbes) #Regression berechnen
> summary(forbes.fit) # Regressionsoutput zeigen
> abline(forbes.fit)  # Regressionsgerade einzeichnen
```

(**c**) Zeichnen Sie den Tukey-Anscombe-Plot (Residuen gegen angepasste Werte) und den Normalplot der Residuen. Gibt es Hinweise, dass die Modellannahmen verletzt sind?

```
> plot(fitted(forbes.fit), resid(forbes.fit), main="Tukey-Anscombe Plot")
> abline(h=0)
> qqnorm(resid(forbes.fit))
```

(**d**) Logarithmieren Sie nun den Druck. Tragen Sie in einem Streudiagramm den logarithmierten Druck gegen die Temperatur auf, berechnen Sie die Regressionsgerade und tragen Sie sie ins Diagramm ein.

```
> forbes[,"Logpress"] <- log(forbes[,"Press"])
```

(**e**) Zeichnen Sie wiederum den Tukey-Anscombe und den Normalplot. Wie steht es nun mit den Modellannahmen? Gibt es Ausreisser?

(**f**) Identifizieren Sie und entfernen Sie den Ausreisser. Berechnen Sie die Regressionsgerade neu und zeichnen Sie nochmals alle Plots. Sind jetzt die Modellvoraussetzungen erfüllt? Ein Ausreisser ist eine Beobachtung, die nicht in das Modell passt (z.B. wegen Tippfehler). Ausreisser identifizieren Sie mit Hilfe des Befehls `identify()`: Dazu schliesse man zuerst alle Graphikfenster. Nach Ausführung des identify Befehls (wie unten beschrieben) mit der linken Maustaste auf den Ausreisser klicken, dann erscheint die Nummer des Ausreissers. **R** fährt nach dem identify Befehl erst weiter, wenn dieser mittels Klicken der mittleren Maustaste in der Graphik beendet worden ist.

```
> plot(fitted(forbes.fit), resid(forbes.fit))
> identify(fitted(forbes.fit), resid(forbes.fit))
> forbes <- forbes[-..,] # Ausreisser entfernen: .. mit Beobachtungsnummer ersetzen
```

Aufgabe 2

In der folgenden Tabelle stehen die Weltrekorde der Männer über 13 verschiedene Laufdistanzen, Stand 1974.

Distanz (m)	100	200	400	800	1000	1500	2000
Zeit (s)	9.9	19.8	43.8	103.7	136	213.1	296.2

Distanz (m)	3000	5000	10000	20000	25000	30000
Zeit (s)	457.6	793.0	1650.8	3464.4	4495.6	5490.4

An diese Daten wurde folgendes Regressionsmodell angepasst:

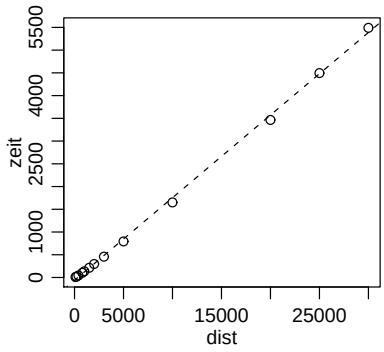
$$\text{Zeit}_i = \beta_0 + \beta_1 \cdot \text{Distanz}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

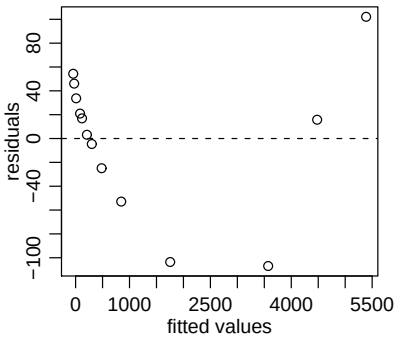
Der Regressionsoutput und die Diagnoseplots sehen folgendermassen aus:

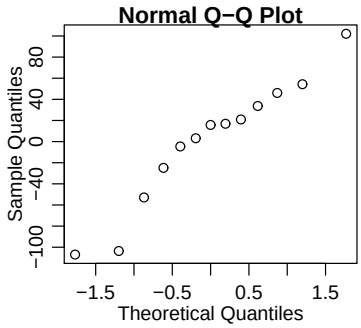
```
Call:
lm(formula = zeit ~ dist)

Residuals:
Min           1Q  Median    3Q    Max
-106.95   -24.90   15.77 33.71 102.08

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -62.59296 21.81098 -2.87 0.0152 *
dist         0.18170  0.00173 105.05 <2e-16 ***

Residual standard error: 62.68 on 11 degrees of freedom
Multiple R-squared: 0.999,Adjusted R-squared: 0.9989
F-statistic: 1.103e+04 on 1 and 11 DF, p-value: < 2.2e-16

Residuals:
    Min      1Q  Median     3Q     Max
-106.95  -24.90   15.77  33.71  102.08
```

(**a**) Gibt es einen signifikanten Zusammenhang zwischen Distanz und Zeit, d.h. ist β_1 signifikant von 0 verschieden?

(**b**) Eines der folgenden 4 Intervalle ist das 95%-Vertrauensintervall für β_1. Welches?







i) [0.1800, 0.1834]

ii) [0.1779, 0.1855]

iii) [0.1765, 0.1869]

iv) [0.1800, 0.1852]

(**c**) Wie gross ist das Residuum der 5. Beobachtung (1000m)?

(**d**) Dürfen wir die berechnete Regressionsgerade benutzen, um zu schliessen, dass 1974 der Weltrekord über 100km (100000m) ungefähr bei 18000s gelegen wäre?

(**e**) Wie gross ist die geschätzte Standardabweichung der Fehler E_i? Was heisst das für die Brauchbarkeit des Modells?

(**f**) Was folgern Sie aus der Darstellung der Residuen gegen angepasste Werte?

(**g**) Formulieren Sie ein Modell, das vermutlich besser zu diesen Daten passen würde.

Aufgabe 3

Bestimmen Sie aufgrund des Datensatzes von E. Hubble (siehe Kapitel 6.1 im Skript) das Alter des Universums. Geben Sie für das Alter des Universums ein 95%-Vertrauensintervall an. Berücksichtigen Sie dabei, dass die Altersbestimmung des Universums aufgrund der Urknall-Theorie davon ausgeht, dass zum Zeitpunkt des Urknalls die Distanz Erde zu galaktischen Nebeln null ist (die Urknall-Theorie beinhaltet also bloss einen unbekannten Parameter). Eine megaparsec-Sekunde pro Kilometer entspricht etwa 979.8 Milliarden Jahren.
R-Hinweis:
Lineare Regression mit Achsenabschnitt bei 0, also mit $\beta_0 = 0$:

```
>lm(y ~  0 + x)
```

Aufgabe 4

Wir betrachten einen betrunkenen Spaziergänger, der an einem Punkt x_0 in einer engen langen Gasse seinen Heimweg antritt. Er setzt seinen Fuss eine Schrittweite X_1 von x_0 entfernt hin (entweder links oder rechts von x_0).Wir fassen X_1 als eine Zufallsvariable mit Erwartungswert

μ und Standardabweichung σ auf; somit ist die Position des Spaziergängers nach einem Schritt ebenfalls eine Zufallsvariable und gegeben durch

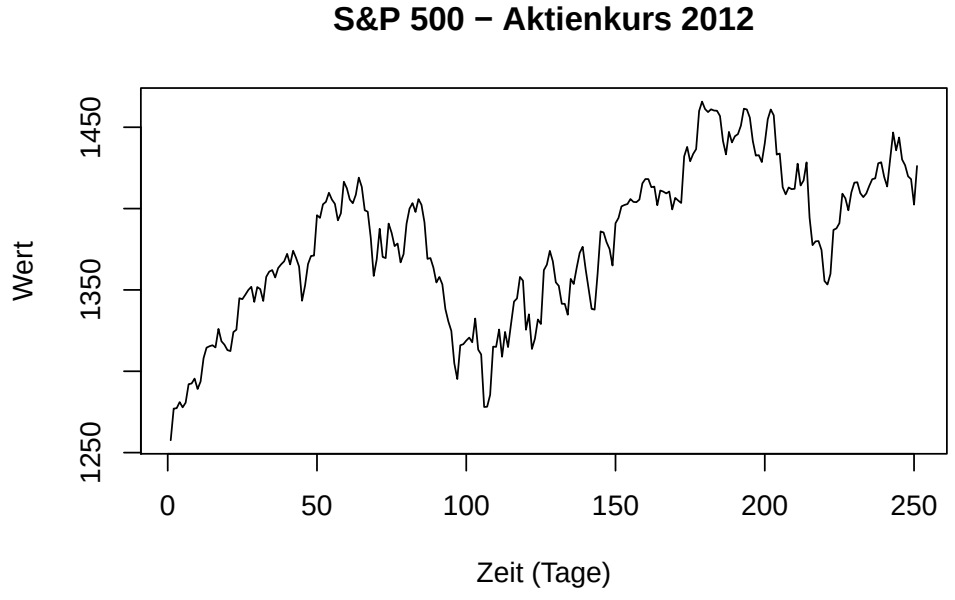
$$S(1) = x_0 + X_1 \, .$$

Nach einem weiteren Schritt der Länge X_2 mit demselben Erwartungswert μ und Standardabweichung σ ist die Position des Spaziergängers $S(2) = x_0 + X_1 + X_2$. Nach n Schritten ist die Position gegeben durch

$$S(n) = x_0 + \sum_{i=1}^{n} X_i \, .$$

(**a**) Bestimmen Sie den Erwartungswert und die Varianz der Position des betrunkenen Spaziergängers nach n Schritten, also $\mathrm{E}(S(n))$ und $\mathrm{Var}(S(n))$. Wie interpretieren Sie den Erwartungswert und die Varianz von $S(n)$?

(**b**) Das Beispiel des betrunkenen Spaziergängers ist ein Beispiel für einen (eindimensionalen) **Random Walk**. Random Walks haben viele Anwendungen in verschiedenen Bereichen der Wissenschaft. Brownsche Bewegung stellt eine Version von Random Walk dar, wobei die Zeitvariable kontinuierlich ist und die Schrittweiten normalverteilte Zufallsvariablen sind. Der Begriff Brownsche Bewegung bezieht sich auf die 1927 vom Biologen Robert Brown durchgeführten Arbeiten, in welchen dieser die zufällige Bewegung von in Wasser suspendierten Pollen beobachtete. Einstein lieferte 1905 eine Erklärung dazu: die Zitterbewegung der Pollen wird durch Stösse mit sich zufällig bewegenden Wassermolekülen verursacht. Die Theorie der Brownschen Bewegung wurde von Louis Bachelier im Jahre 1900 in seiner Doktorarbeit "Theorie de la speculation" entwickelt, die Random Walks mit der Entwicklung von Börsenkursen in Verbindung brachte. Wenn ein Aktienkurs als Random Walk beschrieben wird, dann ist die kurzzeitige Entwicklung des Aktienwertes zufällig und nicht vorhersagbar (siehe dazu die Hypothese effizienter Märkte). In untenstehender Abbildung ist der S&P 500 Kurs (Aktienkurs von 500 führenden Unternehmen) vom Jahre 2012 - bestehend aus 251 Tagen - aufgezeichnet.



Der Werteverlauf von S&P 500 während 2012 beginnt bei einem Wert von 1257.6, die mittlere Schrittweite beträgt 0.483 und die Standardabweichung der 250 Schrittweiten beträgt 11. Simulieren Sie den Werteverlauf als einen Random Walk, indem Sie annehmen, dass die Schrittweiten normalverteilte Zufallsvariablen mit $\mu = 0.483$ und $\sigma = 11$ sind.

R-Hinweis:
Benützen Sie die Funktion `rnorm(...)`, um normalverteilte Zufallsvariablen zu generieren.

