

Statistik und Qualität - Ausarbeitung der ersten Übung

Stefan Dünser

A) TECHNISCHE FERTIGKEITEN

A-1) Handelt es sich bei den vorliegenden statistischen Gesamtheiten um Bestands- oder Bewegungsgrößen?

a) Studierende an einer Hochschule. * Bestandsgröße b) Hochzeiten am Standesamt einer Gemeinde. * Bewegungsgröße c) Bei der Behörde gemeldete Personenkraftwagen. * Bestandsgröße d) Maschinenausfälle in einer Werkstatt. * Bewegungsgröße e) Wartende Kunden vor einem Abfertigungsschalter. * Bestandsgröße

A-2) Im Servicecenter eines Unternehmens werden über einen Zeitraum eines Tages die eingehenden Anrufe aufgezeichnet. Gezählt wird die Anzahl der pro 10-Minuten-Zeitintervall eingehenden Anrufe. Für 40 derartige Zeitintervalle erhält man folgende Ergebnisse:

```
# Erstellen einer Urliste
Liste.vec <- c(0, 0, 1, 3, 4, 1, 2, 2, 1, 1, 1, 2, 3, 0, 2, 0, 1, 3, 1, 2, 2, 0, 1, 1, 6, 1, 0, 2, 3, 1, 1, 4, 2, 3, 2, 0, 3, 0)
Liste.vec

## [1] 0 0 1 3 4 1 2 2 1 1 1 2 3 0 2 0 1 3 1 2 2 0 1 1 6 1 0 2 3 1 1 4 2 3 2 0 3 0
## [39] 1 2
```

a) Was stellt bei dieser Fragestellung die statistische Grundgesamtheit dar? Was sind die beobachteten Merkmale der statistischen Einheiten und wie sind sie skaliert?

- Grundgesamtheit
 - Die Grundgesamtheit setzt sich aus den statistischen Einheiten zusammen. Die Grundgesamtheit sind die Anzahl der 10-Minuten-Zeitintervalle - (40).
- Merkmale
 - Anzahl der 10-Minuten-Zeitintervalle (40)
 - Anzahl der Anrufe über alle Zeitintervalle (65)
- Skalierung
 - Metrische Skalierung (Kardinalskala), die verhältnisskaliert ist

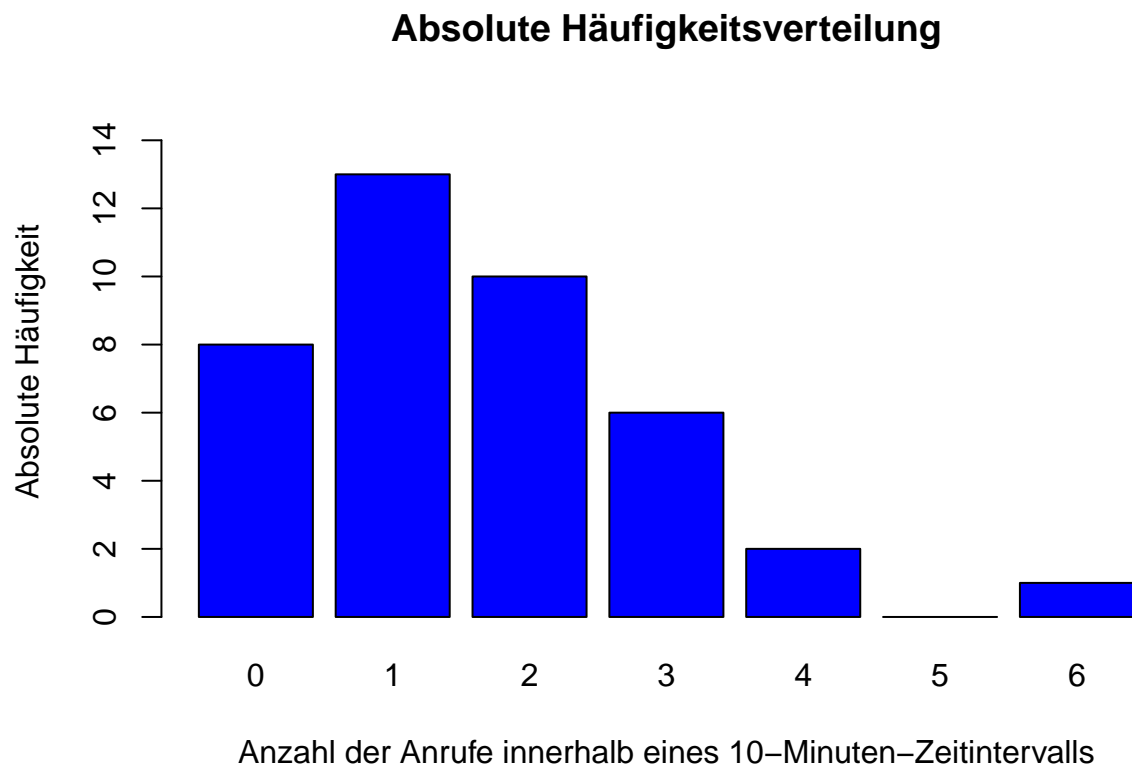
b) Ermittle die absolute und relative Häufigkeitstabelle der eingehenden Anrufe und stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.

Absolute Häufigkeit:

```
table(Liste.vec)
```

```
## Liste.vec
##  0  1  2  3  4  6
##  8 13 10  6  2  1
```

```
# Beim folgenden Barplot muss die x-Achse separat noch einmal mit den Werten von 0 bis 6 initialisiert werden
barplot(table(factor(Liste.vec, levels=c(0,1,2,3,4,5,6))), ylim = c(0,15), xlab = "Anzahl der Anrufe innerhalb eines 10-Minuten-Zeitintervalls")
```

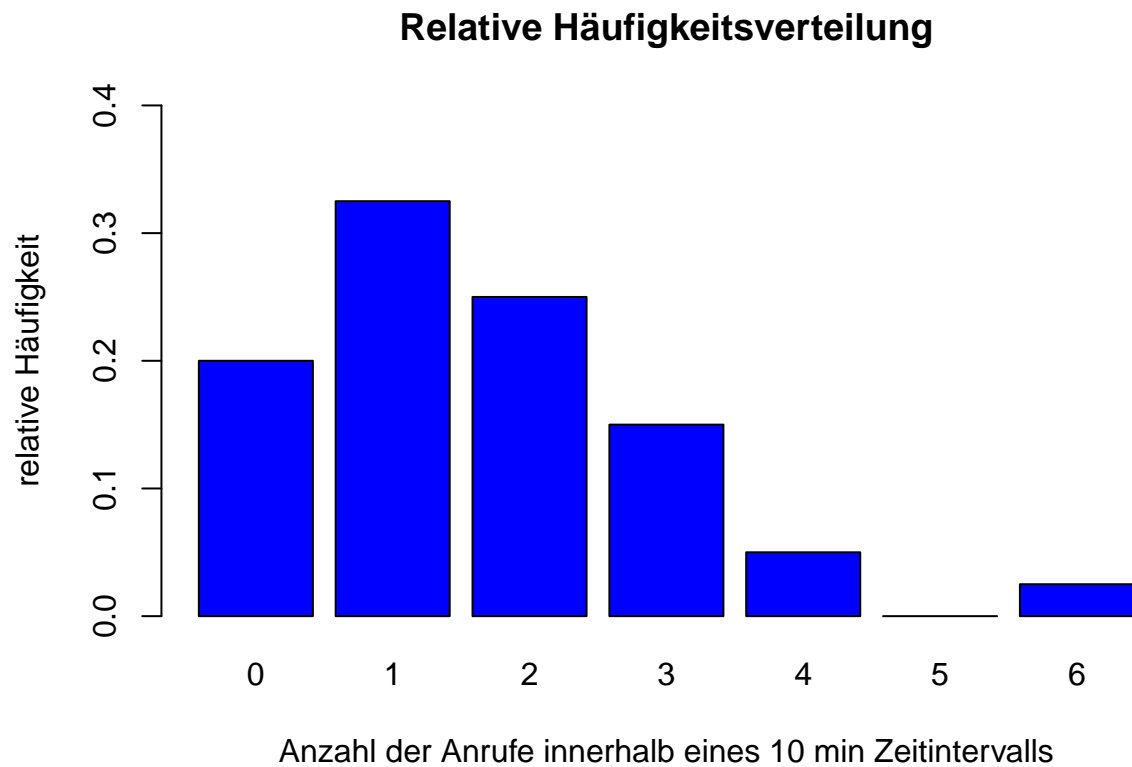


Relative Häufigkeit:

```
# Für die relative Häufigkeit wird die Absolute Häufigkeit durch die Anzahl der Elemente im Vektor dividiert
table(Liste.vec)/length(Liste.vec)
```

```
## Liste.vec
##    0    1    2    3    4    6
## 0.200 0.325 0.250 0.150 0.050 0.025
```

```
# Auch hier wieder die separate Definition des x-Vektors für die Darstellung mit 5.
barplot(table(factor(Liste.vec, levels=c(0,1,2,3,4,5,6)))/sum(table(Liste.vec)), ylim = c(0,0.4), xlab = "Anzahl der Anrufe innerhalb eines 10-Minuten-Zeitintervalls")
```

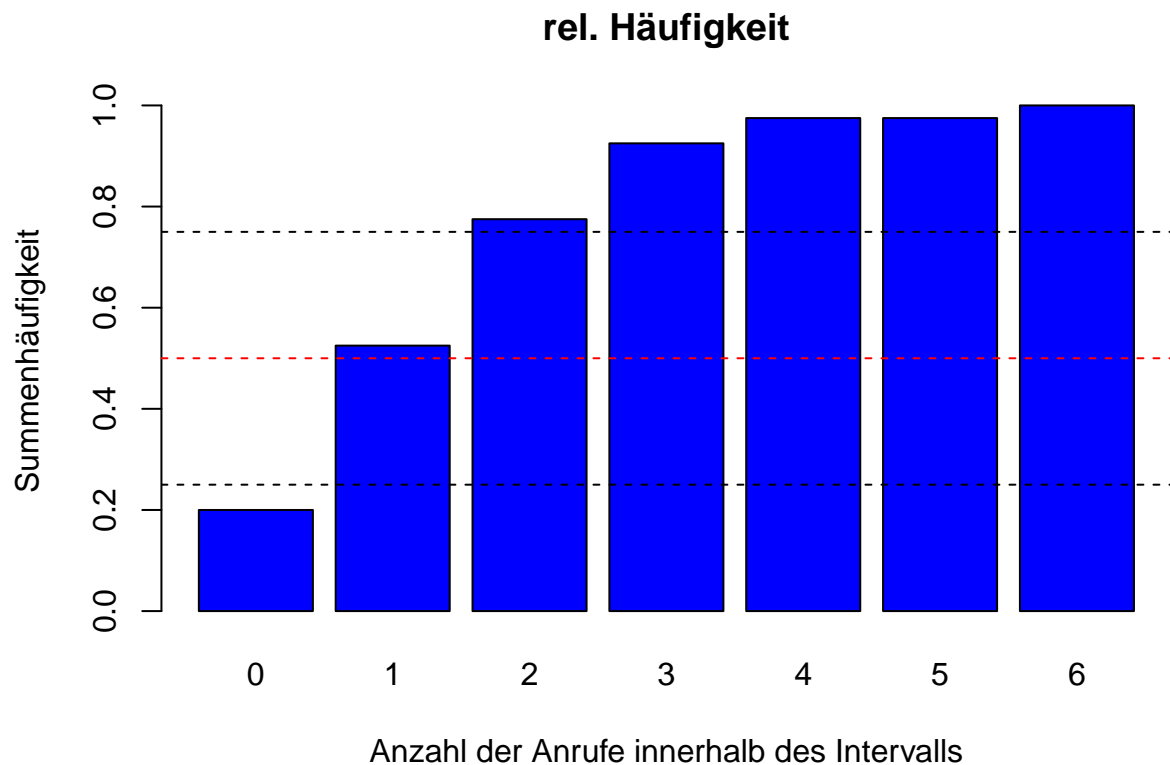


Summenhäufigkeit

```
cumsum(table(Liste.vec))
```

```
## 0 1 2 3 4 6  
## 8 21 31 37 39 40
```

```
barplot(cumsum(table(factor(Liste.vec, levels = c(0,1,2,3,4,5,6))))/sum(table(Liste.vec)), xlab = "Anzahl der Anrufe",  
        abline(0.25, 0, lty = "dashed")  
        abline(0.5, 0, lty = "dashed", col = "red")  
        abline(0.75, 0, lty = "dashed")
```

