

# Tutorial 1

Lars Abt

## Aufgabe 1: Daten einlesen

In diesem Abschnitt lesen wir unsere Titanic-Daten ein. Je nach Ordnerstruktur kann dies unterschiedlich aussehen.

```
getwd()
```

```
[1] "C:/Users/abtla/Documents/WHK/Angewandte Statistik/Muster/Tutorial 1/Code"
```

```
# install.packages("data.table")
library(data.table)
titanic_data <- fread("../Daten/titanic.csv")
# titanic_data <- read.csv("../Daten/titanic.csv")
# Falls keine data.table Struktur gewünscht ist, funktioniert read.csv ebenso.

# data.table::fread("../Daten/titanic.csv")
# mit dieser Methode wird der Datensatz in eine data.table Struktur eingelesen,
# ohne den 'library' Befehl zu benötigen.

# Falls R Skripte benutzt werden, kann der Pfad auch so aussehen:
# titanic_data <- fread("Daten/titanic.csv")
```

Wir haben den Datensatz erfolgreich eingelesen.

## Aufgabe 2: Überlebenschancen

Haben mehr Männer oder Frauen überlebt? Wir überprüfen dies mit Hilfe von “table()”.

```
gender_table <- table(titanic_data$Sex, titanic_data$Survived)
gender_table
```

```
      0    1
female 81 233
male   468 109
```

```
survived_table <- table(titanic_data$Sex[titanic_data$Survived == 1])
survived_table
```

```
female    male
      233     109
```

Es haben 233 Frauen und 109 Männer überlebt. Gestorben sind 81 Frauen und 468 Männer. Frauen hatten somit eine deutlich höhere Überlebensrate.

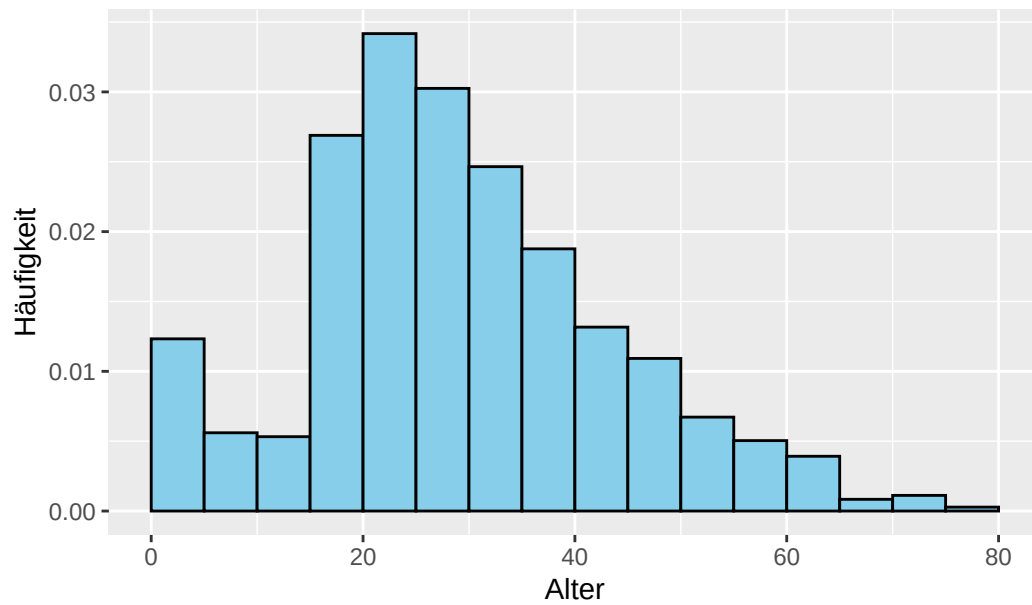
### Aufgabe 3: Histogramme

Wir möchten Histogramme erstellen. Ein Histogramm für die allgemeine Altersverteilung und drei weitere für die verschiedenen Klassen.