```
steps <- rnorm(...,mean=...,sd=...)
sp <- numeric(251)
sp[1] <- ...
for(i in 1:250){
sp[i+1] <- sp[i]+steps[i]
}
plot(sp,type="l",xlab="...",ylab="...", main="...")
```

Aufgabe 5

(**a**) Gegeben sei das Ensemble eines Zufallsprozesses mit in T_0 periodischen Musterfunktionen $s_i(t)$, die um die Zeitspanne ΔT_i gegenüber dem Nullpunkt der Zeitachse versetzt sind:

$$s_i(t) = s(t + \Delta T_i)$$

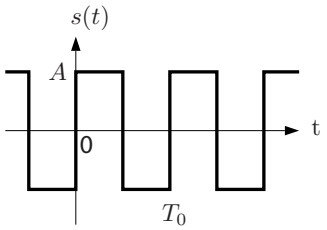
Dabei stelle $\delta = \Delta T_i$ eine kontinuierlich variierende Zufallsvariable mit der Gleichverteilung

$$p(\delta) = \begin{cases} \frac{1}{T}, & -\frac{T}{2} \leq \delta \leq \frac{T}{2} \\ 0, & \text{sonst} \end{cases}$$

dar.

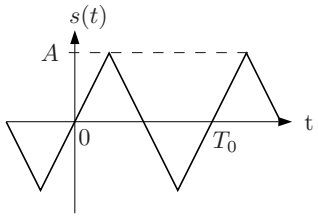
Für die folgenden drei periodischen Signale skizziere man die kumulative Verteilungsfunktion und Wahrscheinlichkeitsdichtefunktionen $F(s)$ und $f(s)$.

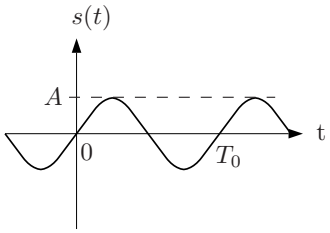
(a) Rechtecksignal



(b) Dreiecksignal

(c) Sinussignal





Hinweis:

Man ermittle in allen drei Fällen die Wahrscheinlichkeit über eine Signalperiode T_0, dass $s(t)$ kleiner oder gleich einer vorgegebenen Schwelle s ist:

$$P[s(t) \leq s] = F(s), \quad -A \leq s \leq A$$

Mit

$$f(s) \triangleq \lim_{\Delta s \to 0} \frac{P(s < \xi \leq (s + \Delta s))}{\Delta s} = \frac{\mathrm{d}F(s)}{\mathrm{d}s} \geq 0$$

gewinnt man daraus die Wahrscheinlichkeitsdichte $f(s)$.

(**b**) Man skizziere die Autokorrelationsfunktion von folgenden periodischen Signalen:

(a) Rechtecksignal

(b) Dreiecksignal

(c) Sinussignal

Aufgabe 6

Man berechne für das Beispiel in Kapitel 7.3.4 die Kurzzeit-Kreuzkorrelationsfunktion des Signals $s(t)$ mit dem Rechtecksignal $r(t)$

$$\varphi_{sr}(\tau) = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} s(t) r(t + \tau) \, \mathrm{d}\tau$$

für

(**a**) $s(t) = r(t) + n(t)$

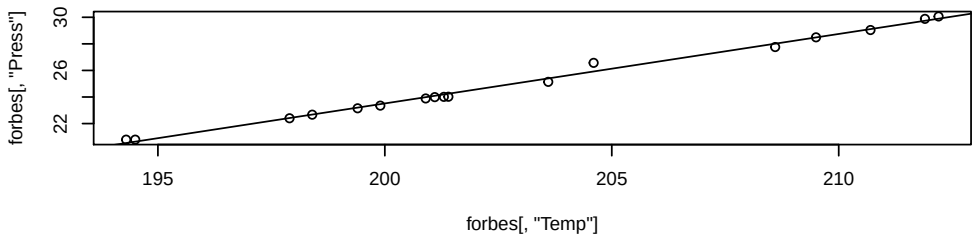
(**b**) $s(t) = n(t)$

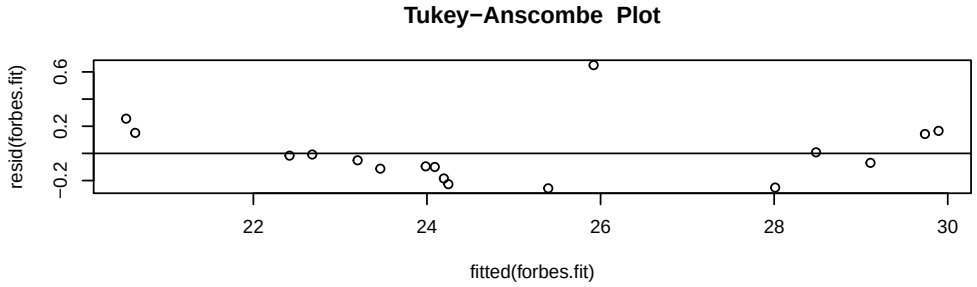
Was ist der Vorteil bei der Detektion des Rechtecksignals mittels Kreuzkorrelationsfunktion?

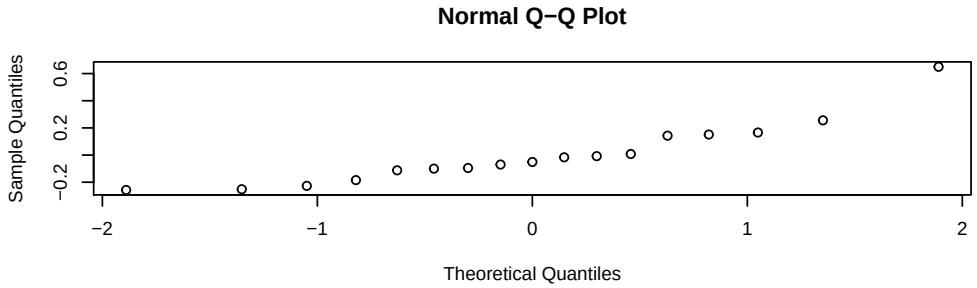
Aufgabe 1

a), b), c) Während ein Blick auf die Originaldaten suggeriert, dass sie ziemlich schön auf einer Geraden liegen, zeigt der Tukey-Anscombe-Plot, dass dem nicht so ist: die Daten weisen eine Krümmung auf. Zusätzlich fällt der Ausreisser in der Mitte oben auf, den wir auch im Normalplot rechts oben wiederfinden. Ansonsten lässt sich über den Normalplot nicht viel sagen.
Im Regressions-Output findet sich keine Hinweis darauf, dass das Modell unpassend sein könnte. Im Gegenteil könnte man meinen, dass das hohe R^2 (0.99) darauf hinweist, dass das Modell gut passt. Es ist aber eben sehr wichtig, sich nicht nur auf den Regressionsoutput zu verlassen, sondern die Daten und die Residuen anzuschauen, um Modellabweichungen festzustellen.

```
> summary(forbes.fit)

Call:
lm(formula = Press ~ Temp, data = forbes)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373    2.05182  -39.51   <2e-16 ***
Temp          0.52289    0.01011   51.74   <2e-16 ***
---
Signif. codes:  0
```

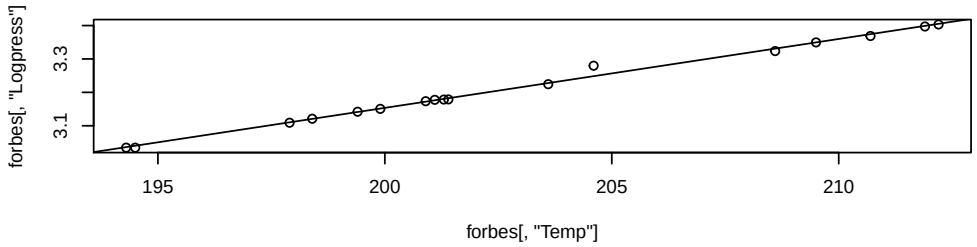


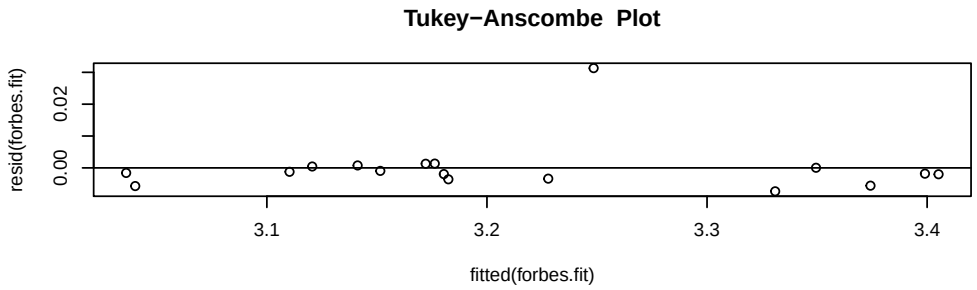


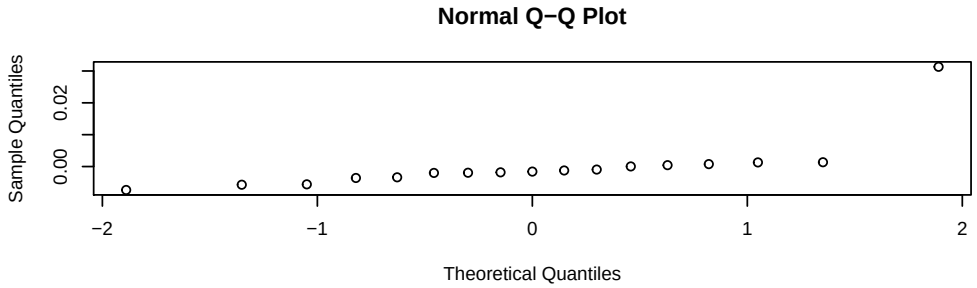


d), e) e) Die logarithmierten Werte liegen schön auf einer Geraden. Sowohl im Tukey-Anscombe wie auch im Normal-Plot ist ausser dem Ausreisser keine Abweichung von den Modellvoraussetzungen festzustellen.

```
> summary(forbes.fit)

Call:
lm(formula = Logpress ~ Temp, data = forbes)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0073622 -0.0033863 -0.0015865  0.0004322  0.0313139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9708662  0.0769377  -12.62 2.17e-09 ***
Temp         0.0206224  0.0003789   54.42  < 2e-16 ***
---
Signif. codes:  0
```

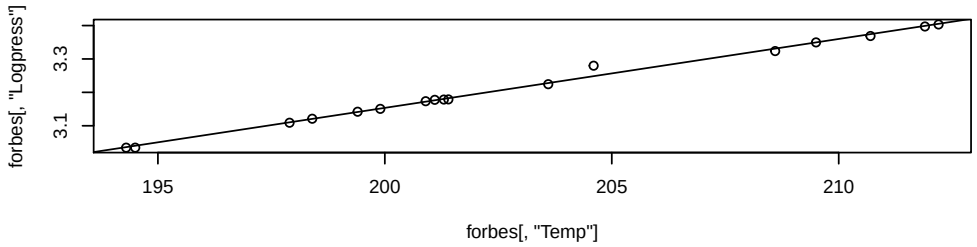


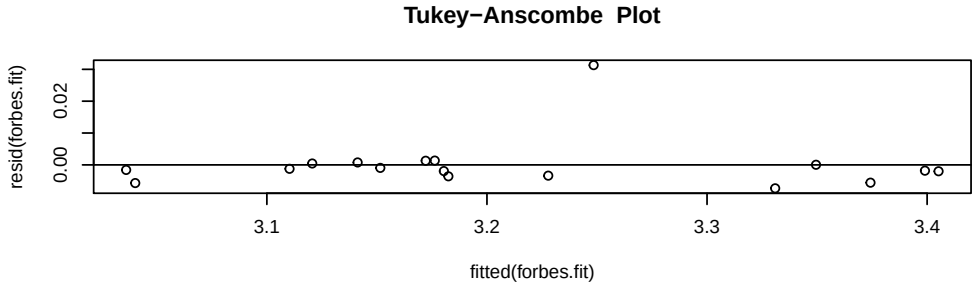


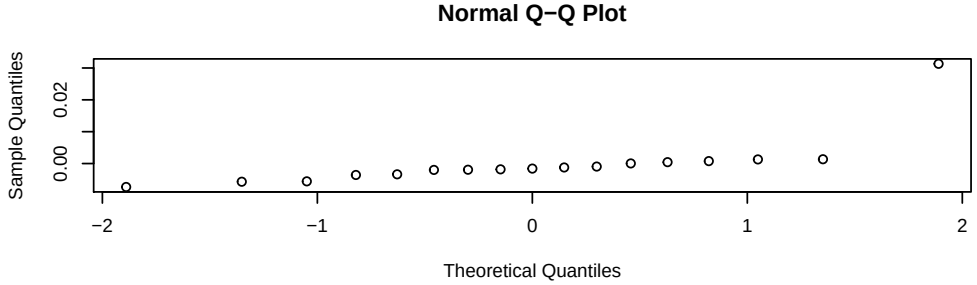


f) Nach Weglassen des Ausreissers sehen alle Plots wunderbar aus. Wie auch in den vorherigen Regressionen ist die erklärende Variable hochsignifikant.

```
> summary(forbes.fit)

Call:
lm(formula = Logpress ~ Temp, data = forbes[-12, ])

Residuals:
       Min         1Q     Median         3Q        Max
-0.0048082 -0.0014595  0.0004546  0.0020358  0.0031219

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9517662  0.0231021   -41.2 5.16e-16 ***
Temp         0.0205186  0.0001138   180.2  < 2e-16 ***
---
Signif. codes:  0
```







Aufgabe 2

(**a**) Der p-Wert ist extrem klein ($< 2 \cdot 10^{-16}$), also ist β_1 signifikant von Null verschieden.

(**b**) Das Vertrauensintervall hat die Form $0.18170 \pm t_{11,0.975} \cdot 0.00173$. Es ist $t_{11,0.975} = 2.201$, also haben wir $[0.1779, 0.1855]$. Dies ist Vertrauensintervall ii).

(**c**) Für 1000m liefert das Modell den Wert $-62.6 + 0.18170 \cdot 1000 = 119.1$. Das Residuum ist daher $136 - 119.1 = 16.9$.

(**d**) Nein, denn für 100km müssten wir eine Extrapolation verwenden. In diesem Bereich haben wir keine Daten.

(**e**) Im Output können wir unter `Residual standard error` 62.68 ablesen. Für kleine Distanzen hat das Modell also einen viel zu grossen relativen Fehler!

(**f**) Man sieht einen sehr deutlichen Trend. Also stimmt das Modell nicht; wir haben einen systematischen Effekt nicht modelliert.

(**g**) Wir müssen den quadratischen Effekt noch berücksichtigen:

$$\text{Zeit}_i = \beta_0 + \beta_1 \cdot \text{Distanz}_i + \beta_2 \cdot \text{Distanz}_i^2 + \varepsilon_i$$

Aufgabe 3

Gemäss der Big Bang Theorie ist die Distanz Y eines galaktischen Nebels gegeben durch

$$Y_i = \beta_1 x_i + E_i \,,$$

wobei x_i die gemessene Fluchtgeschwindigkeit des $i-$ten Nebels bezeichnet. Die Bedeutung von β_1 entspricht dem Alter des Universums. Es wäre falsch, β_1 mit dem Wert 0.001373 zu schätzen, wie wir es in Kapitel 6.1 gemacht haben, da sich diese Schätzung auf ein anderes Modelll beziehen würde - eines mit einem von null verschiedenen Achsenabschnitt, in welchem die Steigung der Regressionsgeraden eine andere Bedeutung hätte. Mit der Methode der kleinsten Quadrate lässt sich der Parameter β_1 für das Urknall Modell bestimmen. Mit `R` findet man