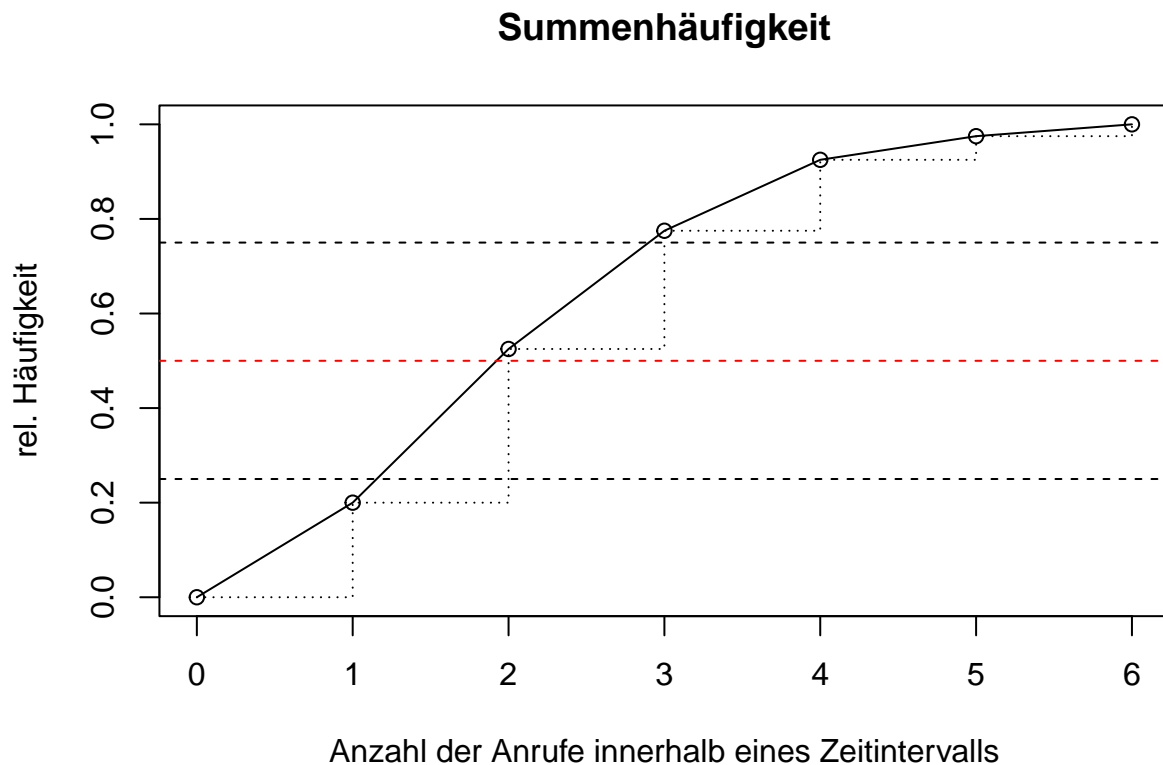


Die Darstellung der Summenhäufigkeit wird meist in einem Liniendiagramm gemacht. Eine solche Darstellung ist in der nachfolgenden Grafik ersichtlich.

Andere Darstellung der Summenhäufigkeit

```
Breaks.vec <- c(0:6)
SummHfk.vec <- cumsum(table(Liste.vec))/sum(table(Liste.vec))

# plot-Funktion erzeugt ein Punktdiagramm
# Für die bessere Lesbarkeit werden die einzelnen Punkte durch eine Linie miteinander verbunden
plot(Breaks.vec, c(0,SummHfk.vec), main = "Summenhäufigkeit", xlab = "Anzahl der Anrufe innerhalb eines
points(Breaks.vec, c(0,SummHfk.vec)) # Darstellung der Werte als Punkte
lines(Breaks.vec, c(0,SummHfk.vec), type = "s", lty = 3) # Verbindungslinie zwischen den Punkten
abline(0.25, 0, lty = "dashed") # unteres Quartil
abline(0.75, 0, lty = "dashed") # oberes Quartil
abline(0.5, 0, lty = "dashed", col = "red") # Median
```



c) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

Arithmetisches Mittel: Das arithmetische Mittel kann als Schwerpunkt der Häufigkeitsverteilung angesehen werden. In diesem Fall beträgt das arithmetische Mittel nach Abschätzung rund 1,5, da somit links und rechts der x-Achse in etwa gleich viele Werte sind.

Median: Der Median wird über die y-Achse ermittelt. Der Median befindet sich dann auf der x-Achse an dem Punkt, an dem 50% des y-Werts erreicht sind. In diesem Fall ist der Median bei 1.

d) Berechne die möglichen Lagemaße (Zentralmaße und Streumaße)

Zentralmaße:

Modus Der Modus beschreibt den Wert, der in der Werteliste am häufigsten vorkommt.

```
# Für den Modus gibt es in R keinen Befehl, weshalb hier die Ermittlung über eine Funktion erfolgt.
getmode <- function(Liste.vec) {
  uniqv <- unique(Liste.vec)
  uniqv[which.max(tabulate(match(Liste.vec, uniqv)))]
}
mode <- getmode(Liste.vec)
mode
```

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

Median

Der Median beschreibt “den Wert in der Mitte”. Er ist der Zentralwert aller Werte in der Werteliste.

```
median(Liste.vec)
```

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

Arithmetisches Mittel

Das arithmetische Mittel, oder Durchschnitt, bezeichnet den Mittelwert aller Werte in einer Werteliste.

```
mean(Liste.vec)
```

```
## [1] 1.625
```

Streu Maße:

Minimum

Das Minimum ist das kleinste Element in einer Liste

```
Minimum <- min(Liste.vec)  
Minimum
```

```
## [1] 0
```

Maximum

Das Maximum ist der größte Wert in einer Liste.

```
Maximum <- max(Liste.vec)  
Maximum
```

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

Spannweite

Die Spannweite misst die Streuung in einer Beobachtung von ordinalskalierten Merkmalen.

```
Spannweite <- range(Liste.vec)  
Spannweite
```

```
## [1] 0 6
```

Quantile

Ähnlich wie der Median teilt ein Quantil die Werteliste in 100 gleich große Einheiten. Der Median an sich ist auch ein Quantil, nämlich das Quantil bei 50. Das Quantil teilt die Werte nicht in 100 sondern in 4 Einheiten.

```
quantile(Liste.vec)
```

```
##    0%   25%   50%   75%  100%  
##    0    1    1    2    6
```

Mittlere absolute Abweichung

Die mittlere absolute Abweichung gibt den Wert an, um die die Stichprobe um den Median schwankt.

```
mean(abs(Liste.vec-mean(Liste.vec)))
```

```
## [1] 1.05625
```

Standardabweichung

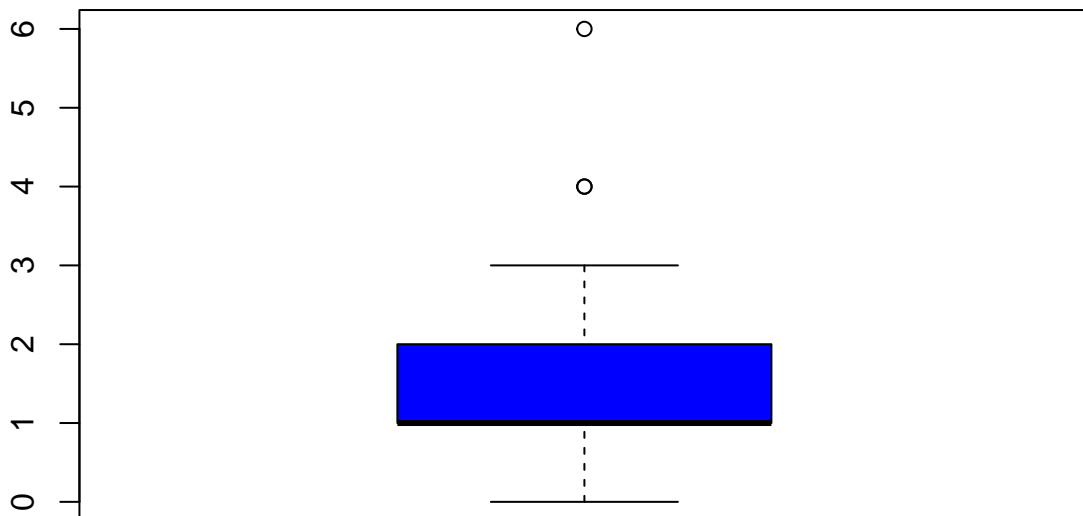
Die Standardabweichung ist ein Maß für die Streuung der Wahrscheinlichkeitsdichte um ihren Schwerpunkt.

```
sd(Liste.vec)
```

```
## [1] 1.333734
```

e) Stelle die Daten in einem Boxplot dar

```
boxplot(Liste.vec, col = "blue")
```



Besonders Auffällig ist bei diesem Boxplot bzw. den Werten aus der Liste, dass der Median und das erste Quartil (25% Quantil) zusammenfallen. Die Werte 4 Anrufe sowie 6 Anrufe innerhalb eines 10-Minuten_Zeitintervalls werden hier als Ausreißer aufgefasst und als Kreis dargestellt.

A-3) Eine Anzahl von 1000 Kleinmotoren weist folgende Lebensdauer auf:

```
Klassen.vec <- c("[0,2]", "(2,4]", "(4,6]", "(6,8]", "(8,10]") # Einteilung in gleiche Klassenbreite
Anzahl.vec <- c(33,276,404,237,50)
```

```
Klassen.vec; Anzahl.vec
```

```
## [1] "[0,2]" "(2,4]" "(4,6]" "(6,8]" "(8,10]"
```

```
## [1] 33 276 404 237 50
```

```
# Darstellung der Daten in einer Tabelle
```

```
Lebensdauer.df <- data.frame(Lebensdauer=Klassen.vec, Anzahl_Motoren=Anzahl.vec)
Lebensdauer.df
```

```
##   Lebensdauer Anzahl_Motoren
## 1      [0,2]           33
## 2      (2,4]          276
## 3      (4,6]          404
## 4      (6,8]          237
## 5      (8,10]          50
```

Ergänzung der vorgegebenen Liste um die Klassenbreite und die Klassenmitte

```
Lebensdauer.df <- data.frame(Lebensdauer=Lebensdauer.df$Lebensdauer, Klassenmitte=c(1,3,5,7,9), Klassenbreite=c(2,2,2,2,2))
Lebensdauer.df
```

```
##   Lebensdauer Klassenmitte Klassenbreite Anzahl_Motoren
## 1      [0,2]           1           2           33
## 2      (2,4]           3           2          276
## 3      (4,6]           5           2          404
## 4      (6,8]           7           2          237
## 5      (8,10]          9           2           50
```

Ergänzung der Tabelle um die relative Häufigkeitsverteilung und die Summenhäufigkeit

```
RelHfgk.vec <- Lebensdauer.df$Anzahl_Motoren / sum(Lebensdauer.df$Anzahl_Motoren)
SumHfgk.vec <- cumsum(RelHfgk.vec)
Lebensdauer.df <- data.frame(Lebensdauer.df, RelHfgk.vec, SumHfgk.vec)
Lebensdauer.df
```

```
##   Lebensdauer Klassenmitte Klassenbreite Anzahl_Motoren RelHfgk.vec SumHfgk.vec
## 1      [0,2]           1           2           33      0.033      0.033
## 2      (2,4]           3           2          276      0.276      0.309
## 3      (4,6]           5           2          404      0.404      0.713
## 4      (6,8]           7           2          237      0.237      0.950
## 5      (8,10]          9           2           50      0.050      1.000
```