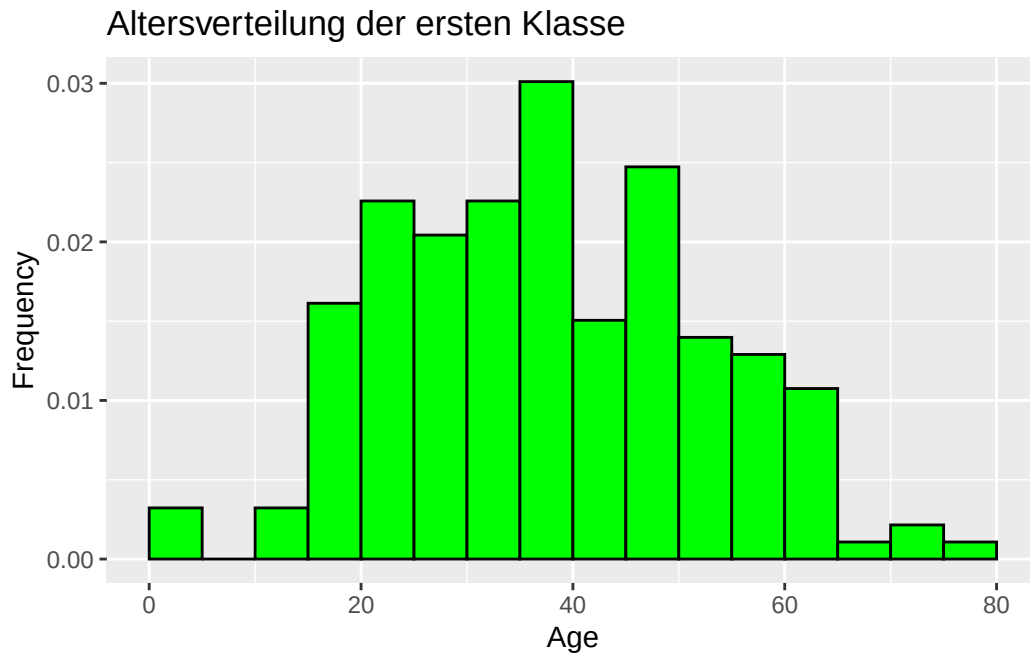
```
# install.packages("ggplot2")
library(ggplot2)

# Allgemeines Histogramm
ggplot(titanic_data, aes(x = Age, y = after_stat(density))) +
  geom_histogram(breaks = seq(0, 80, by = 5), fill = "skyblue", color = "black",
                 na.rm = TRUE) +
  labs(title = "Altersverteilung aller Passagiere",
       x = "Alter",
       y = "Häufigkeit")
```

Altersverteilung aller Passagiere

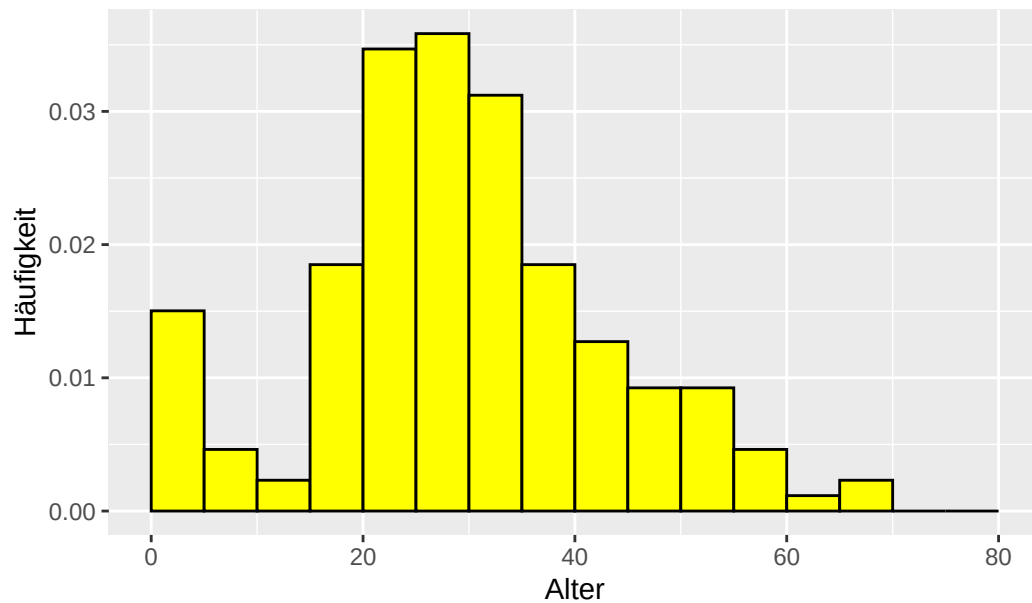


```
# Histogramm für die erste Klasse
first_class <- titanic_data[titanic_data$Pclass == 1,]
ggplot(first_class, aes(x = Age, y = after_stat(density))) +
  geom_histogram(breaks = seq(0, 80, by = 5), fill = "green", color = "black",
    na.rm = TRUE) +
  labs(title = "Altersverteilung der ersten Klasse",
    x = "Age",
    y = "Frequency")
```