```
> recession.velocity <- c(170,290,-130,-70,-185,-220,200,290,270,200,300,-30,
+         650,150,500,920,450,500,500,960,500,850,800,1090)
> distance <- c(0.032,0.034,0.214,0.263,0.275,0.275,0.450,0.500,
+         0.500,0.630,0.800,0.900,0.900,0.900,0.900,1.000,1.100,1.100,
+         1.400,1.700,2.000,2.000,2.000,2.000)
> lin.regr <- lm(distance ~  0 + recession.velocity)
> summary(lin.regr)

Call:
lm(formula = distance ~ 0 + recession.velocity)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76806 -0.06937  0.22932  0.46288  1.03910

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
recession.velocity 0.0019218  0.0001911   10.06 6.87e-10 ***
---
Signif. codes:  0
```

Eine Schätzung für β_1 ergibt also den Wert 0.001922 megaparsec Sekunde pro Kilometer mit Standardfehler, wobei die Anzahl Freiheitsgrade in diesem Fall 23 ist ($n-1$, da es bloss einen geschätzten Wert im Modell gibt). Da das 97.5% Quantil der t-Verteilung mit 23 Freiheitsgraden

```
> qt(0.975,23)

[1] 2.068658
```

ist, lautet das 95%-Vertrauensintervall für den Parameter β_1

$$[0.001922 - 2.069 \cdot 0.0000191, 0.001922 + 2.069 \cdot 0.0000191] = [0.001527, 0.002317]$$

Da eine megaparsec-Sekunde pro Kilometer etwa 979.8 Milliarden Jahren entspricht, ist das 95%-Konfidenzintervall für das Alter - ausgedrückt in Milliarden Jahren - des Universums gegeben durch

$$[1.50, 2.27] .$$

Aufgabe 4

(**a**) Der Erwartungswert der Position $S(n)$ nach n Schritten ist

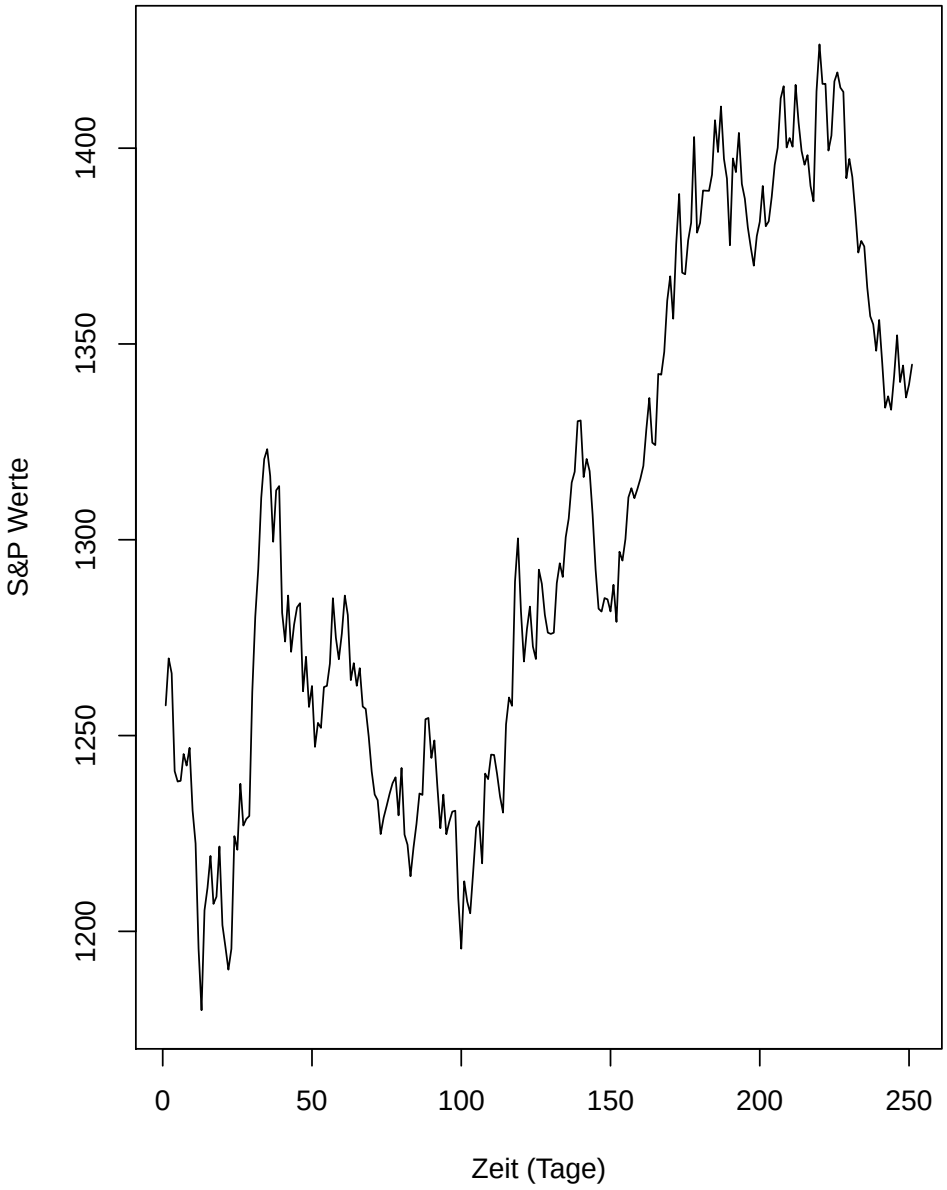
$$\mathrm{E}(S(n)) = x_0 + \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = x_0 + n\mu .$$

Die Varianz der Schrittweiten berechnet sich zu

$$\mathrm{Var}(S(n)) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2 .$$

Der betrunkene Spaziergänger kann also davon ausgehen, dass er sich im Mittel nach n Schritten an der Position $x_0 + n\mu$ befindet, wobei ein Mass für die Unsicherheit durch die Standardabweichung $\sqrt{n}\sigma$ gemessen werden kann. Falls $\mu > 0$, dann wird sich der betrunkene Spaziergänger nach n Schritten mit grosser Wahrscheinlichkeit rechts von seinem Ausgangspunkt x_0 befinden.

(**b**)
```
> steps <- rnorm(250,mean=0.483,sd=11)
> sp <- numeric(251)
> sp[1] <- 1257.7
> for(i in 1:250){
+         sp[i+1] <- sp[i]+steps[i]
+ }
> plot(sp,type="l",xlab="Zeit (Tage)",ylab="S&P Werte")
```



Aufgabe 5

(**a**) Generell gilt $F(s,t) = 0$ für $s < -A$ und $F(s,t) = 1$ für $s \geq A$. Daher sei $-A \leq s < A$. Die Verteilungsfunktion der Gleichverteilung ist nach Integration

$$F_{\Delta}(\delta) = \frac{\delta}{T} + \frac{1}{2} \quad \text{für} \quad -\frac{T}{2} \leq \delta \leq \frac{T}{2}.$$

- Rechtecksignal.

Für $t = 0$ gilt $-\frac{T_0}{2} < -\frac{T}{2} \leq \Delta T \leq \frac{T}{2} < \frac{T_0}{2}$, also

$$F(s,0) = P[S(0) \leq s] = P[f(\Delta T) \leq s] = P[\Delta T \leq 0] = F_\Delta(0) = \frac{1}{2}.$$

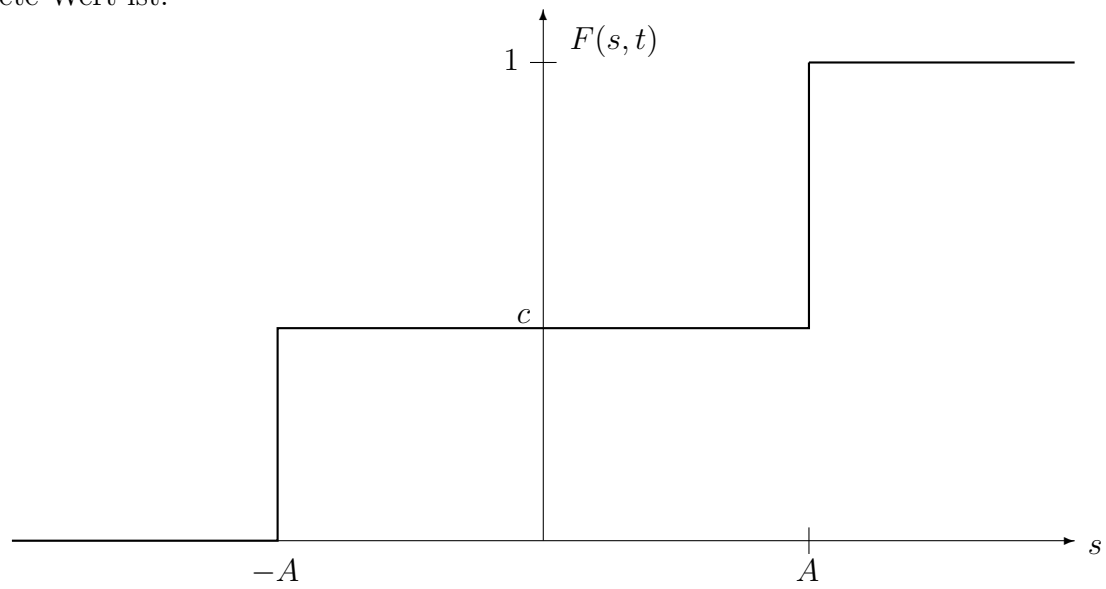
Für $0 < t < \frac{T}{2}$ gilt $-\frac{T_0}{2} < -\frac{T}{2} < t + \Delta T < T < \frac{T_0}{2}$, also

$$F(s,t) = P[S(t) \leq s] = P[f(t + \Delta T) \leq s] = P[t + \Delta T \leq 0] = F_\Delta(-t) = \frac{-t}{T} + \frac{1}{2}.$$

Für $t = \frac{T_0}{4}$ gilt $0 < t + \Delta T \leq \frac{T_0}{2}$, also

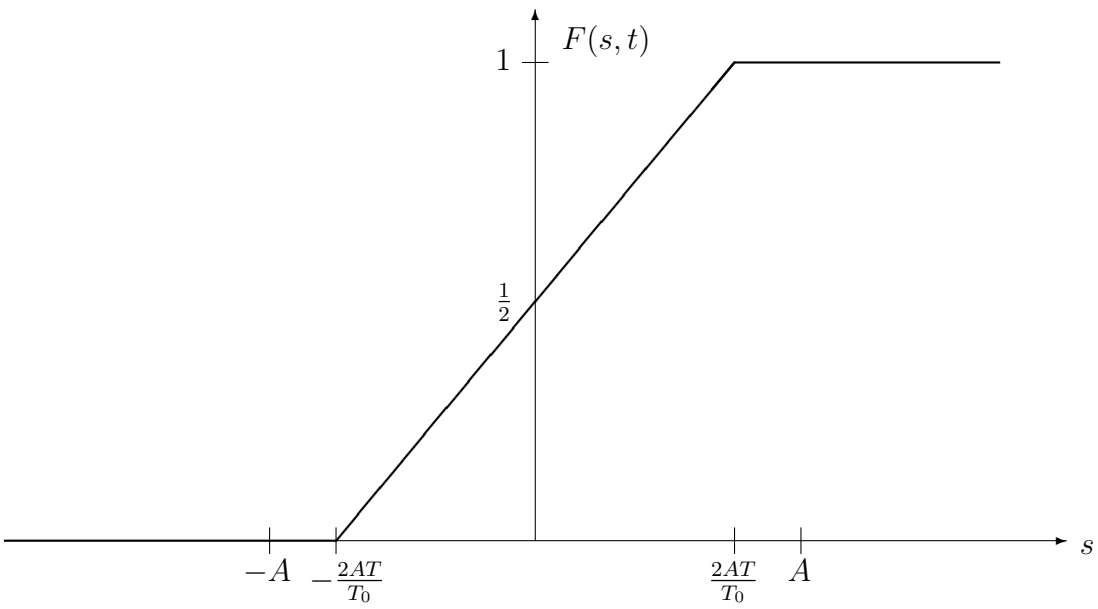
$$F(s, \frac{T_0}{4}) = P[S(\frac{T_0}{4}) \leq s] = P[f(\frac{T_0}{4} + \Delta T) \leq s] = 0.$$

Der Graph sieht also in allen Fällen folgendermassen aus, wobei c der oben berechnete Wert ist:



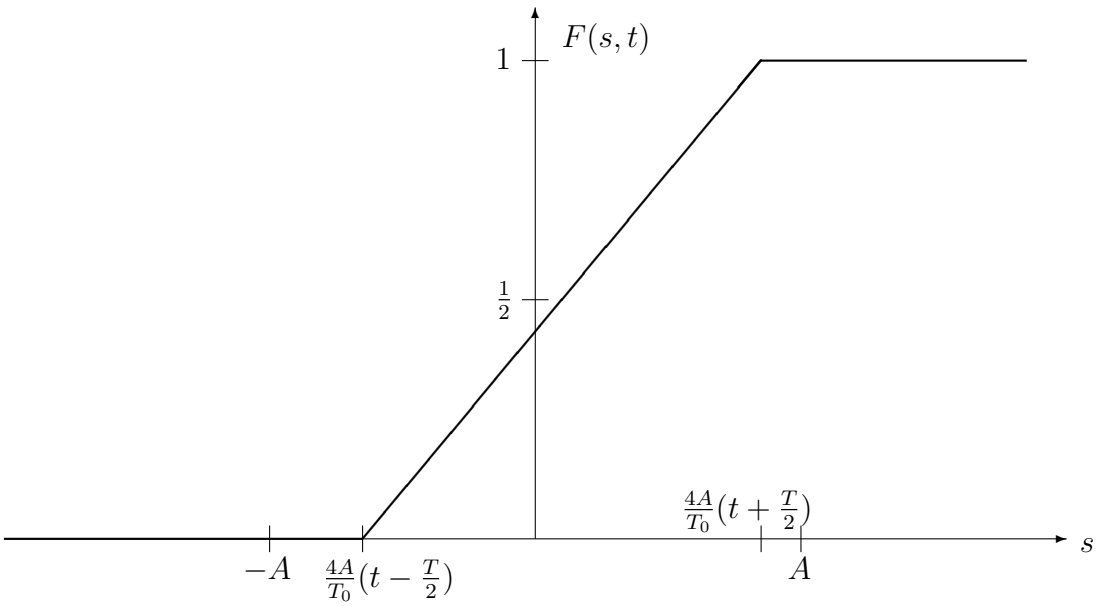
- Dreiecksignal.
  Für $t = 0$ und $-\frac{2AT}{T_0} \leq s \leq \frac{2AT}{T_0}$ gilt

$$\begin{aligned} F(s,0) = P[S(0) \leq s] = P[f(\Delta T) \leq s] &= P\left[\frac{4A}{T_0}\Delta T \leq s\right] \\ = P\left[\Delta T \leq \frac{T_0}{4A}s\right] &= F_\Delta\left(\frac{T_0}{4A}s\right) = \frac{T_0}{4AT}s + \frac{1}{2}. \end{aligned}$$



Für $0 < t \leq \frac{T_0}{4} - \frac{T}{2}$ und $\frac{4A}{T_0}(t - \frac{T}{2}) \leq s \leq \frac{4A}{T_0}(t + \frac{T}{2})$ gilt

$$\begin{aligned} F(s,t) &= P[S(t) \leq s] = P[f(t + \Delta T) \leq s] = P\left[\frac{4A}{T_0}(t + \Delta T) \leq s\right] \\ &= P\left[t + \Delta T \leq \frac{T_0}{4A}s\right] = F_\Delta\left(\frac{T_0}{4A}s - t\right) = \frac{T_0}{4AT}s - \frac{t}{T} + \frac{1}{2}. \end{aligned}$$

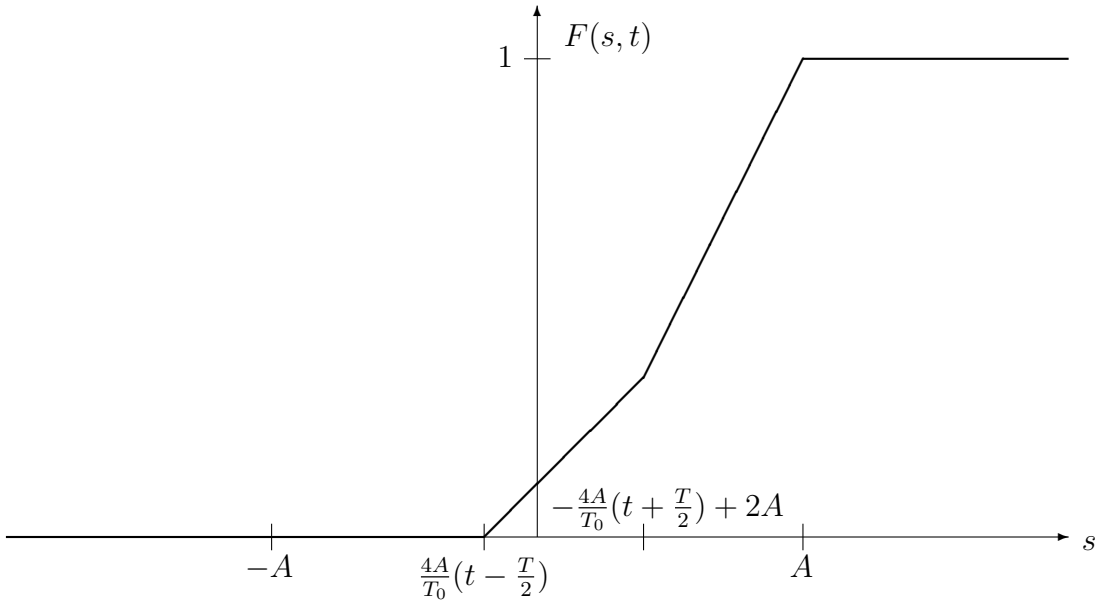


Sei nun $\frac{T_0}{4} - \frac{T}{2} < t < \frac{T_0}{4}$. Für $\frac{4A}{T_0}(t - \frac{T}{2}) \leq s \leq -\frac{4A}{T_0}(t + \frac{T}{2}) + 2A$ gilt

$$\begin{aligned} F(s,t) =& P\left[\frac{4A}{T_0}(t + \Delta T) \leq s \text{ und } t + \Delta T \leq \frac{T_0}{4}\right] \\ &+ \underbrace{P\left[-\frac{4A}{T_0}(t + \Delta T) + 2A \leq s \text{ und } t + \Delta T > \frac{T_0}{4}\right]}_{=0} \\ =& F_\Delta\left(\frac{T_0}{4A}s - t\right) = \frac{T_0}{4AT}s - \frac{t}{T} + \frac{1}{2}. \end{aligned}$$

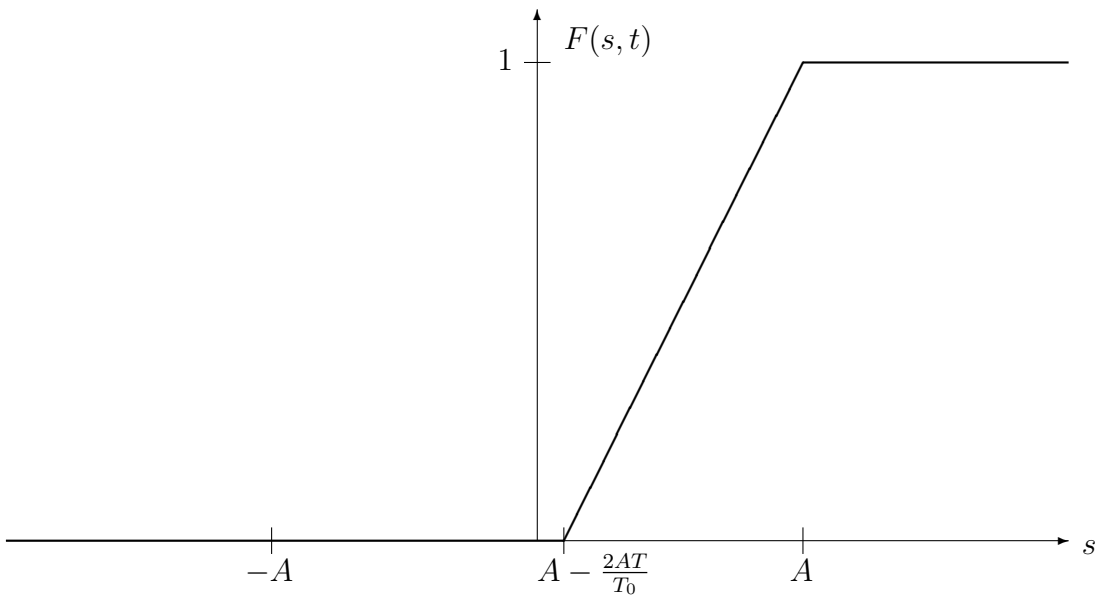
Für $-\frac{4A}{T_0}(t+\frac{T}{2})+2A \le s \le A$ gilt

$$\begin{aligned}
F(s,t) =& P\left[\frac{4A}{T_0}(t+\Delta T) \le s \text{ und } t+\Delta T \le \frac{T_0}{4}\right] \\
&+ P\left[-\frac{4A}{T_0}(t+\Delta T)+2A \le s \text{ und } t+\Delta T > \frac{T_0}{4}\right] \\
=& F_\Delta\left(\frac{T_0}{4A}s - t\right) + 1 - F_\Delta\left(-\frac{T_0}{4A}(s-2A) - t\right) \\
=& \frac{T_0}{4AT}s + 1 + \frac{T_0}{4AT}(s-2A) = \frac{T_0}{2AT}s + 1 - \frac{T_0}{2T}.
\end{aligned}$$



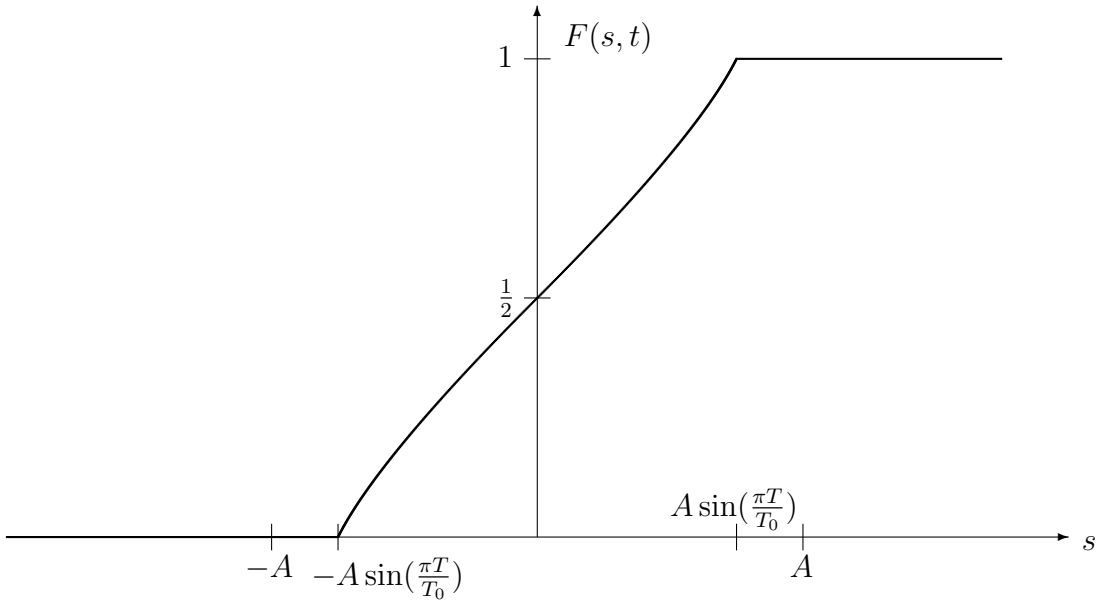
Sei nun $t = \frac{T_0}{4}$. Für $A - \frac{4AT}{T_0} \le s \le A$ gilt

$$\begin{aligned}
F(s,\frac{T_0}{4}) =& P\left[\frac{4A}{T_0}(\frac{T_0}{4}+\Delta T) \le s \text{ und } \frac{T_0}{4}+\Delta T \le \frac{T_0}{4}\right] \\
&+ P\left[-\frac{4A}{T_0}(\frac{T_0}{4}+\Delta T)+2A \le s \text{ und } \frac{T_0}{4}+\Delta T > \frac{T_0}{4}\right] \\
=& F_\Delta\left(\frac{T_0}{4A}s - \frac{T_0}{4}\right) + 1 - F_\Delta\left(-\frac{T_0}{4A}(s-2A) - \frac{T_0}{4}\right) \\
=& \frac{T_0}{4AT}s + 1 + \frac{T_0}{4AT}(s-2A) = \frac{T_0}{2AT}s + 1 - \frac{T_0}{2T}.
\end{aligned}$$



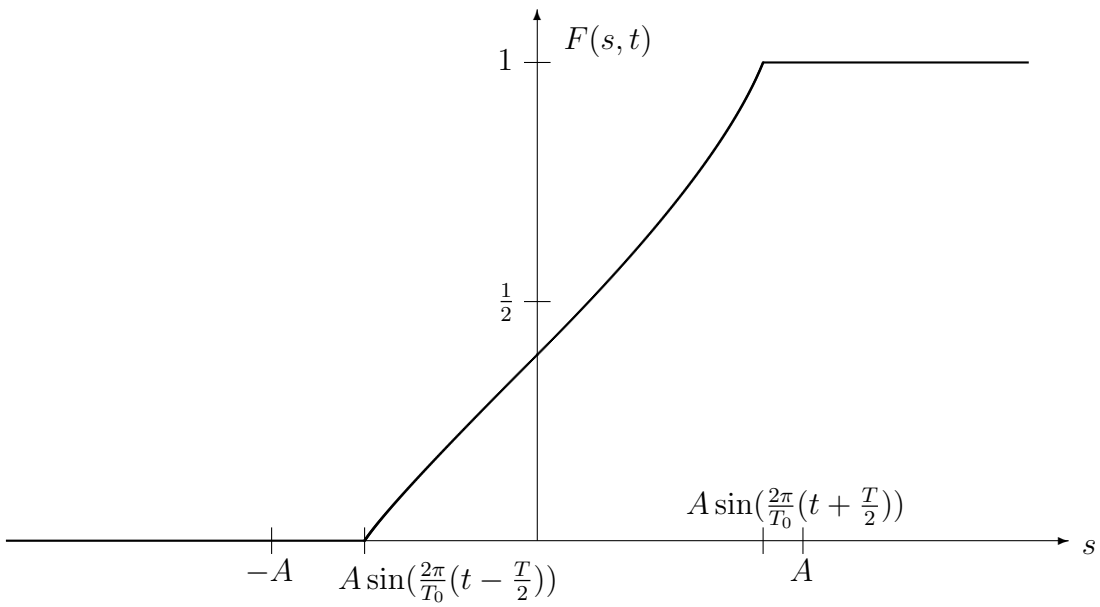
- Sinussignal.
  Für $t = 0$ und $-A\sin(\frac{\pi T}{T_0}) \leq s \leq A\sin(\frac{\pi T}{T_0})$ gilt

$$\begin{aligned} F(s,0) =& P[S(0) \leq s] = P[f(\Delta T) \leq s] = P\left[A\sin\left(\frac{2\pi}{T_0}\Delta T\right) \leq s\right] \\ =& P\left[\Delta T \leq \frac{T_0}{2\pi}\arcsin\frac{s}{A}\right] = F_\Delta\left(\frac{T_0}{2\pi}\arcsin\frac{s}{A}\right) = \frac{T_0}{2\pi T}\arcsin\frac{s}{A} + \frac{1}{2}. \end{aligned}$$



Für $0 < t \leq \frac{T_0}{4} - \frac{T}{2}$ und $A\sin(\frac{2\pi}{T_0}(t - \frac{T}{2})) \leq s \leq A\sin(\frac{2\pi}{T_0}(t + \frac{T}{2}))$ gilt

$$\begin{aligned} F(s,0) =& P\left[A\sin\left(\frac{2\pi}{T_0}(t + \Delta T)\right) \leq s\right] \\ =& P\left[\Delta T \leq \frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right] = F_\Delta\left(\frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right) = \frac{T_0}{2\pi T}\arcsin\frac{s}{A} - \frac{t}{T} + \frac{1}{2}. \end{aligned}$$

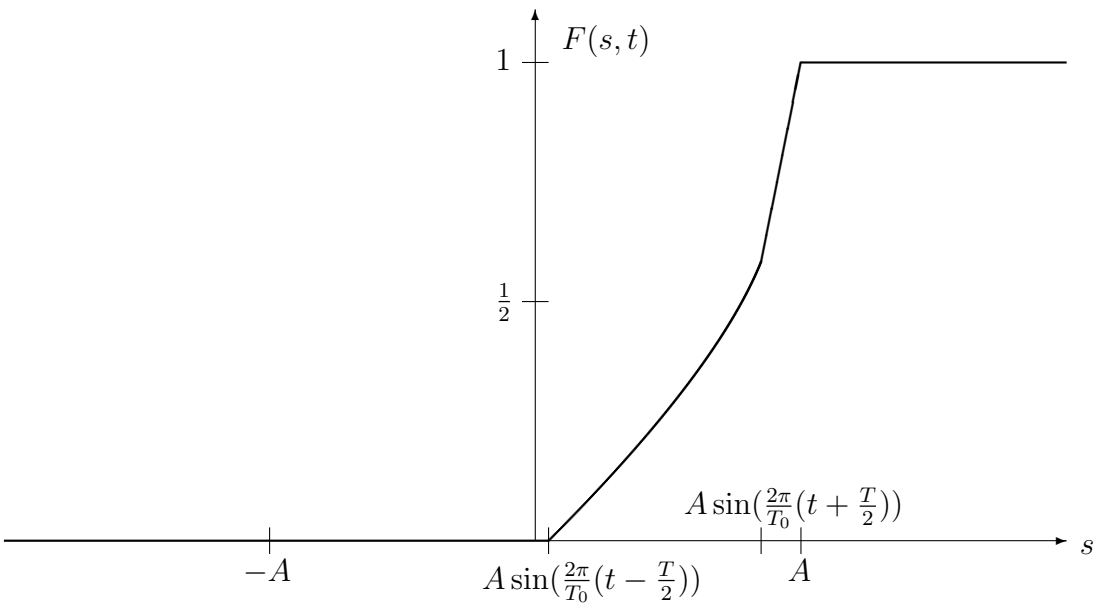


Sei nun $\frac{T_0}{4} - \frac{T}{2} < t < \frac{T_0}{4}$. Für $A\sin\frac{2\pi}{T_0}(t-\frac{T}{2}) \le s \le A\sin\frac{2\pi}{T_0}(t+\frac{T}{2})$ gilt

$$\begin{aligned}
F(s,t) =& P\left[A\sin\big(\frac{2\pi}{T_0}(t+\Delta T)\big) \le s \text{ und } t+\Delta T \le \frac{T_0}{4}\right] \\
&+ \underbrace{P\left[A\sin\big(\frac{2\pi}{T_0}(t+\Delta T)\big) \le s \text{ und } t+\Delta T > \frac{T_0}{4}\right]}_{=0} \\
=& P\left[\Delta T \le \frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right] \\
=& F_\Delta\left(\frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right) \\