Ergänzung der Tabelle um die Häufigkeitsdichte

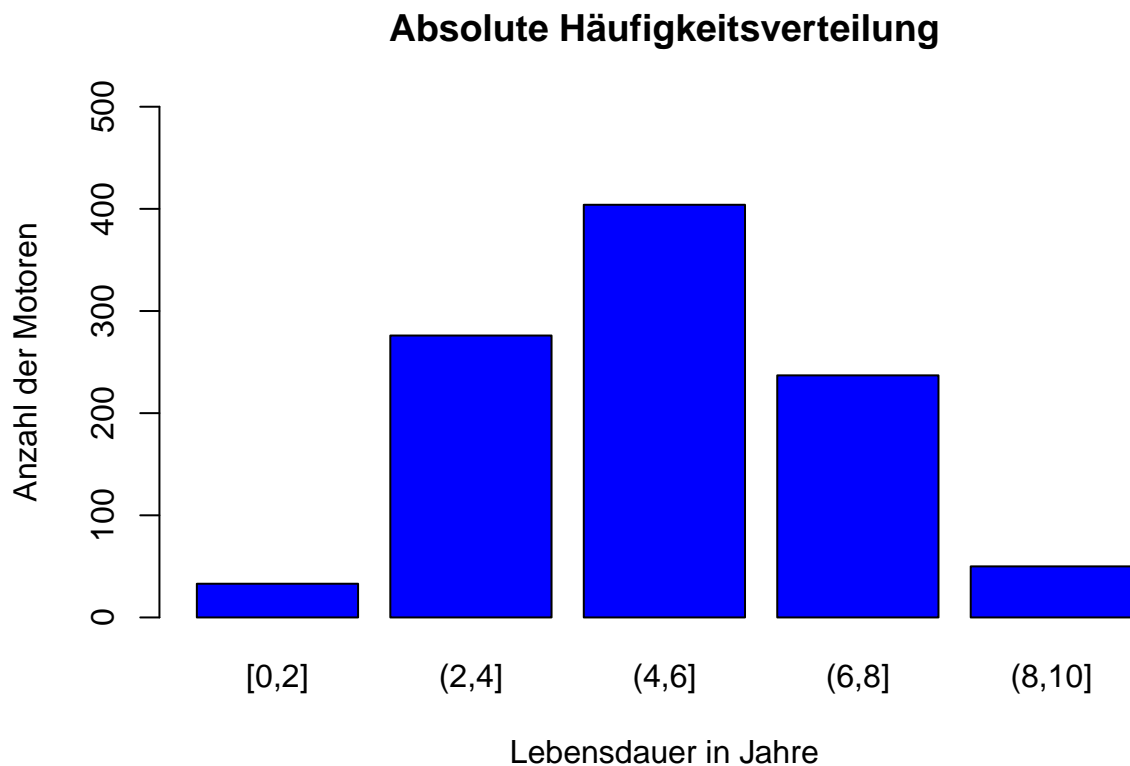

```
HfgkDichte.vec <- Lebensdauer.df$RelHfgk.vec / Lebensdauer.df$Klassenbreite
Lebensdauer.df <- data.frame(Lebensdauer.df, Dichte=HfgkDichte.vec)
Lebensdauer.df
```

```
##   Lebensdauer Klassenmitte Klassenbreite Anzahl_Motoren RelHfgk.vec SumHfgk.vec
## 1      [0,2]           1           2           33      0.033      0.033
## 2      (2,4]           3           2          276      0.276      0.309
## 3      (4,6]           5           2          404      0.404      0.713
## 4      (6,8]           7           2          237      0.237      0.950
## 5      (8,10]          9           2           50      0.050      1.000
##   Dichte
## 1 0.0165
## 2 0.1380
## 3 0.2020
## 4 0.1185
## 5 0.0250
```

a) Stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.

Absolute Häufigkeitsverteilung

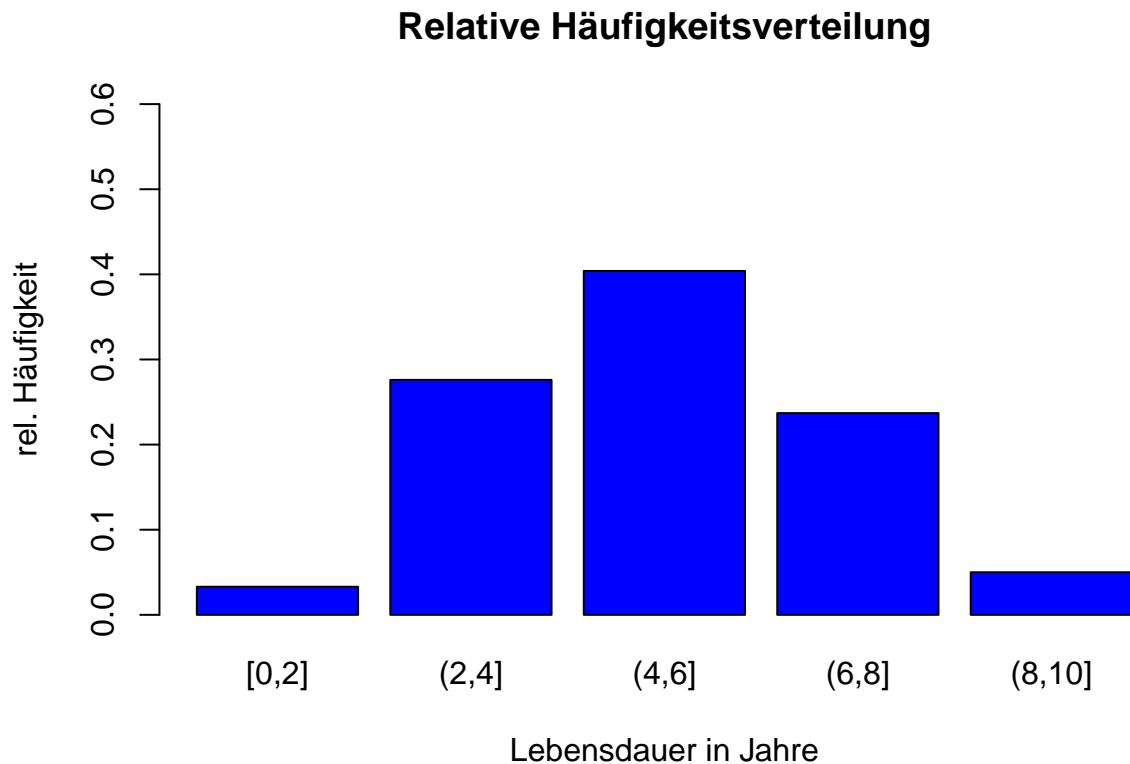
```
# Erstellung einer absoluten Häufigkeitsverteilung mit gleicher Klassenbreite
barplot(Lebensdauer.df$Anzahl_Motoren, names=Lebensdauer.df$Lebensdauer, col = "blue", xlab = "Lebensdauer in Jahre")
```



```
# Für eine bessere Lesbarkeit der Darstellung wurden die Limits der y-Achse nicht automatisch sondern m
```

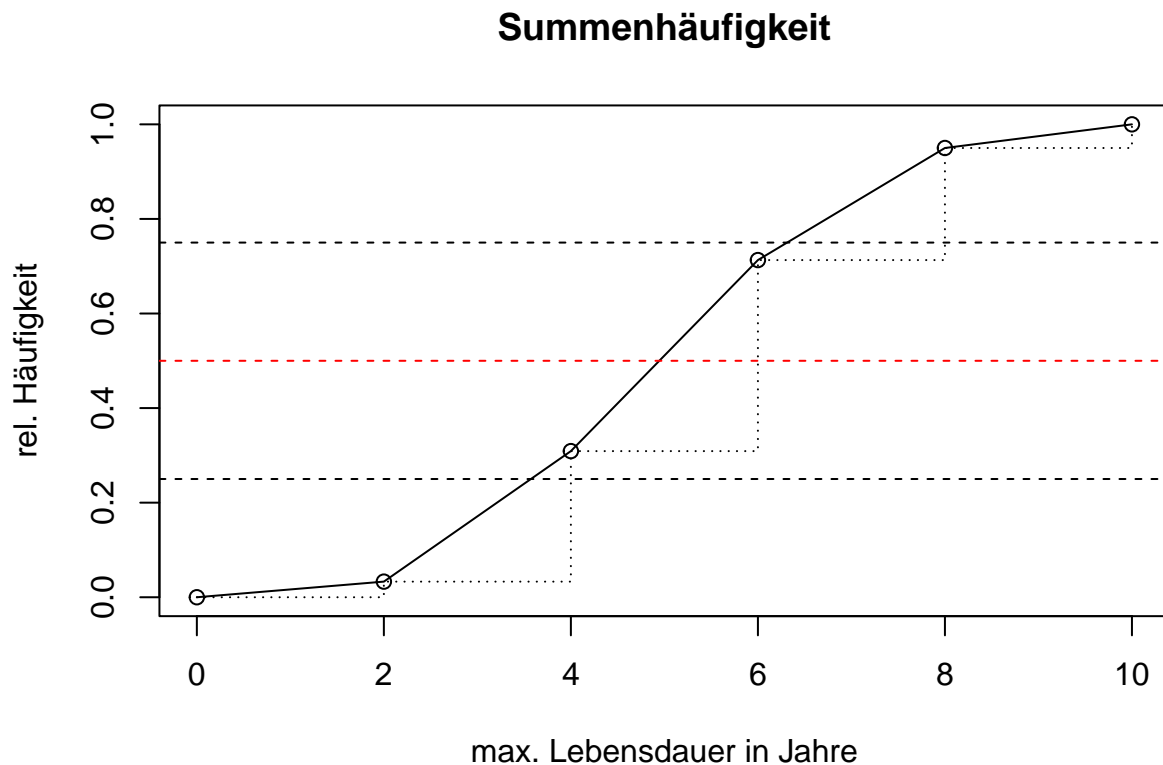
Relative Häufigkeitsverteilung

```
barplot(Lebensdauer.df$Anzahl_Motoren / sum(Lebensdauer.df$Anzahl_Motoren), names=Lebensdauer.df$Lebens
```



Summenhäufigkeit

```
X.vec <- c(0,2,4,6,8,10)
Y.vec <- c(0, Lebensdauer.df$SumHfgk.vec)
plot(X.vec, Y.vec, type = "l", lty = 1, main = "Summenhäufigkeit", xlab = "max. Lebensdauer in Jahre", ylab = "Summenhäufigkeit")
points(X.vec, Y.vec) # Darstellung der Messpunkte als Punkte
lines(X.vec, Y.vec, type = "s", lty = 3) # Darstellung der Steigungsdreiecke
abline(0.25, 0, lty = "dashed") # unteres Quartil
abline(0.5, 0, lty = "dashed", col = "red") # Median
abline(0.75, 0, lty = "dashed") # oberes Quartil
```



b) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

Arithmetisches Mittel - geschätzt:

Das arithmetische Mittel kann als Schwerpunkt der Häufigkeitsverteilung angesehen werden. Da dieses Beispiel konstante Klassenbreiten hat, kann man das arithmetische Mittel so ablesen, dass links und rechts vom arithmetischen Mittel auf der x-Achse in etwa gleich viele Werte vorhanden sind. Das geschätzte arithmetische Mittel beträgt in dem Fall 5.

Median:

Der Median liegt bei 50% auf der y-Achse und der Wert kann auf der x-Achse abgelesen werden. Der Median beträgt schätzungsweise bei 5.

c) Bestimme der Anteil der Motoren mit über 6 Jahren Lebensdauer.

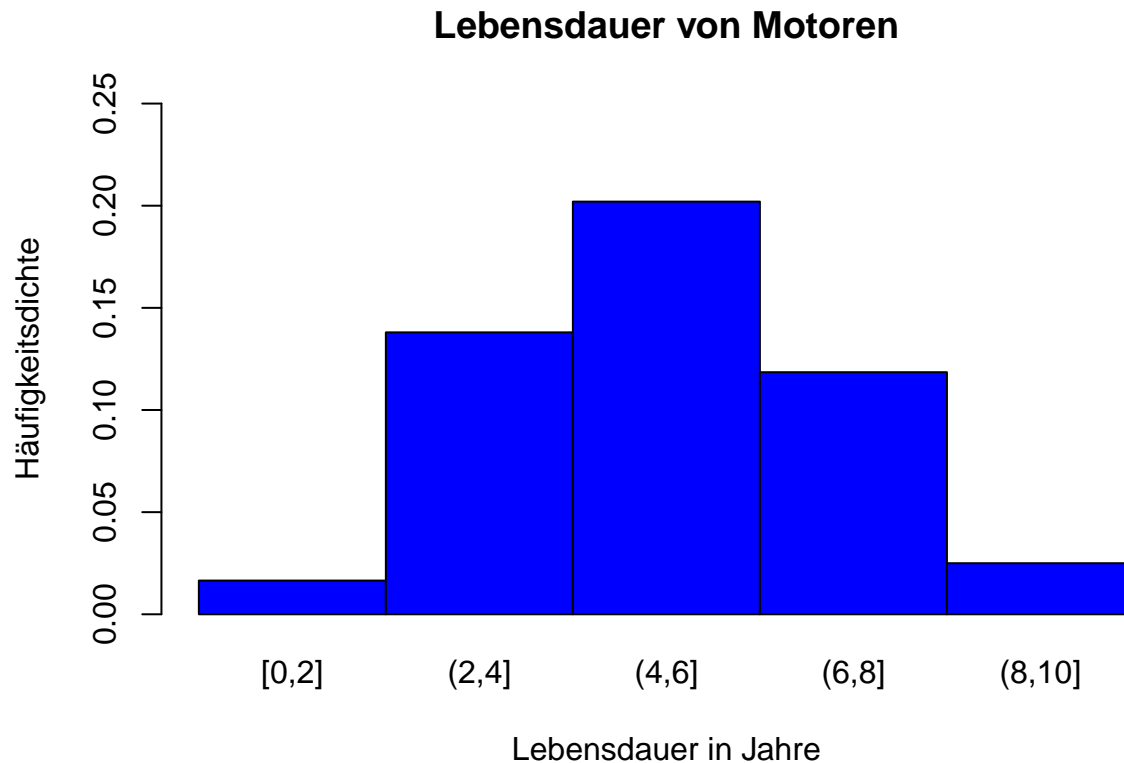
```
LangeLebensdauer.vec <- Lebensdauer.df$Anzahl_Motoren[4] + Lebensdauer.df$Anzahl_Motoren[5]
LangeLebensdauer.vec
```

```
## [1] 287
```

Motoren mit einer Lebensdauer über 6 Jahren befinden sich in der Liste in den letzten beiden statistischen Einheiten. Über die `[]` kann auf spezielle Vektorelemente/Listenelemente zugegriffen werden. In diesem Fall wird auf die Listenelemente 4 und 5 zugegriffen und für die Anzahl der Motoren, die länger als 6 Jahre laufen, addiert. Somit laufen 287 Motoren länger als 6 Jahre.