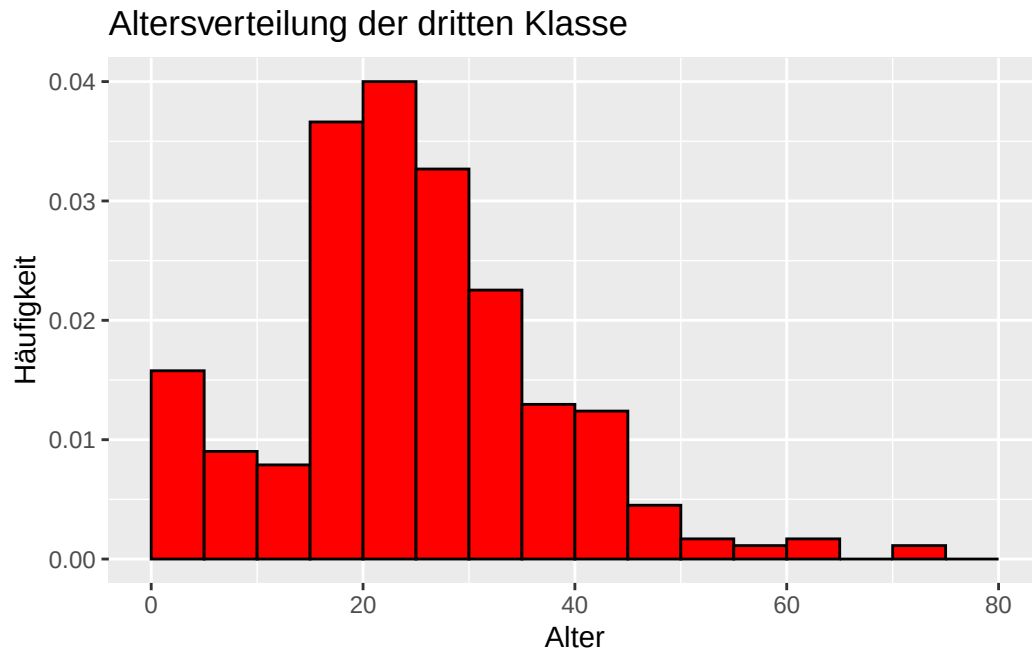


```
# Histogramm für die zweite Klasse
second_class <- titanic_data[titanic_data$Pclass == 2,]
ggplot(second_class, aes(x = Age, y =after_stat(density))) +
  geom_histogram(breaks = seq(0, 80, by = 5), fill = "yellow", color = "black",
    na.rm = TRUE) +
  labs(title = "Altersverteilung der zweiten Klasse",
    x = "Alter",
    y = "Häufigkeit")
```

Altersverteilung der zweiten Klasse

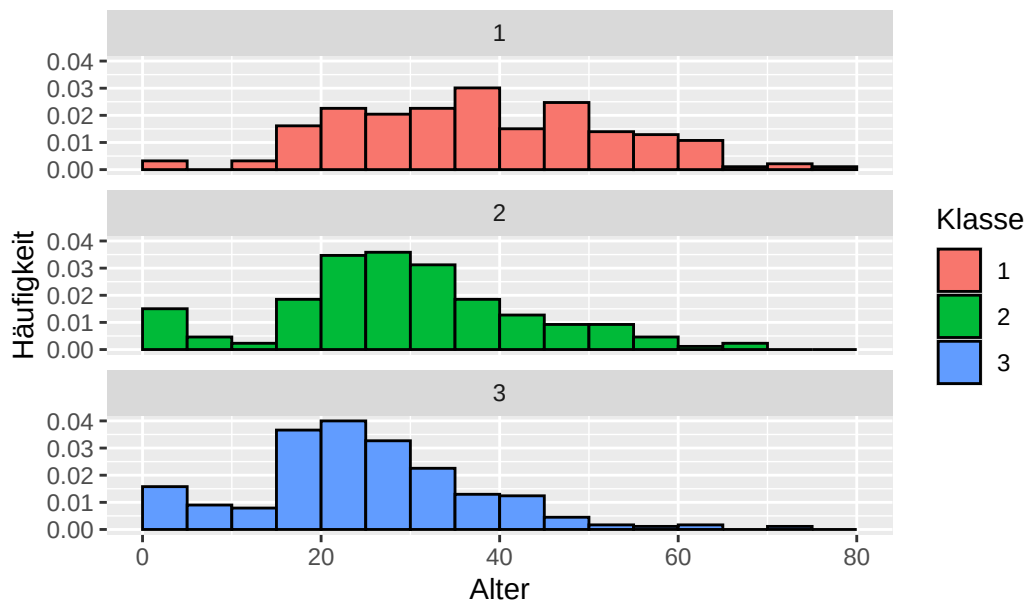


```
# Histogramm für die dritte Klasse
third_class <- titanic_data[titanic_data$Pclass == 3,]
ggplot(third_class, aes(x = Age, y = after_stat(density))) +
  geom_histogram(breaks = seq(0, 80, by = 5), fill = "red", color = "black",
    na.rm = TRUE) +
  labs(title = "Altersverteilung der dritten Klasse",
    x = "Alter",
    y = "Häufigkeit")
```



```
# Histogramme für alle Klassen
ggplot(titanic_data, aes(x = Age, y =after_stat(density),
                        fill = factor(Pclass))) +
  geom_histogram(breaks = seq(0, 80, by = 5), color = "black", na.rm = TRUE) +
  facet_wrap(~factor(Pclass), ncol = 1) +
  labs(title = "Altersverteilung für die verschiedenen Klassen",
       x = "Alter",
       y = "Häufigkeit",
       fill = "Klasse")
```