=& \frac{T_0}{2\pi T}\arcsin\frac{s}{A} - \frac{t}{T} + \frac{1}{2}.
\end{aligned}$$

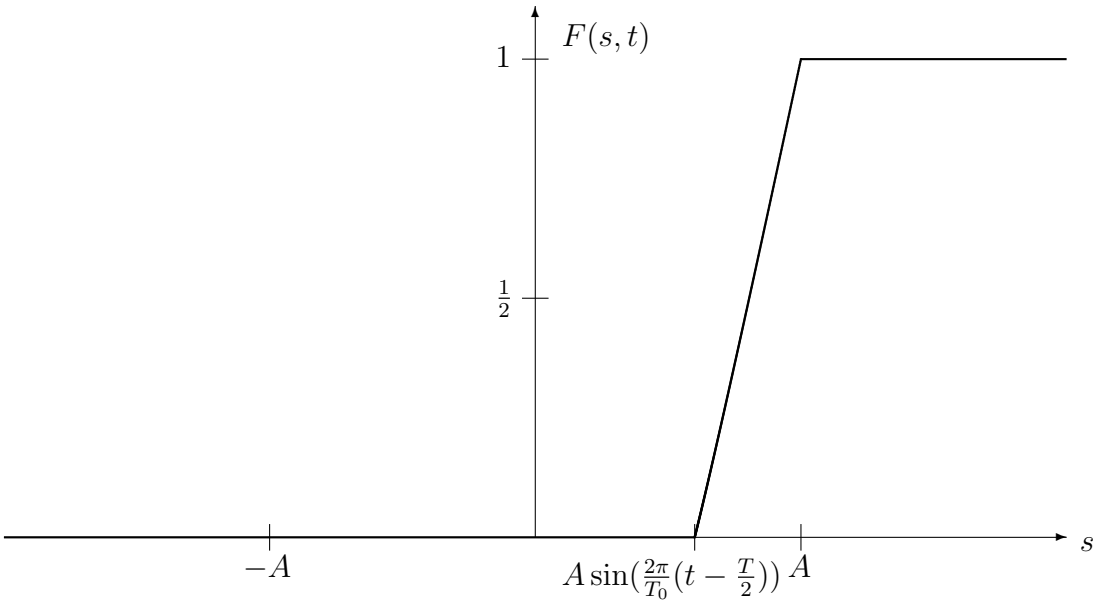
Für $A\sin\frac{2\pi}{T_0}(t+\frac{T}{2}) \le s \le A$ gilt

$$\begin{aligned}
F(s,t) =& P\left[A\sin\big(\frac{2\pi}{T_0}(t+\Delta T)\big) \le s \text{ und } t+\Delta T \le \frac{T_0}{4}\right] \\
&+ P\left[A\sin\big(\frac{2\pi}{T_0}(t+\Delta T)\big) \le s \text{ und } t+\Delta T > \frac{T_0}{4}\right] \\
=& P\left[\Delta T \le \frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right] + P\left[\Delta T \ge \frac{T_0}{2} - \frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right] \\
=& F_\Delta\left(\frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right) + 1 - F_\Delta\left(\frac{T_0}{2} - \frac{T_0}{2\pi}\arcsin\frac{s}{A} - t\right) \\
=& \frac{T_0}{2\pi T}\arcsin\frac{s}{A} - \frac{t}{T} + \frac{1}{2} + 1 - \frac{T_0}{2T} + \frac{T_0}{2\pi T}\arcsin\frac{s}{A} + \frac{t}{T} - \frac{1}{2} \\
=& \frac{T_0}{\pi T}\arcsin\frac{s}{A} + 1 - \frac{T_0}{2T}.
\end{aligned}$$



Sei nun $t = \frac{T_0}{4}$. Eine analoge Rechnung wie oben zeigt für $A\sin\frac{2\pi}{T_0}(\frac{T_0}{4} - \frac{T}{2}) \leq s \leq A$

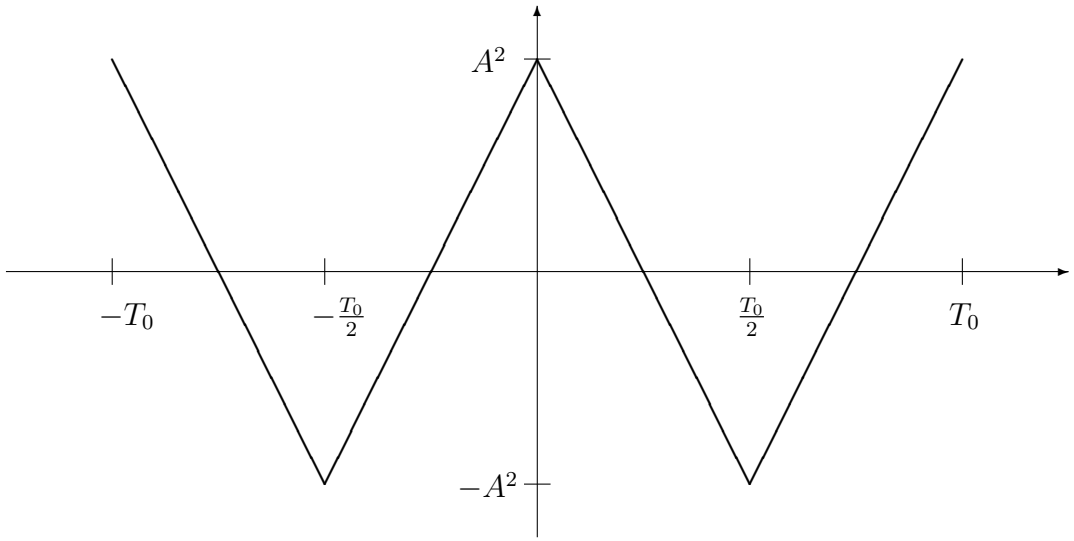
$$F(s, \frac{T_0}{4}) = \frac{T_0}{\pi T} \arcsin\frac{s}{A} + 1 - \frac{T_0}{2T}.$$



(**b**) • Rechtecksignal. Die Autokorrelationsfunktion ist die T_0-periodische Fortsetzung von

$$\varphi(\tau) = \begin{cases} A^2(1 - 4\frac{\tau}{T}), & 0 \leq \tau \leq \frac{T_0}{2} \\ A^2(1 + 4\frac{\tau}{T}), & -\frac{T_0}{2} \leq \tau \leq 0. \end{cases}$$

Der Graph sieht also folgendermassen aus:



- Dreiecksignal.Die Autokorrelationsfunktion ist die T_0-periodische Fortsetzung von

$$\varphi(\tau) = \begin{cases} A^2(1 - 4\frac{\tau}{T}), & 0 \leq \tau \leq \frac{T_0}{2} \\ A^2(1 + 4\frac{\tau}{T}), & -\frac{T_0}{2} \leq \tau \leq 0. \end{cases}$$

  Der Graph sieht also folgendermassen aus:

Aufgabe 6

Lösung auf Anfrage erhältlich.

Stochastik # Testklausur **HSLU**

Vorbesprechung: Ferienserie 2013

Aufgabe 1

(**a**) Auf wieviele verschiedene Arten können fünf Kinder in Reih und Glied stehen? (120)

(**b**) Nehmen Sie an, an einer Geburtstagsparty befinden sich 56 Personen. Wie gross ist die Wahrscheinlichkeit, dass mindestens eine zweite Person ebenfalls Geburstag am selben Tag wie das Geburtstagskind hat? (0.988)

(**c**) Wieviele Leute müssen Sie fragen, um mit einer 50% Wahrscheinlichkeit auf eine weitere Person zu treffen, die am gleichen Tag Geburtstag hat wie Sie. (253)

(**d**) Sie nehmen an einem Glücksspiel teil: Sie werfen eine faire Münze. Bei Kopf gewinnen Sie 200 CHF, bei Zahl verlieren Sie 100 CHF. Wenn Sie dieses Spiel zehnmal spielen, welchen Gewinn/Verlust erwarten Sie? (+500 CHF)

Aufgabe 2

Ein Analog-Digital-Wandler oder A/D-Wandler ist ein elektronisches Gerät oder Bauteil zur Umsetzung analoger Eingangssignale in digitale Daten bzw. einen Datenstrom, der dann weiterverarbeitet oder gespeichert werden kann. Analog-Digital-Wandler sind elementare Bestandteile fast aller Geräte der modernen Kommunikations- und Unterhaltungselektronik wie beispielsweise Mobiltelephonen, Digitalkameras, oder Camcordern. Zudem werden sie in der Messwerterfassung in industriellen Anwendungen, Maschinen und in technischen Alltagsgegenständen wie Autos oder Haushaltsgeräten eingesetzt. Nach der A/D-Wandlung eines analogen Signals stehen die Daten als binäre Rechteckfolgen direkt für eine weitere Verarbeitung, Speicherung oder Übertragung zur Verfügung. Während in Verarbeitungskomponenten im allgemeinen alle Bits parallel vorliegen, wird insbesondere bei Übertragungen digitaler Signale eine serielle Darstellung der Bits hintereinander in verschachtelter Form bevorzugt. Die Übertragung von digitalen Signalen ist häufig fehleranfällig.
Wir nehmen an, dass ein Bit, der durch einen Übertragungskanal gesendet wird, mit einer Wahrscheinlichkeit von $\pi = 0.1$ fehlerhaft empfangen wird. Wir nehmen ebenfalls an, dass die n Übertragungsversuche unabhängig voneinander sind. Wir bezeichnen mit X die Anzahl fehlerhaft übertragener Bits.

(**a**) Geben Sie die Verteilung von X inklusive Parameter an.

(**b**) Wie gross ist die Wahrscheinlichkeit, dass es in den nächsten vier übertragenen Bits keinen, resp. zwei Fehler hat? Geben Sie diese beiden Ereignisse als Mengen an, indem Sie ein fehlerhaft empfangenes Bit mit E, und ein korrekt empfangenes Bit mit O bezeichnen. (0.0486;0.6561)

(**c**) Wie gross ist die Wahrscheinlichkeit, dass höchstens 3 fehlerhafte Bits empfangen werden? (0.9999)

(**d**) Wieviele fehlerhaft empfangene Bits erwarten Sie bei 100 übertragenen Bits? Wie gross ist die Varianz? Wie interpretieren Sie die Varianz? (10;9)

Sie testen ein neues Material für den Übertragungskanal und stellen fest, dass von 100 übertragenen Bits 6 fehlerhaft empfangen werden. Es soll nun getestet werden, ob das neue Material (auf dem Signifikanzniveau von 5%) signifikant weniger fehlerhafte Bits überträgt als das Material mit einer Fehlerrate von $\pi = 0.1$.

(**e**) Was sind die Null- und die Alternativhypothese?

(**f**) Führen Sie einen geeigneten Test durch. Geben Sie den Verwerfungsbereich für X und den Testentscheid an.

(**g**) Wie gross ist die Macht des Tests bei konkreter Alternative $\pi = 0.05$?

(**h**) Berechnen Sie das 95%-Konfidenzintervall für π? Wie interpretieren Sie das Konfidenzintervall? ($[0, 0.11]$)

Aufgabe 3

In einer Studie (*Human Factors*, 1962, pp. 375-380) wurden $n = 14$ Personen aufgefordert, zwei Automobile mit sehr unterschiedlichen Wendekreisen und Radständen rückwärts einzuparkieren. Die dafür benötigte Zeit in Sekunden ist in folgender Tabelle aufgetragen. Es stellt sich nun die Frage, ob Automobil 1 oder 2 schneller rückwärts einparkiert werden kann.

Person	Automobil 1	Automobil 2
1	37.0	17.8
2	25.8	20.2
3	16.2	16.8
4	24.2	41.4
5	22.0	21.4
6	33.4	38.4
7	23.8	16.8
8	58.2	32.2
9	33.6	27.8
10	24.4	23.2
11	23.4	29.6
12	21.2	20.6
13	36.2	32.2
14	29.8	53.8

Tabelle 1: Zum Rückwärts-Parkieren benötigte Zeit (Sekunden) zweier Fahrzeuge.

(**a**) Handelt es sich um einen gepaarten oder einen ungepaarten Test?

(**b**) Geben Sie die Null- und die Alternativhypothese an.

(**c**) Geben Sie eine Schätzung für die Varianz σ^2 der Differenz an.

(**d**) Führen Sie den geeigneten t-Test auf dem Signifikanzniveau von 5% durch: Bestimmen Sie den Wert der Teststatistik T, den Verwerfungsbereich für T und den Testentscheid. (Wenn Sie obige Aufgabe nicht lösen konnten, benutzen Sie im folgenden als Ersatzwert $\sigma^2 = 150$.) ($t = 0.358$)

(**e**) Bestimmen Sie ein zweiseitiges 90%-Vertrauensintervall für μ_D. ($[-4.79, 7.22]$)

(**f**) Angenommen, der Wert $\sigma^2 = 160$ wäre nicht aus den Daten geschätzt, sondern bekannt: Wie lautet dann das zweiseitige 90%-Vertrauensintervall?

Aufgabe 4

Übelkeit und Erbrechen sind häufige Nebenwirkungen, die während einer Krebs Chemotherapie auftreten können. Solche Nebenwirkungen können die Eignung eines Patienten, sich einer langfristigen Chemotherapie zu unterziehen, beeinträchtigen und in Frage stellen. In einer randomisierten Doppelblindstudie (Chang, A.E., *et al.*, "Delta-9-Tetrahydrocannibinol as an Antiemetic in Cancer Patients Receiving High-Dose Methotrexate") wurde die Wirkung von Marihuana zur Verminderung von Nebenwirkungen untersucht. 15 Krebspatienten wurde nach jeder der ersten drei Chemotherapie Behandlungen jeweils zufällig entweder Marihuana oder ein Placebo verabreicht. Dann wurde in den darauffolgenden drei Chemotherapie Behandlungen zur anderen entgegengesetzten Nachbehandlung gewechselt. Bei der Nachbehandlung wurde Marihuana und Placebo entweder in Form von identischen Pillen oder Zigaretten verabreicht. In untenstehender Tabelle ist die Gesamtzahl von Erbrechen- und Würgereflexen nach der jeweiligen Nachbehandlung zusammengestellt. Es stellt sich also die Frage, ob Marihuana die Nebenwirkungen bei einer Chemotherapie reduziert.