Lebensdauer von Motoren

```
# Ein Histogramm kann über eine Barplot gezeichnet werden, wenn die Abstände zwischen den Klassen auf 0  
barplot(Lebensdauer.df$Dichte, names=Lebensdauer.df$Lebensdauer, main = "Lebensdauer von Motoren", xlab
```



d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße).

```
Lebensdauer.vec <- c(rep(1,33), rep(3,276), rep(5,404), rep(7,237), rep(9,50))
```

Zentralmaße:

Modus

```
getmode <- function(Lebensdauer.vec) {  
  uniqv <- unique(Lebensdauer.vec)  
  uniqv[which.max(tabulate(match(Lebensdauer.vec, uniqv)))]  
}  
Modus <- getmode(Lebensdauer.vec)  
Modus
```

```
## [1] 5
```

Der Modus wird hier durch die Zahl 5 repräsentiert. Da diese Zahl innerhalb eines Intervalls liegt, beträgt der Modus (4,6].

Median

```
median(Lebensdauer.vec)
```

```
## [1] 5
```

Mittelwert

```
Mittelwert <- sum(Lebensdauer.df$RelHfgk.vec * Lebensdauer.df$Klassenmitte)
Mittelwert
```

```
## [1] 4.99
```

Streuumaße:

Minimum

```
min(Lebensdauer.vec)
```

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

Maximum

```
max(Lebensdauer.vec)
```

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

Spannweite

```
range(Lebensdauer.vec)
```

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

Quantile

```
quantile(Lebensdauer.vec)
```

```
##    0%   25%   50%   75%  100%
##     1     3     5     7     9
```

Standardabweichung

```
sd(Lebensdauer.vec)
```

```
## [1] 1.83937
```

Mittlere absolute Abweichung

```
mean(abs(Lebensdauer.vec-mean(Lebensdauer.vec)))
```

```
## [1] 1.36182
```

In diesem Beispiel ist der Modus 5. Da es 5 Jahre Lebensdauer nicht gibt, ist die richtige Antwort, dass der Modus das einseitig offene Intervall (4,6] ist.

A-4) Ein technisches Servicecenter zeichnet an 100 Tagen die Häufigkeit der Einsätze auf. Es ergibt sich folgende Tabelle:

```
AnzahlTage.vec <- c(16,48,27,9)
AnzahlEinsätze.vec <- c("[0,10]", "(10,20]", "(20,30]", "(30,80]")
Service.df <- data.frame(Einsätze = AnzahlEinsätze.vec, Tage = AnzahlTage.vec)
RelHfk.vec <- Service.df$Tage / sum(Service.df$Tage)
SumHfk.vec <- cumsum(RelHfk.vec)
# Die Tabelle wird kontinuierlich mit Werten aufgefüllt.
# In der data.frame Funktion wird immer wieder Service.df mit aufgenommen. Dies dient dazu, die bisheri.
Service.df <- data.frame(Service.df, relHfk = RelHfk.vec, sumHfk = SumHfk.vec)
Service.df <- data.frame(Service.df, Klassenbreite=c(10,10,10,50), Klassenmitte=c(5,15,25,55))

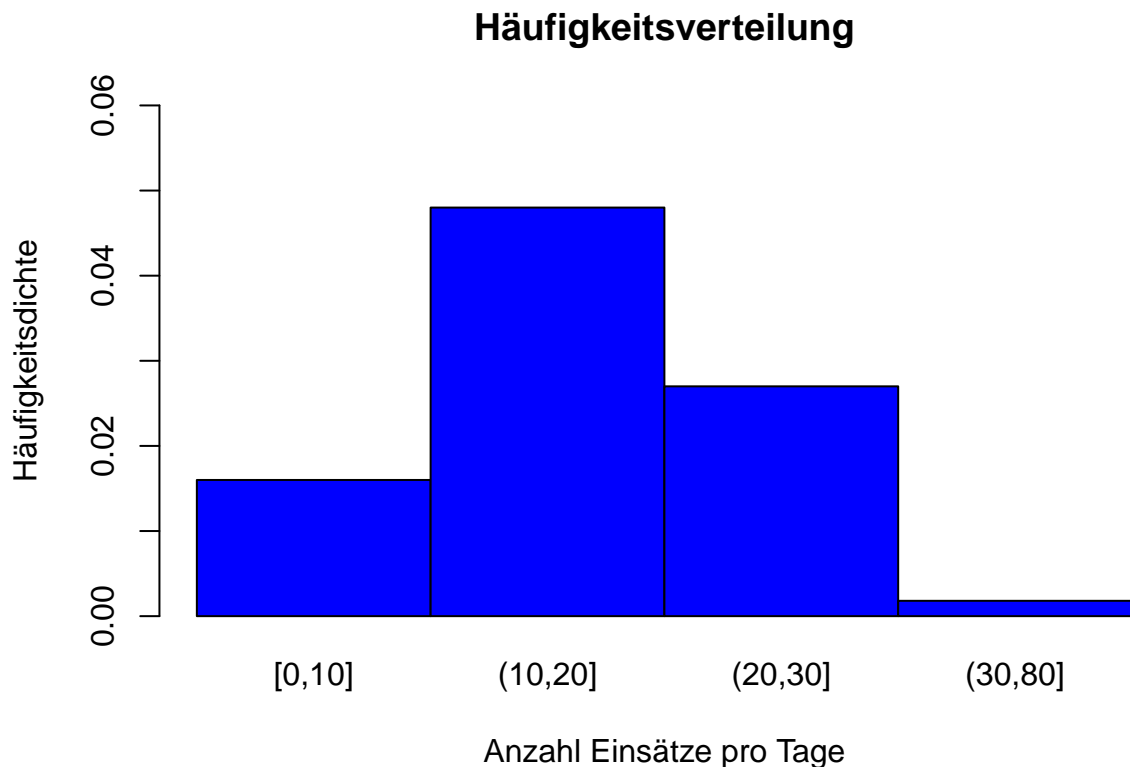
HfkDichte.vec <- Service.df$relHfk / Service.df$Klassenbreite
Service.df <- data.frame(Service.df, Dichte=HfkDichte.vec)
Service.df
```

##	Einsätze	Tage	relHfk	sumHfk	Klassenbreite	Klassenmitte	Dichte
## 1	[0,10]	16	0.16	0.16	10	5	0.0160
## 2	(10,20]	48	0.48	0.64	10	15	0.0480
## 3	(20,30]	27	0.27	0.91	10	25	0.0270
## 4	(30,80]	9	0.09	1.00	50	55	0.0018

a) Stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.

Häufigkeitsverteilung

```
barplot(Service.df$Dichte, names=Service.df$Einsätze, col = "blue", main = "Häufigkeitsverteilung", xlab = "Einsätze", ylab = "Dichte")
```

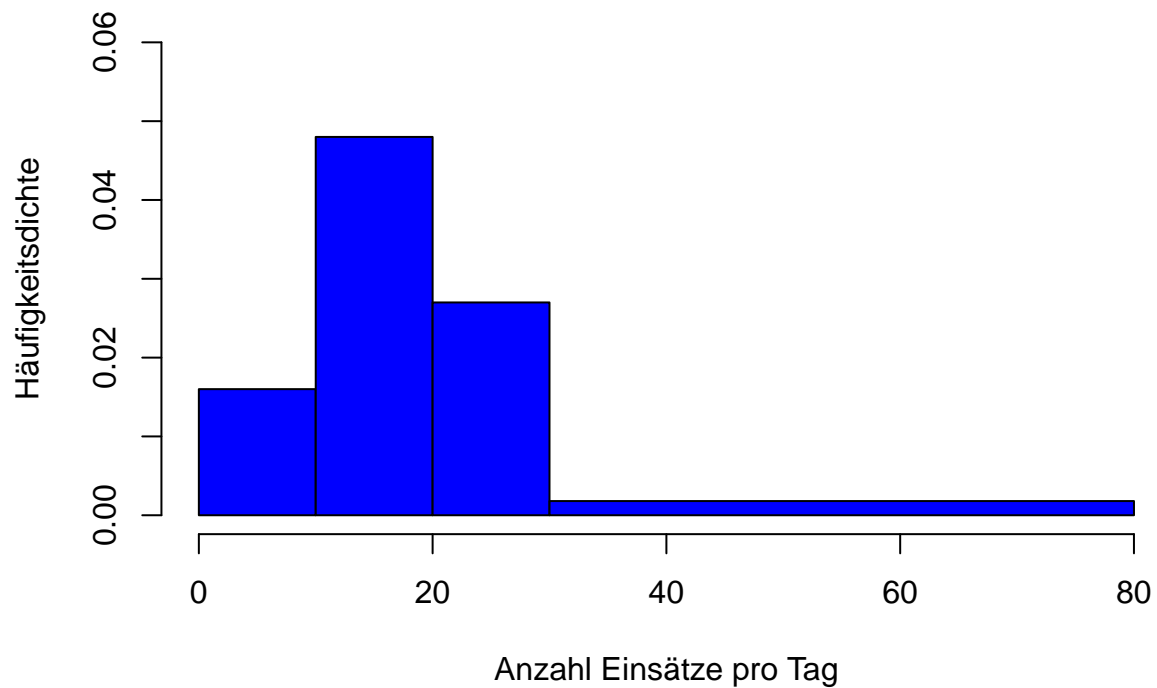


Die oben dargestellte Häufigkeitsverteilung bildet die statistischen Einheiten in einheitlichen Klassenbreiten ab. Die Darstellung lässt den Trugschluss zu, dass die letzte statistische Einheit die gleiche Breite wie die anderen Klassen aufweist. Im Gegensatz zu den anderen Einheiten ist die letzte statistische Einheit nicht 10 sondern 50 Einsätze pro Tag breit. In diesem Fall entspricht die Darstellung nicht einer Häufigkeitsdichteverteilung, was auch bedeutet, dass die y-Achse nicht richtig beschriftet ist.

In der nachfolgenden Darstellung werden die Intervalle ignoriert und die absolute Anzahl der Einsätze pro Tag angezeigt. Diese Darstellung kann einfacher verständlich und besser anschaulich zeigen, mit welcher Häufigkeit Anrufe pro Tag auftreten. Die Darstellung entspricht auch der Häufigkeitsdichtefunktion.