### Altersverteilung für die verschiedenen Klassen



```
# GGPlot2 Befehle können unterschiedlich aussehen. Insbesondere der "aes"-Aspekt  
# kann auch an anderer Stelle eingefügt werden.
```

### Aufgabe 4: Ticketpreise

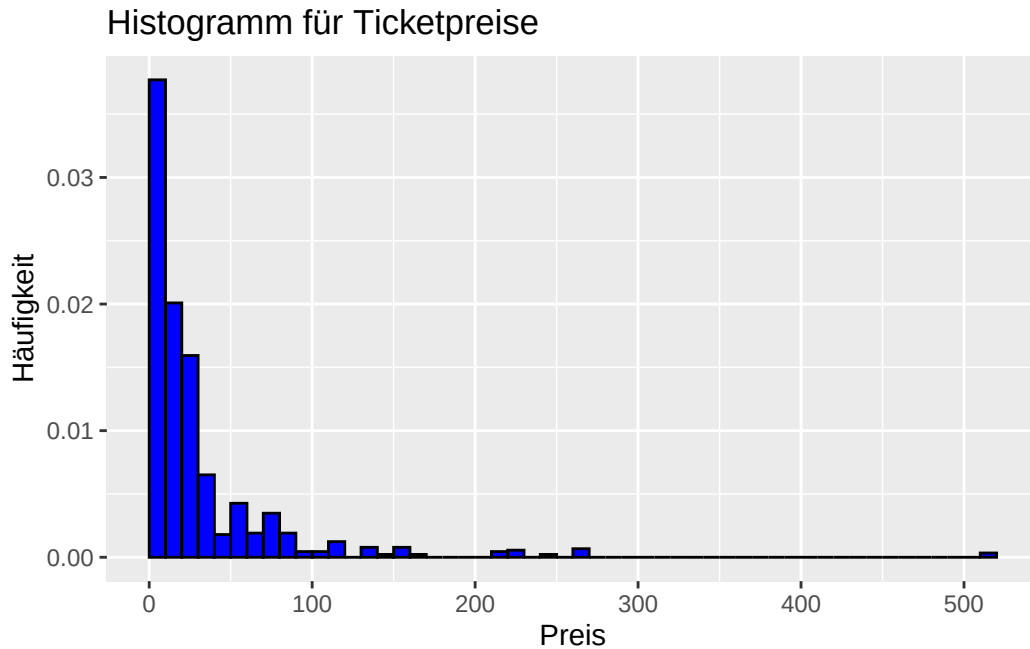
Abschließend möchten wir noch mit den Ticketpreisen arbeiten. Insbesondere interessiert uns, wie hoch die Wahrscheinlichkeit ist, dass wir zufällig eine Person mit einem Ticketpreis von über 100\$ ziehen. Zu welcher Verteilung könnte dies passen?

```
fare_over_100 <- sum(titanic_data$Fare > 100)  
  
probability_over_100 <- fare_over_100 / nrow(titanic_data)  
  
print(probability_over_100)
```

```
[1] 0.05948373
```

Bei zufälliger Ziehung eines Passagiers haben wir eine Chance von ungefähr 6%, dass dieser einen Ticketpreis über 100€ hatte.

```
ggplot(titanic_data, aes(x = Fare, y = after_stat(density))) +
  geom_histogram(breaks = seq(0, 520, by=10), fill = "blue", color = "black",
    na.rm = TRUE) + labs(title = "Histogramm für Ticketpreise",
    x = "Preis",
    y = "Häufigkeit")
```



Dies kann an eine Exponentialverteilung erinnern. Diese beschreibt jedoch normalerweise Wartezeiten. Es gibt keine eindeutige Antwort.