(**a**) Handelt es sich um einen gepaarten oder ungepaarten Test? Begründen Sie.

(**b**) Geben Sie die Null- und Alternativhypothese an.

(**c**) Welchen Test wenden Sie an? Begründen Sie die Kriterien zur Beantwortung dieser Frage.

(**d**) Wie wahrscheinlich ist die Nullhypothese? Geben Sie den (einseitigen) p-Wert an. (0.0008308)

Patient	Marihuana	Placebo
1	15	23
2	25	50
3	0	0
4	0	99
5	4	31
6	2	21
7	1	79
8	4	113
9	9	53
10	0	0
11	22	61
12	11	18
13	0	12
14	0	6
15	0	5

Tabelle 2: Anzahl Erbrechen- und Würgevorfälle bei 15 sich in Chemotherapie befindenden Krebspatienten bei Nachbehandlung mit Marihuana und Placebo.

Aufgabe 5

In folgender Tabelle ist der durchschnittliche Weinkonsum (in Liter pro Person und Jahr) und die Mortalität aufgrund von Herz-Kreislauferkrankungen (Anzahl Todesfälle pro 1000 Personen zwischen 55 und 64 Jahren pro Jahr) in 18 industrualisierten Ländern zusammengestellt (A.S.St.Leger, A.L.Chochrane, and F.Moore, "Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine."*Lancet, 1979*). Es stellt sich nun die Frage, ob diese Daten suggerieren, dass es einen Zusammenhang zwischen der Mortalitätsrate aufgrund von Herzkreislauferkrankung und Weinkonsum gibt. Wir legen den Daten folgendes Modell zugrunde

$$\text{mort}_i = \beta_0 + \beta_1 \cdot \text{wine}_i + E_i \ , \qquad E_i \sim \mathcal{N}(0, \sigma^2) \ .$$

(**a**) Bestimmen Sie $\widehat{\beta}_1$. (-0.07608)

(**b**) Bestimmen Sie $\widehat{\beta}_0$. Was für eine Bedeutung hat β_0 in diesem Beispiel?(7.68655)

(**c**) Bestimmen Sie den Standardfehler von $\widehat{\beta}_1$. (0.47332)

(**d**) Mit wievielen Freiheitsgraden wurde der "residual standard error" berechnet? (16)

(**e**) Berechnen Sie das 99%-Konfidenzintervall für β_1 (ohne Normalapproximation).

(**f**) Kann die Nullhypothese $H_0 \ : \ \beta_1 = 0$ auf dem 5% Signifikanzniveau verworfen werden? (ja)

Land	Weinkonsum	Mortalität Herzerkrankung
Norwegen	2.8	6.2
Schottland	3.2	9.0
Grossbritannien	3.2	7.1
Irland	3.4	6.8
Finnland	4.3	10.2
Kanada	4.9	7.8
Vereinigte Staaten	5.1	9.3
Niederlande	5.2	5.9
New Zealand	5.9	8.9
Dänemark	5.9	5.5
Schweden	6.6	7.1
Australien	8.3	9.1
Belgien	12.6	5.1
Deutschland	15.1	4.7
Österreich	25.1	4.7
Schweiz	33.1	3.1
Italien	75.9	3.2
Frankreich	75.9	2.1

Tabelle 3: Weinkonsumation (Liter pro Person pro Jahr) und Mortalität aufgrund von Herzkreislauferkrankung (Todesfälle pro 1000) in 18 Ländern.

(**g**) Wie hoch ist das Residuum für die Schweiz in unserem Modell? (2.068)

(**h**) Welche Aussage trifft aufgrund des Tukey-Anscombe Plots und des Normal-Q-Q Plots zu?

1. Alle Modellannahmen sind erfüllt.
2. Die Fehlervarianz scheint nicht konstant zu sein, aber die Normalverteilungsannahme ist plausibel.
3. Die Fehlervarianz scheint konstant zu sein, aber die Normalverteilungsannahme scheint nicht zuzutreffen.
4. Sowohl konstante Fehlervarianz als auch Normalverteilungsannahme treffen nicht zu.

Aufgabe 6

Die folgenden Aufgaben sind zufällig angeordnet und nicht nach Schwierigkeitsgrad sortiert.

1. Seien A und B stochastisch unabhängige Ereignisse. Dann gilt für das Komplement A^C von A:
   (a) A^C und B sind stochastisch unabhängig.
   (b) A^C und B sind nicht stochastisch unabhängig.

(c) Es kann keine Aussage über die stochastische Abhängigkeit von A^C und B gemacht werden.

2. Zwei Ereignisse A und B schliessen sich aus. Welche Aussage trifft immer zu?

   (a) $P(A \cup B) = P(A) + P(B)$.

   (b) $P(A) = P(B)$.

   (c) $P(A) + P(B) < 1$.

   (d) $P(A) + P(B) = 1$.

3. In einer Kiste seien drei Spielkarten. Eine davon ist auf beiden Seiten schwarz, eine auf beiden Seiten weiss und eine auf einer Seite schwarz und auf der anderen Seite weiss. Es wird zufällig eine Karte gezogen und auf den Tisch gelegt. Angenommen, die Karte ist auf der sichtbaren Seite schwarz. Wie gross ist die Wahrscheinlichkeit, dass sie auf der anderen Seite weiss ist?

   (a) $1/3$.

   (b) $1/2$.

   (c) $1/4$.

   (d) $2/3$.

4. Für eine Zufallsvariable Z gelte $\mathrm{E}[Z] = 1$ und $\mathrm{Var}[Z] = 2$. Welchen Wert hat $\mathrm{E}[3Z + 2]$?

   (a) 5 .

   (b) 6 .

   (c) 7 .

   (d) 8 .

5. Für eine Zufallsvariable Z gelte $\mathrm{E}[Z] = 1$ und $\mathrm{Var}[Z] = 2$. Welchen Wert hat $\mathrm{Var}[3Z+2]$?

   (a) 6 .

   (b) 8 .

   (c) 18 .

   (d) 20 .

6. Eine Zufallsvariable $X \in [0, \infty]$ habe die Dichtefunktion $f(x) = \lambda^2 x e^{-\lambda x}$, $\lambda > 0$. Was ist die dazugehörige kumulative Verteilungsfunktion?

   (a) $F(x) = \lambda x$, $x \in [0, \infty[$.
   (b) $F(x) = 1 - \exp(-\lambda x)$, $x \in [0, \infty[$.
   (c) $F(x) = 1 - (1 + \lambda x)e^{-\lambda x}$, $x \in [0, \infty[$.
   (d) $F(x) = \Phi(x)$, $x \in \mathbb{R}$.

7. Eine Zufallsvariable $X \in \mathbb{R}$ habe die kumulative Verteilungsfunktion

$$F(x) = \alpha \frac{\exp(\lambda x)}{1 + \exp(\lambda x)} ,$$

   wobei $\lambda > 0$. Welchen Wert hat α?

   (a) $\alpha = \frac{1}{4}$.
   (b) $\alpha = \frac{1}{2}$.
   (c) $\alpha = 1$.
   (d) Keine Aussage möglich.

8. Wir testen mit einem Binomialtest auf dem Signifikanzniveau 0.05, ob eine Münze gefälscht wurde, so dass sie häufiger Kopf zeigt ($H_0 : \pi = 0.5$, $H_A : \pi > 0.5$). Wie gross ist die Wahrscheinlichkeit, dass wir die Münze als gefälscht bezeichnen (H_0 wird verworfen), wenn sie in Wahrheit fair (H_0 ist richtig) ist?

   (a) Mindestens 0.05 .
   (b) Mindestens 0.95 .
   (c) Höchstens 0.05 .
   (d) Höchstens 0.95 .

9. Die Wahrscheinlichkeit, dass eine zufällig gewählte Person die Krankheit K hat, sei $P(K) = 0.01$. Für diese Krankheit wurde ein Test T entwickelt. Eine kranke Person wird mit Wahrscheinlichkeit $P(T|K) = 0.9$ positiv getestet, eine gesunde Person mit Wahrscheinlichkeit $P(T|K^C) = 0.1$. Nun wird eine zufällig gewählte Person positiv getestet. Wie gross ist ungefähr die Wahrscheinlichkeit, dass die Person krank ist?

   (a) 0.0041 .
   (b) 0.041 .
   (c) 0.0083 .
   (d) 0.083 .

10. Gegeben sei eine Zufallsvariable $X \sim$ Binomial(n, π), wobei $n = 200$ und $\pi = 0.01$. Welche Approximation eignet sich am besten für die Verteilung von X?

    (a) Poisson-Verteilung .
    (b) Normal-Verteilung .
    (c) Exponential-Verteilung .

Aufgabe 1

(**a**) Es gibt $5! = 120$ verschiedene Arten, wie sich 5 Kinder in einer Reihenfolge aufstellen können.

(**b**) Wir bezeichnen mit X die Anzahl Personen, die Geburtstag haben am selben Tag wie das Geburtstagskind. Unter der Annahme, dass jeder Tag mit gleicher Wahrscheinlichkeit ein Geburtstag ist, folgt für die Verteilung von X

$$X \sim \text{Bin}(n = 55, \pi = \frac{1}{365}) .$$

Die Wahrscheinlichkeit, dass mindestens jemand anderes von den übrigen $n = 55$ Personen am selben Tag Geburtstag hat, ist

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{55}{0} \left(\frac{1}{365}\right)^0 \cdot \left(\frac{364}{365}\right)^{55} = 0.14 .$$

(**c**) Wir bezeichnen mit A wiederum das Ereignis, dass unter den n angefragten Personen genau **eine** Person am selben Tag Geburtstag hat. Wiederum ist es einfacher, das Ereignis A^C zu betrachten, dass niemand am selben Tag Geburtstag hat. Wiederum enthält der Grundraum Ω 365^n Elemente. Daraus folgt

$$P(A^C) = \frac{364^n}{365^n} ,$$

und somit

$$P(A) = 1 - \frac{364^n}{365^n} .$$

$P(A)$ ist 0.5 für