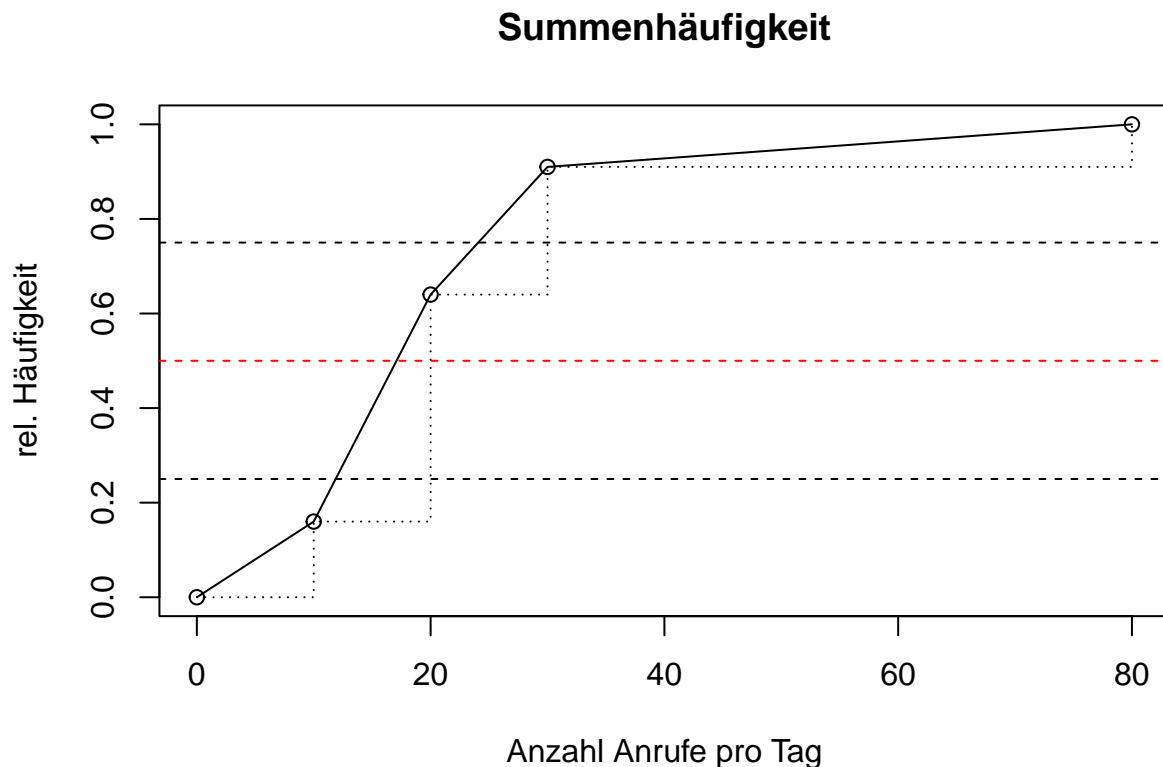
```
Service.vec <- c(rep(5,16), rep(15,48), rep(25,27), rep(55,9))
X2a4.vec <- c(0,10,20,30,80)
hist(Service.vec, X2a4.vec, main = "Häufigkeitsverteilung (andere Darstellung)", xlab = "Anzahl Einsätze pro Tag")
```

Häufigkeitsverteilung (andere Darstellung)



Summenhäufigkeit

```
Xa4.vec <- c(0,10,20,30,80)
Ya4.vec <- c(0, Service.df$sumHfk)
plot(Xa4.vec, Ya4.vec, main = "Summenhäufigkeit", xlab = "Anzahl Anrufe pro Tag", ylab = "rel. Häufigkeit",
     points(Xa4.vec, Ya4.vec)
lines(Xa4.vec, Ya4.vec, type = "s", lty = 3) # aktivieren der Darstellung der Dreiecke für die Klassen
abline(0.25, 0, lty = "dashed")             # Linie für das untere Quartil
abline(0.75, 0, lty = "dashed")             # Linie für das obere Quartil
abline(0.5, 0, lty = "dashed", col = "red")  # Linie für den Median
```

b) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

Arithmetisches Mittel: Das arithmetische Mittel wird auf 19 geschätzt. Das arithmetische Mittel beschreibt den Schwerpunkt der Verteilung. Da die Klassenbreiten bei dieser Verteilung nicht konstant sind, kann das arithmetische Mittel nicht exakt durch die Methode der “gleich viele Werte links und rechts des arithmetischen Mittels auf der x-Achse” angewandt werden. Die Häufigkeitsverteilung (andere Darstellung) liefert allerdings die Möglichkeit, das arithmetische Mittel anhand der Fläche abzuschätzen. Das arithmetische Mittel liegt dort, wo die Flächen links und rechts davon gleich groß sind. In diesem Fall kann das arithmetische Mittel auf 20 geschätzt.

Median: Der Median kann an der y-Achse ablesen werden. Genauer gesagt an der Stelle, an der die 50% Marke bei der relativen Häufigkeit im Summenhäufigkeitsdiagramm liegt. In diesem Fall wird der Median auf 16 geschätzt.

c) Bestimme der Anteil der Tage mit über 20 Einsätzen.

```
Stress.vec <- Service.df$relHfk[3] + Service.df$relHfk[4]
```

Die Fragestellung verlangt nach der expliziten Angabe der Einsätze über 20 Einsätze Pro Tage. Durch Addition der relativen Häufigkeit der dritten und vierten statistischen Einheiten kann die Häufigkeit bestimmt werden, mit der mehr als 20 Einsätze pro Tag absolviert werden müssen. Bei der dritten statistischen Einheit wird der Intervall mit der unteren Grenze von 20 angegeben. Eine genaue Angabe, wie häufig Tage mit über

20 Einsätzen sind, kann daher an dieser Stelle nicht gemacht werden, da ebendieses Intervall auch 20 Einsätze enthält, in der Fragestellung aber explizit nach ÜBER 20 Einsätzen gefragt wird. Ohne Rücksichtnahme auf diese Unstimmigkeit beträgt die Häufigkeit von mehr als 20 Einsätzen pro Tag rund 36 %.

d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße). Welcher Aspekt könnte hier problematisch sein? Warum?

Zentralmaße:

Modus

```
getmode <- function(Service.vec) {  
  uniqv <- unique(Service.vec)  
  uniqv[which.max(tabulate(match(Service.vec, uniqv)))]  
}  
getmode(Service.vec)
```

```
## [1] 15
```

Median

```
median(Service.vec)
```

```
## [1] 15
```

Mittelwert

```
Mittelwert <- sum(Service.df$RelHfgk.vec * Service.df$Klassenmitte)  
Mittelwert
```

```
## [1] 0
```

Streumaße:

Minimum

```
min(Service.vec)
```

```
## [1] 5
```

Maximum

```
max(Service.vec)
```

```
## [1] 55
```

Spannweite

```
range(Service.vec)
```

```
## [1] 5 55
```

Quantile

```
quantile(Service.vec)
```

```
## 0% 25% 50% 75% 100%  
## 5 15 15 25 55
```

Standardabweichung

```
sd(Service.vec)
```

```
## [1] 12.90642
```

Mittlere absolute Abweichung

```
mean(abs(Service.vec-mean(Service.vec)))
```

```
## [1] 9.216
```

Die Aussagekraft des Medians für die Mitte der Daten ist für diese Datenreihe zielführender, da das arithmetische Mittel auch Ausreißer mit berücksichtigt, was beim Median nicht der Fall ist.

B) VERSTÄNDNISFRAGEN

B-1) Zeige, dass das arithmetische Mittel unter dem Schwerpunkt der Häufigkeitsfunktion liegt.

B-1)

$M_1 = M_2$ $M \dots$ Sind die Klassen gleich groß, ist der Schwerpunkt dort, wo sich die Momente aufheben

$$\int_{x_0}^{x_1} (x_1 - x) \cdot f(x) dx = \int_{x_1}^{x_2} (x - x_1) \cdot f(x) dx$$

$$\int_{x_0}^{x_1} (x_1 - x) \cdot f(x) dx - \int_{x_1}^{x_2} (x - x_1) \cdot f(x) dx = \int_{x_0}^{x_2} (x_1 - x) \cdot f(x) dx = 0$$

$$\int_{x_0}^{x_1} x_1 \cdot f(x) dx - \int_{x_0}^{x_1} x \cdot f(x) dx = \int_{x_1}^{x_2} x_1 \cdot f(x) dx + \int_{x_1}^{x_2} x_2 \cdot f(x) dx = \int_{x_0}^{x_2} (x_1 - x) \cdot f(x) dx = 0$$

$$\int_{x_0}^{x_1} x_1 \cdot f(x) dx - \int_{x_0}^{x_2} x \cdot f(x) dx = \int_{x_0}^{x_2} (x_1 - x) \cdot f(x) dx = 0$$

$$\int_{x_0}^{x_1} x_1 \cdot f(x) dx - \int_{x_0}^{x_2} x \cdot f(x) dx = \int_{x_0}^{x_2} x_1 \cdot f(x) dx - \int_{x_0}^{x_2} x \cdot f(x) dx = 0$$

$$\int_{x_0}^{x_1} x_1 \cdot f(x) dx - \int_{x_0}^{x_2} x \cdot f(x) dx = 0$$

$$x_1 \int_{x_0}^{x_1} f(x) dx = \int_{x_0}^{x_2} x \cdot f(x) dx$$

$$x_1 = f(x) \cdot \int_{x_0}^{x_1} x dx$$

$$x_1 = f(x_{i1}) \cdot \int_{x_0}^{x_{i1}} x dx + f(x_{i2}) \cdot \int_{x_{i1}}^{x_{i2}} x dx + \dots + f(x_{in-1}) \cdot \int_{x_{in-1}}^{x_{in}} x dx$$

$$x_1 = f(x_{i1}) \cdot \left[\frac{x^2}{2} \right]_{x_0}^{x_{i1}} + f(x_{i2}) \cdot \left[\frac{x^2}{2} \right]_{x_{i1}}^{x_{i2}} + \dots + f(x_{in-1}) \cdot \left[\frac{x^2}{2} \right]_{x_{in-1}}^{x_{in}}$$

$$x_1 = f(x_{i1}) \cdot \frac{x_{i1}^2 - x_0^2}{2} + f(x_{i2}) \cdot \frac{x_{i2}^2 - x_{i1}^2}{2} + \dots + f(x_{in-1}) \cdot \frac{x_{in}^2 - x_{in-1}^2}{2}$$

$$x_1 = \sum_{i=1}^N f(x_i) \cdot \frac{x_i^2 - x_{i-1}^2}{2}$$

$x_i, x_{i-1} \dots$ Klassenmitte mit konstantem $f(x)$ innerhalb der Klasse

$$x_1 = \sum_{i=1}^N f(x_i) \cdot \frac{(x_i - x_{i-1})(x_i + x_{i-1})}{2}$$

$$x_1 = \sum_{i=1}^N f(x_i) \cdot (x_i - x_{i-1}) \cdot \frac{x_i + x_{i-1}}{2}$$

$\Delta x_i \dots$ Klassenbreite

$$x_1 = \sum_{i=1}^N f(x_i) \cdot \Delta x_i \cdot \underbrace{\left(x_{i-1} + \frac{x_i - x_{i-1}}{2} \right)}_{x_m}$$

$x_m \dots$ Klassenmitte

$$x_1 = \sum_{i=1}^N f(x_i) \cdot \Delta x_i \cdot x_{m,i}$$

$\Delta x_i \dots$ einzeln betrachtet, ergibt jede Merkmalsausprägung $\Delta x_i = 1 \Rightarrow x_i - x_{i-1} = 5 - 4 = 1$

$$x_1 = \sum_{i=1}^N f(x_i) \cdot \Delta x_i \cdot x_i$$

$$x_1 = \sum_{i=1}^N f(x_i) \cdot x_i \equiv \mu$$

B-2) Verschiebungssatz zur Berechnung der Standardabweichung

B-2)

$$\mu = \sum_{i=1}^n f(a_i) \cdot a_i$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot (a_i - \mu)^2$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot (a_i^2 - 2a_i\mu + \mu^2)$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 + \sum_{i=1}^n f(a_i) \cdot (-2a_i\mu) + \sum_{i=1}^n f(a_i) \cdot \mu^2$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - 2\mu \cdot \underbrace{\sum_{i=1}^n f(a_i) \cdot a_i}_{\mu} + \mu^2 \cdot \underbrace{\sum_{i=1}^n f(a_i)}_1$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - 2\mu \cdot \mu + \mu^2 \cdot 1$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - 2\mu^2 + \mu^2$$

$$\underline{\underline{G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - \mu^2}}$$

B-3) Das komma-separierte File "Fehlerquote.csv" enthält das Prüfergebnis von 50 Bauteilen auf Funktionstüchtigkeit. Dabei steht der Eintrag "0" für ein fehlerfreies Bauteil und "1" für ein fehlerhaftes Bauteil.

```
# Einlesen einer .csv Datei über den read.csv Befehl
Fehlerquote.df <- read.csv("Fehlerquote.csv", sep = ";", dec = ",", header = TRUE)
```

a) Welche Skalierung hat dieses Merkmal?

Die Skalierung dieses Merkmals entspricht einem normalskalierten dichotomen Merkmal.

b) Stelle die Messergebnisse in einer Häufigkeitstabelle und grafisch dar.