```
> log(0.5)/(log(364/365))

[1] 252.652
```

Somit müssen 253 Personen gefragt werden, um mit einer 50% Wahrscheinlichkeit eine Person anzutreffen mit dem gleichen Geburtstag.

(**d**) Wir bezeichnen mit X_i den Gewinn/Verlust beim $i-$ten Wurf. Falls Kopf geworfen wird, dann ist $X_i = 200$, falls Zahl geworfen wird, dann ist $X_i = -100$. Mit X bezeichnen wir den Gesamtgewinn/verlust

$$X = \sum_{i=1}^{10} X_i .$$

Da es sich um eine faire Münze handelt, ist $P(X_i = 200) = P(X_i = -100) = 0.5$. Folglich ist der erwartete Gewinn

$$\begin{aligned} E(X) &= \sum_{i=1}^{10} \left(\sum_{x_i \in W_{X_i}} P(X_i = x_i) \cdot x_i \right) \\ &= \sum_{i=1}^{10} \left(P(X_i = 200) \cdot 200 + P(X_i = -100) \cdot (-100) \right) \\ &= 10 \cdot (0.5 \cdot 200 - 0.5 \cdot 100) \\ &= 500\,. \end{aligned}$$

Aufgabe 2

(**a**) Die Anzahl fehlerhaft übertragener Bits X bei n unabhängigen Übertragungsversuchen und bei einer Fehlerwahrscheinlichkeit π folgt einer Binomialverteilung: $X \sim \text{Binomial}(n, \pi)$. Die Punktwahrscheinlichkeit ist gegeben durch

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}\,,$$

wobei n die Anzahl Übertragungsversuche bezeichnet und $\pi = 0.1$ ist.

(**b**) Die Wahrscheinlichkeit, dass bei $n = 4$ Übertragungsversuchen genau $x = 0$ Bits fehlerhaft übermittelt wurden, ist

$$\binom{4}{0} \pi^0 (1 - \pi)^4\,,$$

was mit `R` den Wert

```
> choose(4,0)*0.1^0*(1-0.1)^4

[1] 0.6561
```

oder

```
> dbinom(0,4,prob=0.1)

[1] 0.6561
```

ergibt. Für $x = 2$ finden wir

```
> choose(4,2)*0.1^2*(1-0.1)^2

[1] 0.0486
```

oder

```
> dbinom(2,4,prob=0.1)

[1] 0.0486
```

Das Ereignis, dass $X = 0$ ist, besteht aus folgendem Elementarereignis:

$$\{OOOO\}$$

Die Wahrscheinlichkeit des Elementarereignisses $\{OOOO\}$ ist

$$P(OOOO) \overset{\text{Unabhängigkeit}}{=} P(O) \cdot P(O) \cdot P(O) \cdot P(O) = (0.9)^4 = 0.6561 \, .$$

Das Ereignis mit $X = 0$ hat also die Wahrscheinlichkeit:

$$P(X = 0) = 1 \cdot 0.6561 = 0.6561 \, .$$

Das Ereignis, dass $X = 2$ ist, besteht aus folgenden sechs Elementarereignissen:

$$\{EEOO, EOEO, EOOE, OEEO, OEOE, OOEE\}$$

Unter der Annahme, dass die Übertragungsversuche unabhängig voneinander sind, ist die Wahrscheinlichkeit vom Elementarereignis $\{EEOO\}$

$$P(EEOO) = P(E) \cdot P(E) \cdot P(O) \cdot P(O) = (0.1)^2 \cdot (0.9)^2 = 0.0081 \, .$$

Jedes der sechs Elementarereignisse mit $X = 2$ hat dieselbe Wahrscheinlichkeit. Folglich gilt

$$P(X = 2) = 6 \cdot 0.0081 = 0.0486 \, .$$

Die Anzahl Elementarereignisse mit $X = 2$ berechnet sich mit dem Binomialkoeffizienten $\binom{4}{2}$

```
> choose(4,2)

[1] 6
```

(**c**) Die Wahrscheinlichkeit, dass höchstens 3 Bits fehlerhaft übertragen worden sind, ist

$$P(X \leq 3) = \sum_{i=0}^{3} P(X = i) = \sum_{i=0}^{3} \binom{4}{i} \pi^i (1 - \pi)^{4-i} \, .$$

Mit `R` erhalten wir

```
> pbinom(3,4,prob=0.1)

[1] 0.9999
```

(**d**) Der Erwartungswert für eine binomialverteilte Zufallsvariable ist

$$E(X) = n \cdot \pi \,,$$

also für $X \sim \text{Bin}(n = 100, \pi = 0.1)$ erhalten wir $E(X) = 100 \cdot 0.1 = 10$. Für die Varianz gilt folgende Beziehung

$$\text{Var}(X) = n\pi(1 - \pi) \,,$$

somit erhalten wir $\text{Var}(X) = 100 \cdot 0.1 \cdot (1 - 0.1) = 9$. Die Varianz ist ein Mass für die Streuung um den Mittelwert der Anzahl fehlerhaft übertragener Bits.

(**e**) Die Nullhypothese lautet

$$H_0 \;:\; \pi_0 = 0.1 \,.$$

Die Alternativhypothese ist dann

$$H_A \;:\; \pi < 0.1 \,.$$

(**f**) **1.** Modell: X ist die Anzahl fehlerhaft übertragener Bits, $X \sim \text{Bin}(100, \pi)$.

2. Die Nullhypothese ist $H_0 : \pi = 0.1$, die Alternative ist $H_A : \pi < 0.1$.

3. Die Teststatistik (unter der Annahme von H_0) ist $T : P(T = t) = \binom{100}{t} 0.1^t 0.9^{100-t}$

4. Das Signifikanzniveau ist $\alpha = 0.05$.

5. Verwerfungsbereich: Die Werte in der Tabelle können mit `R` ermittelt werden

	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
$P(T \leq t)$	0.000027	0.00032	0.001944	0.00784	0.02371	0.0575	0.1172

```
> pbinom(0:10,100,prob=0.1)
 [1] 0.0000265614 0.0003216881 0.0019448847 0.0078364871 0.0237110827
 [6] 0.0575768865 0.1171556154 0.2060508618 0.3208738884 0.4512901654
[11] 0.5831555123
```

Daher ist der Verwerfungsbereich $K = \{0, 1, 2, 3, 4\}$.

6. Testentscheid: Da $6 \notin K$ wird H_0 nicht verworfen. Eine geringere Fehlerrate mit dem neuem Material für den Übertragungskanal kann nicht nachgewiesen werden.

Der Test lässt sich ebenfalls mit `R` durchführen

```
> binom.test(6,100,p=0.1,alternative="less")

        Exact binomial test

data:  6 and 100
number of successes = 6, number of trials = 100, p-value = 0.1172
alternative hypothesis: true probability of success is less than 0.1
95 percent confidence interval:
 0.0000000 0.1149853
sample estimates:
probability of success
                  0.06
```

(**g**) Unter der Alternativhypothese ist $X \sim \text{Bin}(100, 0.05)$ und die Macht des Tests ist

$$P(X \leq 4)\,,$$

da 4 die obere Grenze des Verwerfungsbereichs ist. Somit ist die Macht des Tests

```
> pbinom(4,100,p=0.05)

[1] 0.4359813
```

(**h**) Das (einseitige) Vertrauensintervall für π entnehmen wir dem R-Output des Binomialtests:

$$[0, 0.11]\,.$$

Das (zweiseitige) Vertrauensintervall für π lässt sich mit der Normalapproximation berechnen

$$I \approx \frac{x}{n} \pm 1.96\sqrt{\frac{x}{n}\left(1 - \frac{x}{n}\right)\frac{1}{n}}\,,$$

also I=[0.0135,0.1065] . Da $\pi = 0.1$ im Vertrauensintervall ist, kann die Nullhypothese nicht verworfen werden.

Aufgabe 3

(**a**) Da dasselbe Individuum (Versuchseinheit) zwei unterschiedliche Automobile (Versuchsbedingungen) testet, handelt es sich um einen **gepaarten** Test.

(**b**) Die Nullhypothese lautet: die Einparkierzeiten der beiden Automobile unterscheiden sich nicht. Bilden wir die Differenz zwischen jedem Paar von Beobachtungen $D_i = X_i - Y_i$ mit $i = 1, \ldots, n$, dann bezeichnet $\mu_D = \mu_X - \mu_Y$ den Mittelwert der Differenz, und die Nullhypothese lautet somit $H_0 \;:\; \mu_D = \mu_0 = 0$. Wir betrachten eine zweiseitige Alternativhypothese: $H_A \;:\; \mu_D \neq 0$.

(**c**) Die Varianz des Differenzendatenvektors d_i kann geschätzt werden durch

$$\frac{1}{14-1}\sum_{i=1}^{14}(d_i - \bar{d}_{14})^2\,,$$

wobei $\bar{d}_{14}$ den geschätzten oder arithmetischen Mittelwert der Differenzen bezeichnet:

$$\bar{d}_{14} = \frac{1}{14}\sum_{i=1}^{14} d_i\,.$$

Somit ergibt sich für $\widehat{\sigma}^2_D$

```
> x <- c(37.0,25.8,16.2,24.2,22.0,33.4,23.8,58.2,33.6,24.4,23.4,21.2,36.2,29.8)
> y <- c(17.8,20.2,16.8,41.4,21.4,38.4,16.8,32.2,27.8,23.2,29.6,20.6,32.2,53.8)
> d <- x - y
> var(d)
```

```
[1] 160.9075

> sd(d)^2

[1] 160.9075
```

(**d**) Unter der Nullhypothese sind die Differenzen zwischen jedem Beobachtungspaar verteilt gemäss

$$D_i \text{ i.i.d} \sim \mathcal{N}(\mu_0, \sigma_D^2) \,.$$

Wir definieren die Teststatistik als

$$T = \frac{\bar{D}_n - \mu_0}{\widehat{\sigma_{\bar{D}_n}}} = \frac{\bar{D}_n - \mu_0}{\widehat{\sigma_D}/\sqrt{n}} \,,$$

wobei $\sigma_{\bar{D}_n}$ der **Standardfehler** des arithmetischen Mittelwertes $\bar{D}_n$ ist. Unter der Nullhypothese H_0 ist die Verteilung von T gegeben durch

$$T \;\sim\; t_{n-1} \,.$$

Auf dem **Signifikanzniveau** $\alpha = 5\%$ ergibt sich folgender **Verwerfungsbereich**

$$K = \{t \;:\; t < t_{0.025,13} \;\text{ oder }\; t > t_{0.975,13}\} = [-\infty, -2.16] \cup [2.16, \infty]$$

Die 2.5% und 97.5% Quantilen der t_{13}-Verteilung lässt sich mit `R` folgendermassen bestimmen

```
> qt(0.025,13)

[1] -2.160369

> qt(0.975,13)

[1] 2.160369
```

Der **Wert der Teststatistik** ist

$$t = \frac{\bar{D}_n - \mu_0}{\widehat{\sigma_D}/\sqrt{n}} = \frac{1.21 - 0}{12.7/\sqrt{14}} = 0.358 \,.$$

Da der Wert der Teststatistik t nicht im Verwerfungsbereich K liegt, lautet der **Testentscheid**: die Nullhypothese wird auf dem Signifikanzniveau $\alpha = 5\%$ beibehalten. Die Mechanik des ersten Automobils ermöglicht also nicht signifikant schnelleres Einparkieren.

(**e**) Das 90% Vertrauensintervall für den (wahren) Mittelwert $\mu_D = \mu_X - \mu_Y$ ermitteln wir durch folgende Ungleichung

$$\begin{aligned}
\bar{D}_n - t_{0.95,13} \cdot \widehat{\sigma_D}/\sqrt{n} \;&\leq \mu_D \leq\; \bar{D}_n + t_{0.95,13} \cdot \widehat{\sigma_D}/\sqrt{n} \\
\bar{d}_{14} - t_{0.95,13} \cdot \widehat{\sigma_D}/\sqrt{14} \;&\leq \mu_D \leq\; \bar{d}_{14} + t_{0.95,13} \cdot \widehat{\sigma_D}/\sqrt{14} \\
1.21 - 1.77 \cdot 12.687/\sqrt{14} \;&\leq \mu_D \leq\; 1.21 + 1.77 \cdot 12.687/\sqrt{14}
\end{aligned}$$

Also ist das 90% Vertrauensintervall $[-4.79, 7.21]$. Das 95% Quantil der t_{13}-Verteilung ermittelt sich mit `R`

```
> qt(0.95,13)