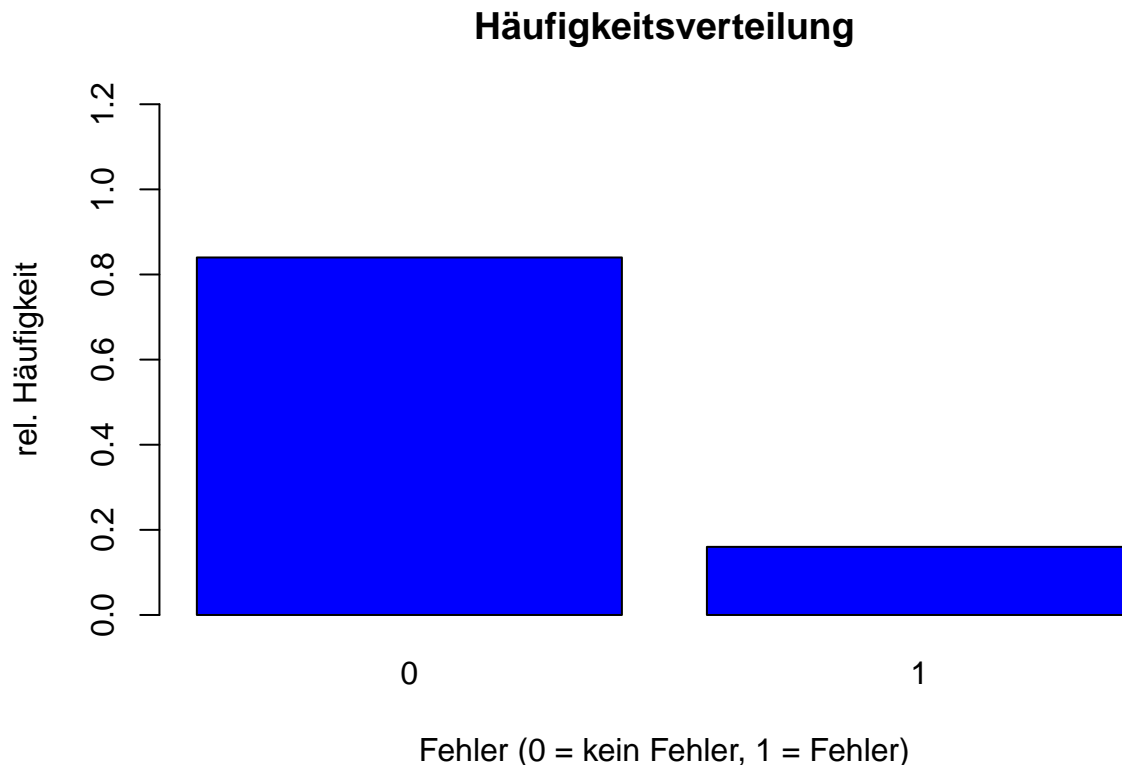
Häufigkeitstabelle:

```
Fehlerquote.vec <- table(Fehlerquote.df$fehlerhaft)/length(Fehlerquote.df$fehlerhaft)
# Spalte 1 - Zustand 0/1:
Zustand.vec <- c("fehlerfrei - 0", "fehlerhaft - 1")
# Spalte 2 - rel. Häufigkeit
Fehlerhaft.vec <- sum(Fehlerquote.df$fehlerhaft)/length(Fehlerquote.df$Nr.)
RelHfkFehlerquote.vec <- c((1-Fehlerhaft.vec), Fehlerhaft.vec)

HfgTabelle.df <- data.frame(Zustand = Zustand.vec, rel.Häufigkeit = RelHfkFehlerquote.vec)
HfgTabelle.df
```

```
##          Zustand rel.Häufigkeit
## 1 fehlerfrei - 0          0.84
## 2 fehlerhaft - 1          0.16
```

```
barplot(Fehlerquote.vec, main = "Häufigkeitsverteilung", xlab = "Fehler (0 = kein Fehler, 1 = Fehler)",
```



c) Wie kann man in diesem Beispiel das arithmetische Mittel berechnen und wofür steht es in diesem Fall?

Das arithmetische Mittel gibt hier nur an, wie viele Teile fehlerhaft sind, da die Bauteile mit 0, also fehlerfreie Bauteile, nicht miteinberechnet werden.

d) Wie groß ist die Standardabweichung σ ? Leite eine Formel her und zeige, wie man sie in diesem Fall einfach aus dem arithmetischen Mittel errechnen kann.

Standardabweichung

```
sd(Fehlerquote.df$fehlerhaft)
```

```
## [1] 0.370328
```

B-3)d)

$$\mu = \sum_{i=1}^n f(a_i) \cdot a_i \quad G^2 = \sum_{i=1}^n f(a_i) \cdot (a_i - \mu)^2$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot (a_i - \mu)^2$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot (a_i - \mu)(a_i + \mu)$$

$$G^2 = \sum_{i=1}^n f(a_i) (a_i^2 - 2a_i\mu + \mu^2)$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - 2 \cdot f(a_i) \cdot a_i \cdot \mu + \underbrace{f(a_i) \cdot \mu^2}_{\mu^2}$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - 2 \cdot \mu \cdot \mu + \mu^2$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - 2\mu^2 + \mu^2$$

$$G^2 = \sum_{i=1}^n f(a_i) \cdot a_i^2 - \mu^2$$

Standardabweichung über arithmetisches Mittel

Anwendung der Formel, die in oben hergeleitet worden ist.

```
Varianz.vec <- sum((Fehlerquote.df$fehlerhaft - mean(Fehlerquote.df$fehlerhaft))^2/length(Fehlerquote.d
```

```
Standardabweichung.vec <- sqrt(Varianz.vec)
```

```
Standardabweichung.vec
```

```
## [1] 0.3666061
```

e) Wie müssen die Verteilungen in diesem Fall sein, damit die Streuung maximal bzw. minimal wird?

Die Streuung ist maximal, wenn gleich viele Teile fehlerhaft sind wie Teile fehlerfrei. Die Verteilung müsste also 50% fehlerhaft zu 50 % fehlerfrei sein. Dann ist die Streuung 0. Die Streuung ist minimal, wenn kein Teil fehlerhaft ist bzw. wenn alle Teile fehlerhaft sind.

B-4) Das komma-separierte File “Verteilungsvergleich.csv” enthält in 4 Spalten die Daten von folgenden Messreihen: Ergebnis von 40 Würfeln mit einem Würfel (Annahme: gleichverteilt), die Zeitspanne (in Minuten) zwischen 40 vorbeifahrenden Autos (Annahme: exponentialverteilt), die Länge von 40 Telefongesprächen in Minuten (Annahme: normalverteilt) und die Länge von 40 Holzstiften in cm (Annahme: normalverteilt). Vergleiche die vier verschiedenen Verteilungen in den folgenden Fragen:

```
Verteilungsvergleich.df <- read.csv("Verteilungsvergleich.csv", sep = ";", dec = ",", header = TRUE)
```

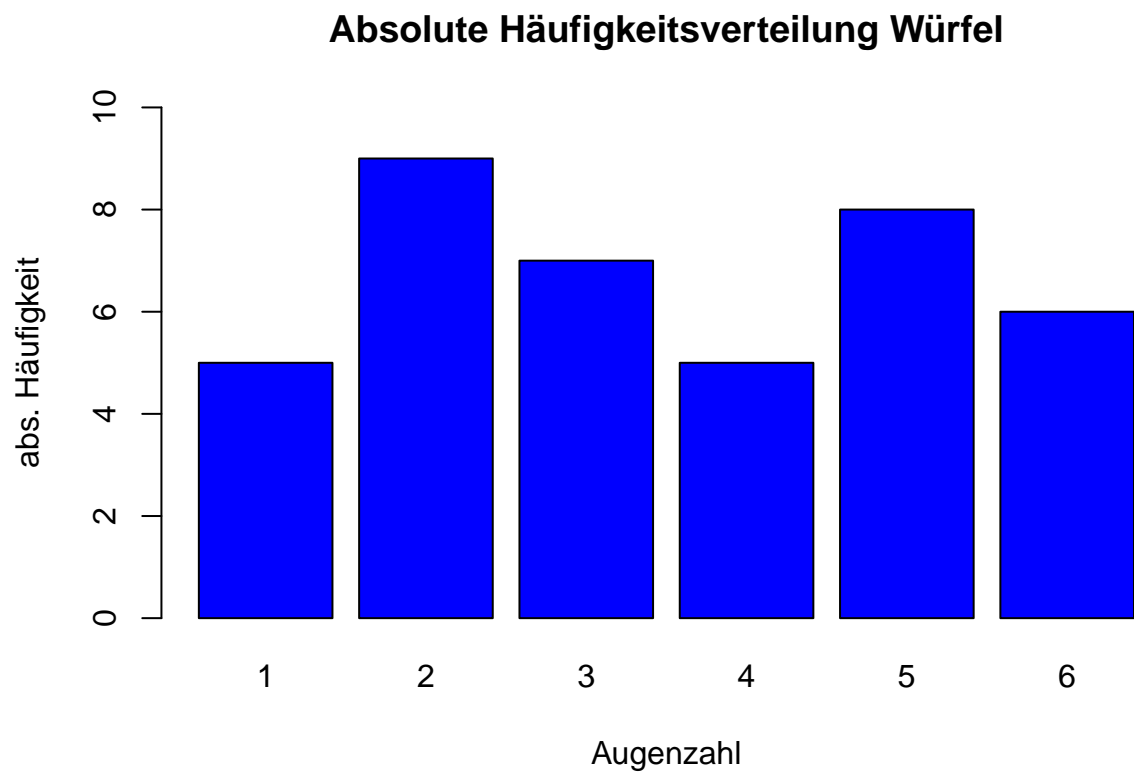
a) Erstelle für jede Messreihe eine Häufigkeitstabelle sowie eine grafische Darstellung der Häufigkeitsverteilung und der Summenhäufigkeit.

Würfel

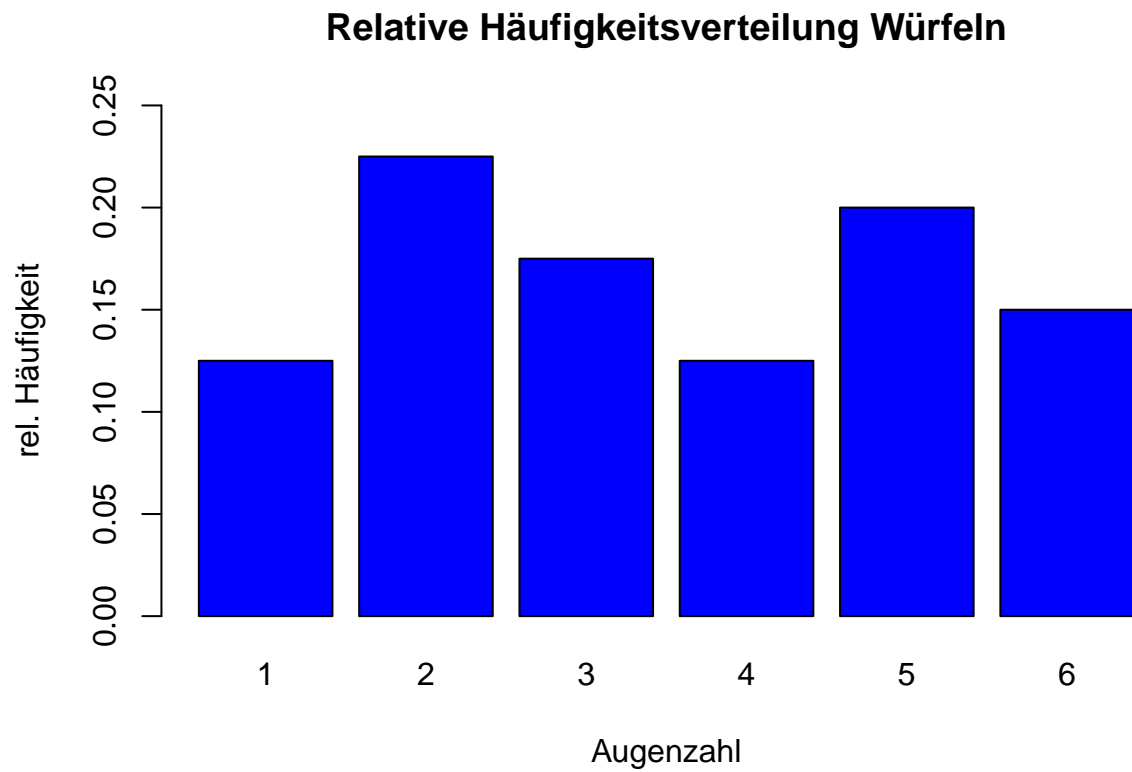
```
Würfel.df <- data.frame(Augenzahl = (1:6), abs.Häufigkeit = tabulate(Verteilungsvergleich.df$Wuerfel),  
Würfel.df <- data.frame(Würfel.df, Summenhfk. = cumsum(Würfel.df$rel.Häufigkeit))  
Würfel.df
```

##	Augenzahl	abs.Häufigkeit	rel.Häufigkeit	Summenhfk.
## 1	1	5	0.125	0.125
## 2	2	9	0.225	0.350
## 3	3	7	0.175	0.525
## 4	4	5	0.125	0.650
## 5	5	8	0.200	0.850
## 6	6	6	0.150	1.000

```
barplot(Würfel.df$abs.Häufigkeit, names = Würfel.df$Augenzahl, ylim = c(0,10), main = "Absolute Häufigk
```