[1] 1.770933
```

Das 90% Vertrauensintervall beinhaltet den Wert $\mu_D = 0$, d.h., dieser Wert ist nicht inkonsistent mit den Daten.

(**f**) Die Annahme der Normalverteilung der Einparkzeiten mit bekannter Standardabweichung σ führt zu einem z-Test. Die Teststatistik Z ist nun normalverteilt: Unter Annahme der Nullhypothese gilt

$$Z = \frac{\bar{D}_n - \mu_0}{\sigma_{\bar{D}_n}} = \frac{\bar{D}_n - \mu_0}{\sigma_D/\sqrt{n}} \sim \mathcal{N}(0,1)\,.$$

Also ist das 90% Vertrauensintervall in diesem Fall

$$\begin{aligned} \bar{D}_n - \Phi(0.95)\sigma_D/\sqrt{n} &\leq \mu_D \leq \bar{D}_n + \Phi(0.95)\sigma_D/\sqrt{n} \\ 1.21 - 1.64 \cdot 12.687/\sqrt{14} &\leq \mu_D \leq 1.21 + 1.64 \cdot 12.687/\sqrt{14} \\ -4.35 &\leq \mu_D \leq 6.77 \end{aligned}$$

Der Wert von $\Phi(0.95)$, das 95%-Quantil der Standardnormalverteilung, ermittelt sich mit `R` durch

```
> qnorm(0.95)

[1] 1.644854
```

Der t-Test kann mit `R` auch direkt durchgeführt werden

```
> t.test(x,y,paired=TRUE,alternative="two.sided",conf.level=0.9)

        Paired t-test

data:  x and y
t = 0.3582, df = 13, p-value = 0.726
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -4.789516  7.218087
sample estimates:
mean of the differences
               1.214286
```

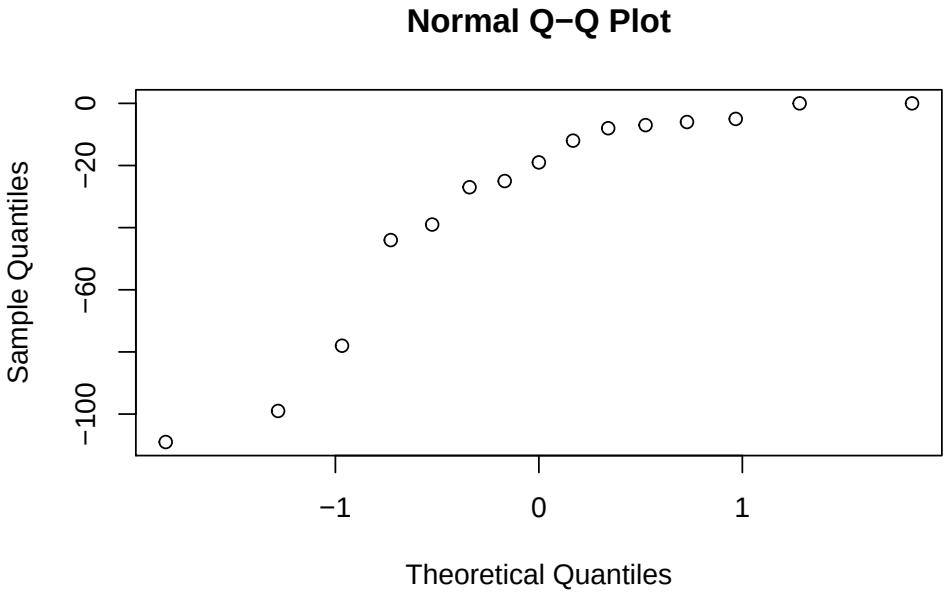
Aufgabe 4

(**a**) Beim selben Patienten werden zwei verschiedene Versuchsbedingungen getestet – einmal die Nachbehandlung mit Marihuana, dann die Nachbehandlung mit einem Placebo, es handelt sich also um einen **gepaarten** Test.

(**b**) Die Nullhypothese lautet: Es gibt keinen Unterschied zwischen den beiden Nachbehandlungen in Bezug auf Anzahl Erbrechen- und Würgereflexen. Mit X_i bezeichnen wir die Anzahl Erbrechen- und Würgereflexe beim $i-$ten Patienten nach der Behandlung mit Marihuana, mit Y_i die Anzahl Erbrechen- und Würgereflexe beim $i-$ten Patienten nach der Behandlung mit einem Placebo. Bilden wir die Differenz zwischen jedem Paar von Beobachtungen $D_i = X_i - Y_i$ mit $i = 1, \ldots, n$, dann bezeichnet $\mu_D = \mu_X - \mu_Y$ den Mittelwert der Differenz, und die Nullhypothese lautet somit $H_0 \ : \ \mu_D = \mu_0 = 0$. Da wir zeigen möchten, dass die Nachbehandlung mit Marihuana zu einer Reduktion der Anzahl Erbrechen- und Würgereflexen führt, betrachten wir die einseitige Alternativhypothese: $H_A \ : \ \mu_D < 0$.

(**c**) Um zu entscheiden, welchen Test wir durchführen können, müssen wir überprüfen, ob die Daten normalverteilt sind. Wir zeichnen deswegen einen Normal Q-Q Plot auf.

```
> x <- c(15,25,0,0,4,2,1,4,9,0,22,11,0,0,0)
> y <- c(23,50,0,99,31,21,79,113,53,0,61,18,12,6,5)
> qqnorm(x-y)
```



Aufgrund der S-förmigen Verteilung der Datenpunkte im Normal-Q-Q-Plot können wir nicht davon ausgehen, dass die Daten normalverteilt ist. Der t-Test bietet sich hier also nicht an. Entweder führen Sie in diesem Fall also einen Wilcoxon-Test oder Vorzeichen-Test durch.

(**d**) Wir führen einen Wilcoxon-Test durch (streng genommen müssten wir noch überprüfen, ob die Daten symmetrisch verteilt sind). Wir benützen dazu `R`:

```
> wilcox.test(x,y,paired=TRUE,alternative="less")

        Wilcoxon signed rank test with continuity correction

data:  x and y
```

```
V = 0, p-value = 0.0008308
alternative hypothesis: true location shift is less than 0
```

Aufgrund des p-Wertes (0.0008308) können wir schliessen, dass das Ergebnis der Doppelblindstudie sehr unwahrscheinlich ist unter der Annahme, dass die Nullhypothese wahr ist. Wir schliessen also daraus, dass bei einer Chemotherapie die Nachbehandlung mit Marihuana zu einer signifikant geringeren Anzahl Erbrechen- und Würgereflexen führt. Der Vorzeichentest würde für den p-Wert 0.00369 liefern (13 von 15 Differenzen sind grösser als null).

Aufgabe 5

(a)(b)(c)(d) Die Werte können direkt mit R ermittelt werden :

```
> wine <- c(2.8,3.2,3.2,3.4,4.3,4.9,5.1,5.2,5.9,5.9,6.6,8.3,12.6,15.1,
+         25.1,33.1,75.9,75.9)
> mort <- c(6.2,9.0,7.1,6.8,10.2,7.8,9.3,5.9,8.9,5.5,7.1,9.1,5.1,4.7,
+         4.7,3.1,3.2,2.1)
> reg <- lm(mort~wine)
> summary(reg)

Call:
lm(formula = mort ~ wine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0683 -1.3616 -0.2138  1.4897  2.8406

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.68655    0.47332  16.240 2.31e-11 ***
wine        -0.07608    0.01700  -4.475 0.000383 ***
---
Signif. codes:  0
```

Aus dem R-Output folgt, dass $\widehat{\beta}_1 = -0.07608$ und $\widehat{\beta}_0 = 7.68655$. β_0 ist der y-Achsenabschnitt und kann interpretiert werden als die Anzahl Todesfälle pro 1000 aufgrund von Herzkreislauferkrankung in einem Land, in dem kein Alkohol getrunken wird (z.B. Saudi-Arabien). Dies ist allerdings eine Extrapolation, bei der Vorsicht geboten ist. Der Standardfehler von $\widehat{\beta}_1$ ist 0.01700 und der residual standard error wurde aufgrund von $18 - 2 = 16$ Freiheitsgraden berechnet.

(**e**) Das 99% Vertrauensintervall für β_1 hat die Form

$$\widehat{\beta}_1 \pm t_{0.995,n-2} \cdot \widehat{s.e.}(\hat{\beta}_1) \,,$$

also $-0.07608 \pm 2.920782 \cdot 0.01700 : [-0.1257333, -0.02642671]$. Da der Wert 0 nicht in diesem Vertrauensintervall enthalten ist, können wir davon ausgehen, dass es einen Zusammenhang zwischen Weinkonsum und Todesfälle aufgrund von Herzkreislauferkrankungen gibt.

(**f**) Der p-Wert aufgrund der Nullhypothese $H_0 : \beta_1 = 0$ kann aus dem `R`-Output ermittelt werden: 0.000383. Die Nullhypothese kann also verworfen werden.

(**g**) Das Resdiduum für die Schweiz berechnet sich wie folgt

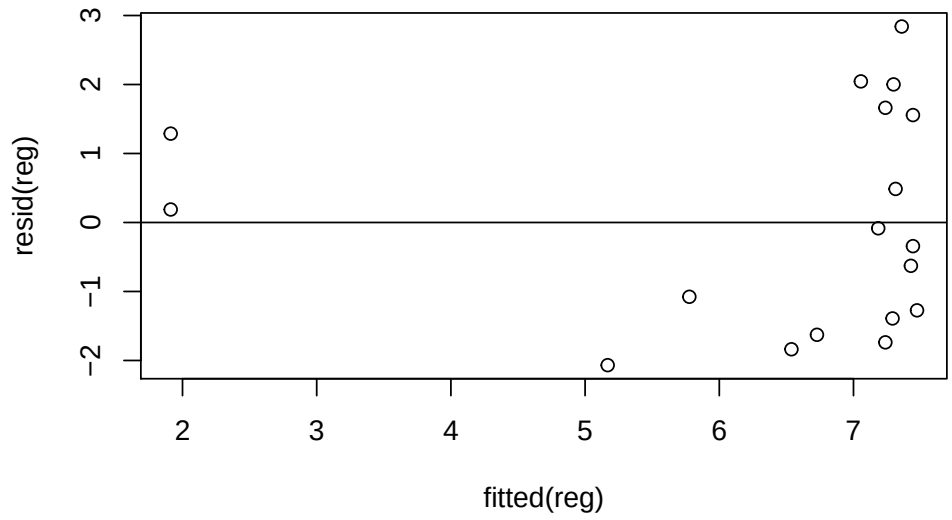
$$R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \,,$$

was für die Schweiz

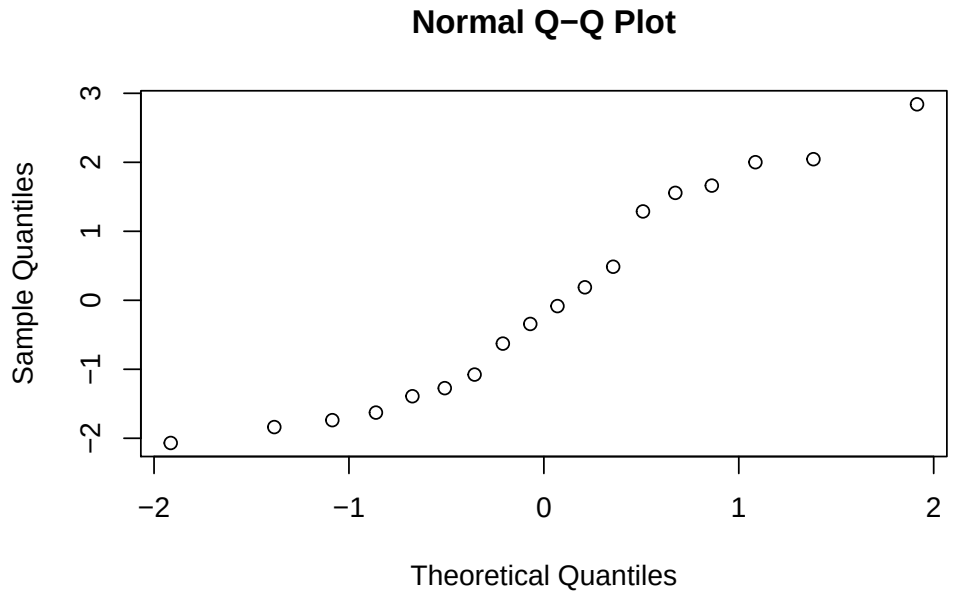
$$R_i = 3.1 - (7.68655 - 0.07608 \cdot 33.1) = -2.068$$

ergibt.

(**g**)
```
> plot(fitted(reg),resid(reg))
> abline(h=0)
```




```
> qqnorm(resid(reg))
```



Aus dem Tukey-Anscombe Plot schliessen wir, dass die Fehlervarianz nicht konstant ist (U-förmige Kurve). Allerdings können wir aus dem Normal Q-Q Plot schliessen, dass die Normalverteilungsannahme zutrifft.

Aufgabe 6

1. (a), denn

   $P(A^C \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = (1 - P(A))P(B) = P(A^C)P(B)$.

2. (a), denn $P(A \cap B) = 0$.

3. (a). Da es drei Karten mit je zwei Seiten gibt, gibt es a priori 6 mögliche Elementarereignisse. Davon fallen drei weg, weil wir wissen, das die Karte oben schwarz ist, und von diesen drei Ereignissen führt nur eines (das, welches die schwarz-weisse Karte beinhaltet) zum geforderten Resultat.

4. (a). Nutzen Sie die Linearität des Integrals.

5. (c). Nutzen Sie die Linearität des Integrals.

6. (c). Nutzen Sie $\frac{d}{dx}F(x) = f(x)$.

7. (c). Wenn $x \longrightarrow \infty$, muss gelten, dass $F(x) \longrightarrow 1$. Das ist nur mit $\alpha = 1$ möglich.

8. (c), Definition des Signifikanzniveaus.

9. (d), denn $P(K|T) = P(T|K)P(K)/(P(T|K)P(K) + P(T|K^C)P(K^C))$.

10. (a), da n gross und π klein.