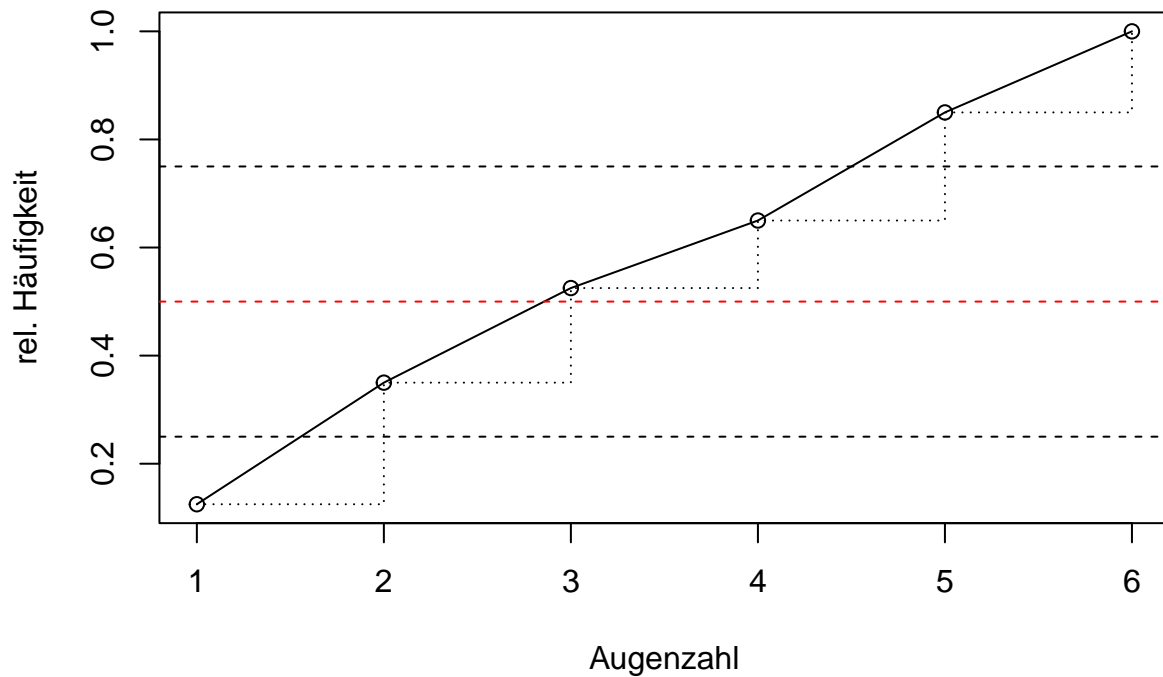



```
barplot(Würfel.df$rel.Häufigkeit, names = Würfel.df$Augenzahl, col = "blue", main = "Relative Häufigkei
```



```
plot(Würfel.df$Augenzahl, Würfel.df$Summenhfk., main = "Summenhäufigkeit Würfel", xlab = "Augenzahl", ylab = "rel. Häufigkeit",  
points(Würfel.df$Augenzahl, Würfel.df$Summenhfk.)  
lines(Würfel.df$Augenzahl, Würfel.df$Summenhfk., type = "s", lty = 3)  
abline(0.25, 0, lty = "dashed")  
abline(0.75, 0, lty = "dashed")  
abline(0.5, 0, lty = "dashed", col = "red")
```

Summenhäufigkeit Würfel



Wartezeit

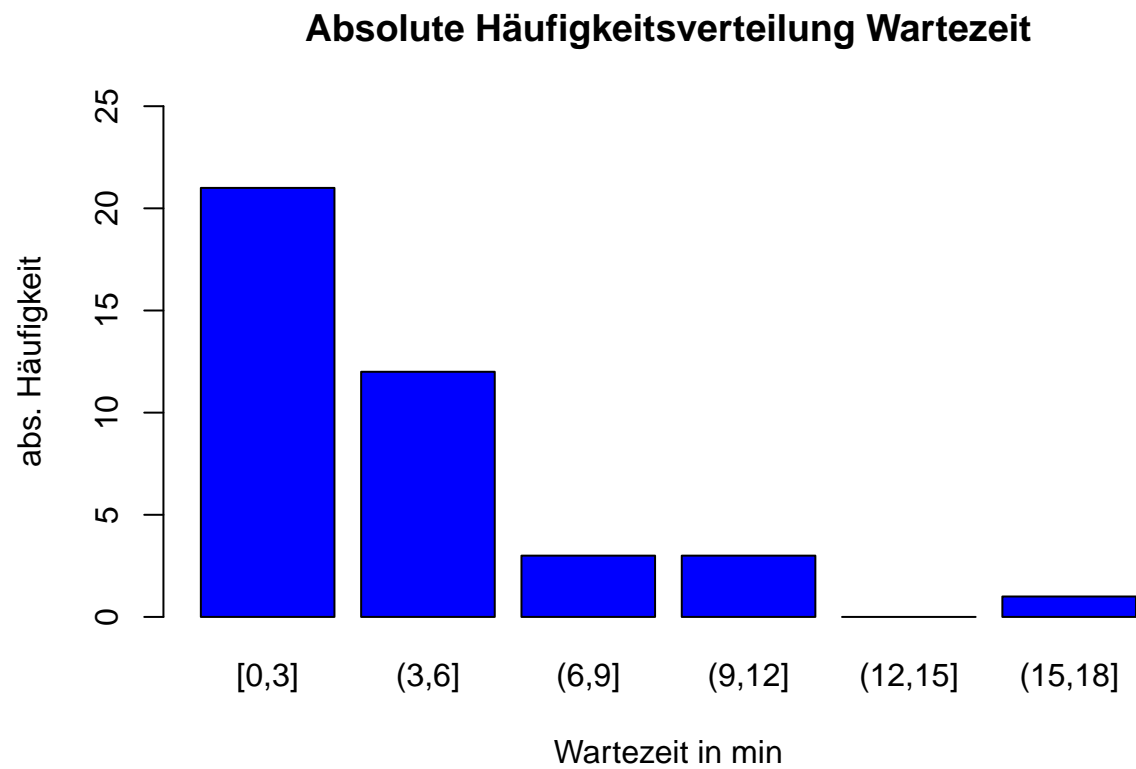
```
Zeit.vec <- c(0,3,6,9,12,15,18)
Zeitintervalle.vec <- c("[0,3]", "(3,6]", "(6,9]", "(9,12]", "(12,15]", "(15,18]")
Klasse.vec <- tabulate(cut(Verteilungsvergleich.df$Wartezeit_min, breaks = Zeit.vec))
Wartezeit.df <- data.frame(Zeit = Zeitintervalle.vec, Klassenbreite = rep(3,6), Klassenmitte = c(1.5,4.5,7.5,10.5,13.5,16.5))

Wartezeit.df <- data.frame(Wartezeit.df, Summenhfk. = cumsum(Wartezeit.df$rel.Häufigkeit), Dichte = Wartezeit.df$abs.Häufigkeit/3)
Wartezeit.df
```

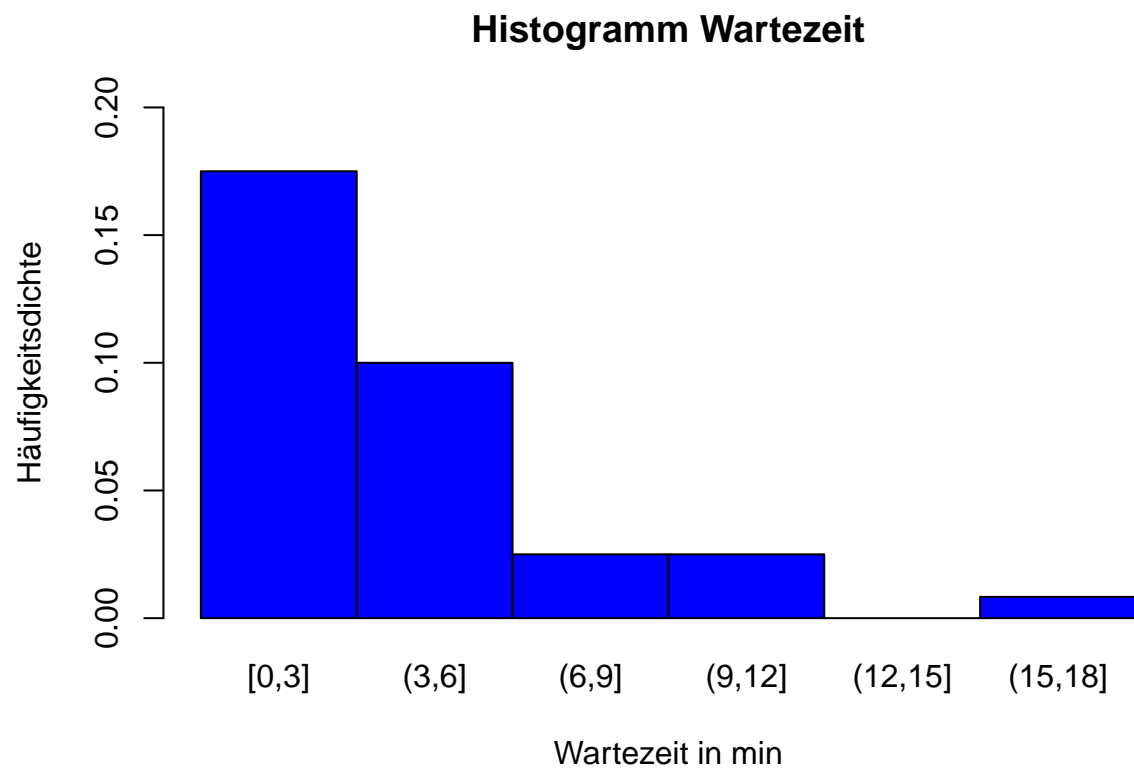
##	Zeit	Klassenbreite	Klassenmitte	abs.Häufigkeit	rel.Häufigkeit	Summenhfk.
## 1	[0,3]	3	1.5	21	0.525	0.525
## 2	(3,6]	3	4.5	12	0.300	0.825
## 3	(6,9]	3	7.5	3	0.075	0.900
## 4	(9,12]	3	10.5	3	0.075	0.975
## 5	(12,15]	3	13.5	0	0.000	0.975
## 6	(15,18]	3	16.5	1	0.025	1.000

##	Dichte
## 1	0.175000000
## 2	0.100000000
## 3	0.025000000
## 4	0.025000000
## 5	0.000000000
## 6	0.008333333

```
barplot(Wartezeit.df$abs.Häufigkeit, names = Wartezeit.df$Zeit, main = "Absolute Häufigkeitsverteilung Wartezeit")
```

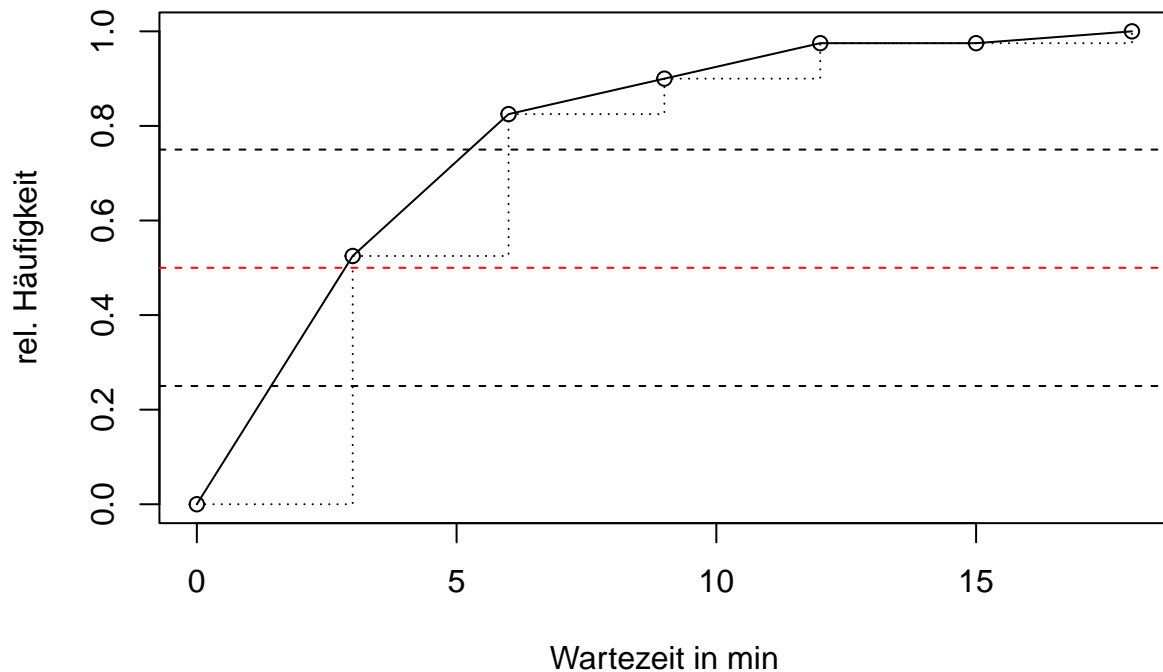


```
barplot(Wartezeit.df$Dichte, names = Wartezeit.df$Zeit, main = "Histogramm Wartezeit", xlab = "Wartezeit in min", ylab = "Dichte")
```



```
plot(c(0,3,6,9,12,15,18), c(0,Wartezeit.df$Summenhfk.), main = "Summenhäufigkeit Wartezeit", xlab = "Wartezeit in min", ylab = "Häufigkeitsdichte", col = "blue", lty = 1)
points(c(0,3,6,9,12,15,18), c(0,Wartezeit.df$Summenhfk.))
lines(c(0,3,6,9,12,15,18), c(0,Wartezeit.df$Summenhfk.), type = "s", lty = 3)
abline(0.25, 0, lty = "dashed")
abline(0.75, 0, lty = "dashed")
abline(0.5, 0, lty = "dashed", col = "red")
```

Summenhäufigkeit Wartezeit



Telegespräche

```
Dauer.vec <- c(0,2,4,6,8,10,12,14,16)
Dauer_Intervall.vec <- c("[0,2]", "(2,4]", "(4,6]", "(6,8]", "(8,10]", "(10,12]", "(12,14]", "(14,16]")
Dauer_Klasse.vec <- tabulate(cut(Verteilungsvergleich.df$Telgesprae_min, breaks = Dauer.vec))
Gesprächsdauer.df <- data.frame(Gesprächsdauer = Dauer_Intervall.vec, Klassenbreite = rep(2,8), Klassenmitte = rep(1,8))

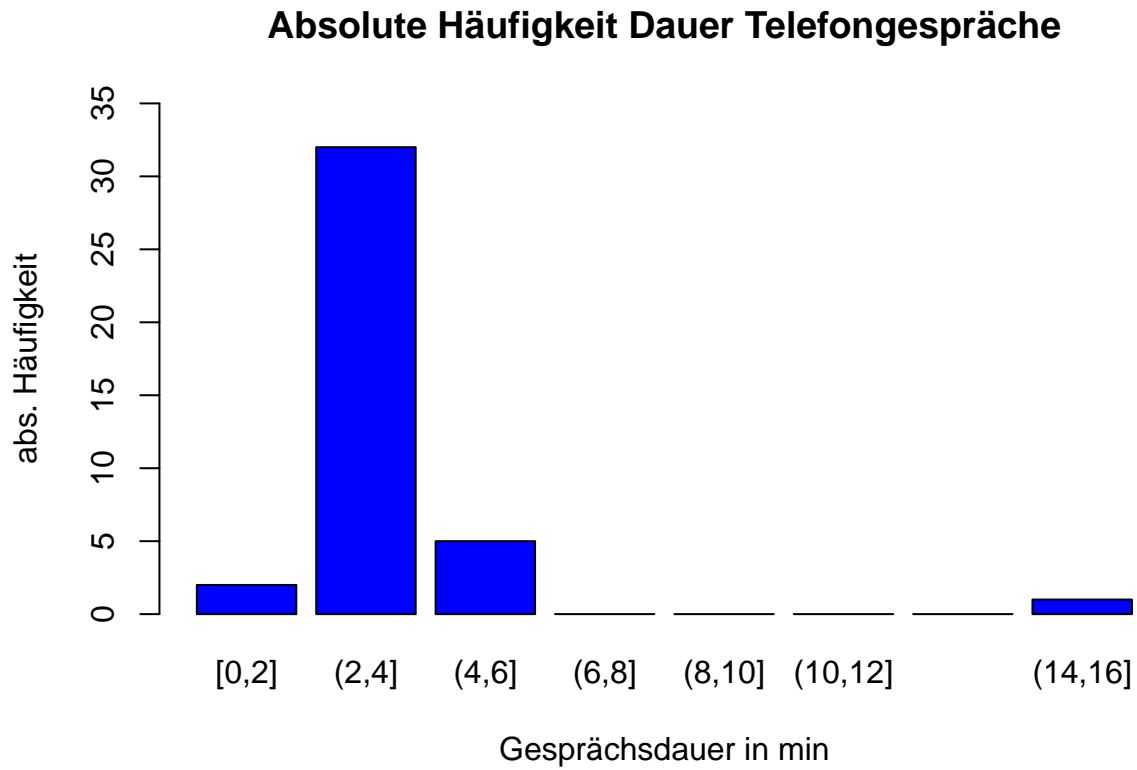
Gesprächsdauer.df <- data.frame(Gesprächsdauer.df, Summenhfk. = cumsum(Gesprächsdauer.df$rel.Häufigkeit))

Gesprächsdauer.df
```

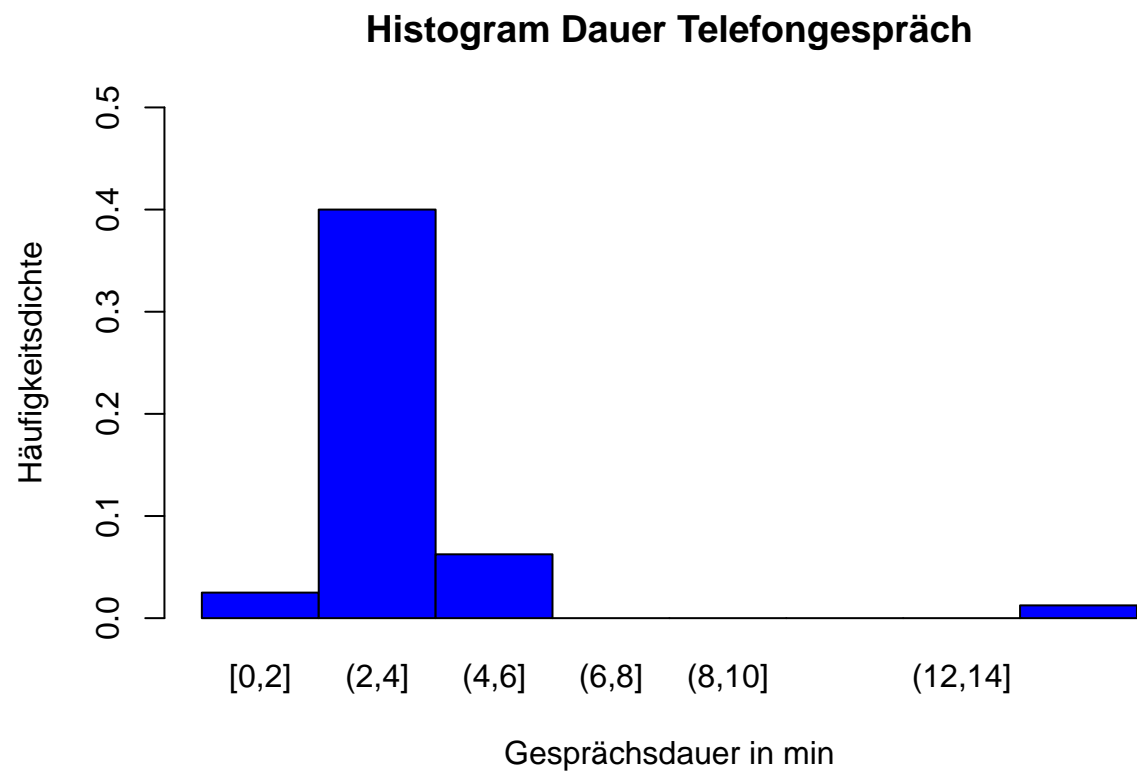
##	Gesprächsdauer	Klassenbreite	Klassenmitte	abs.Häufigkeit	rel.Häufigkeit
## 1	[0,2]	2	1	2	0.050
## 2	(2,4]	2	3	32	0.800
## 3	(4,6]	2	5	5	0.125
## 4	(6,8]	2	7	0	0.000
## 5	(8,10]	2	9	0	0.000
## 6	(10,12]	2	11	0	0.000
## 7	(12,14]	2	13	0	0.000
## 8	(14,16]	2	15	1	0.025
##	Summenhfk. Dichte				
## 1	0.050 0.0250				
## 2	0.850 0.4000				
## 3	0.975 0.0625				
## 4	0.975 0.0000				
## 5	0.975 0.0000				

```
## 6      0.975 0.0000
## 7      0.975 0.0000
## 8      1.000 0.0125
```

```
barplot(Gesprächsdauer.df$abs.Häufigkeit, names = Gesprächsdauer.df$Gesprächsdauer, ylim = c(0,35), col
```

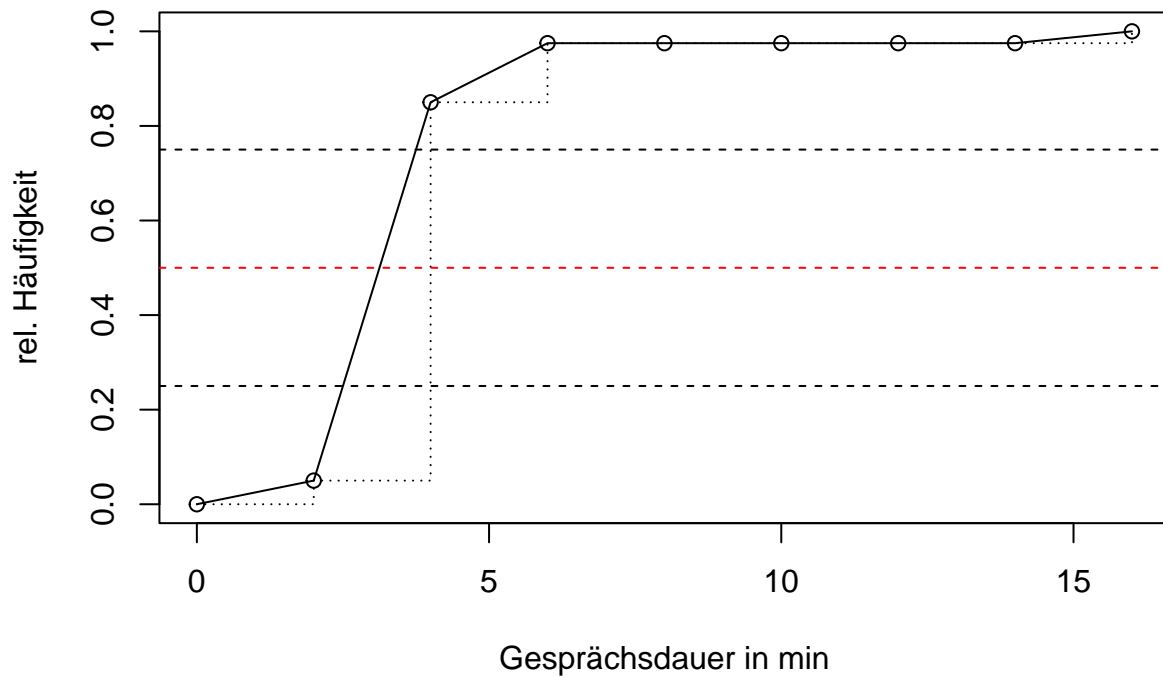


```
barplot(Gesprächsdauer.df$Dichte, names = Gesprächsdauer.df$Gesprächsdauer, ylim = c(0,0.5), col = "blue")
```



```
plot(c(0,2,4,6,8,10,12,14,16), c(0,Gesprächsdauer.df$Summenhfk.), main = "Summenhäufigkeit Telefongespräch",
points(c(0,2,4,6,8,10,12,14,16), c(0,Gesprächsdauer.df$Summenhfk.))
lines(c(0,2,4,6,8,10,12,14,16), c(0,Gesprächsdauer.df$Summenhfk.), type = "s", lty = 3)
abline(0.25, 0, lty = "dashed")
abline(0.75, 0, lty = "dashed")
abline(0.5, 0, lty = "dashed", col = "red")
```


Summenhäufigkeit Telefongespräch



Stiftlänge

```
Länge.vec <- c(3.3,3.4,3.5,3.6,3.7,3.8)
Länge_Intervall.vec <- c("[3.3,3.4]", "(3.4,3.5]", "(3.5,3.6]", "(3.6,3.7]", "(3.7,3.8]")
Länge_Klasse.vec <- tabulate(cut(Verteilungsvergleich.df$Stiftlaenge_cm, breaks = Länge.vec))
Stiftlänge.df <- data.frame(Stiftlänge = Länge_Intervall.vec, Klassenbreite = rep(0.1,5), Klassenmitte = rep(3.35,5))

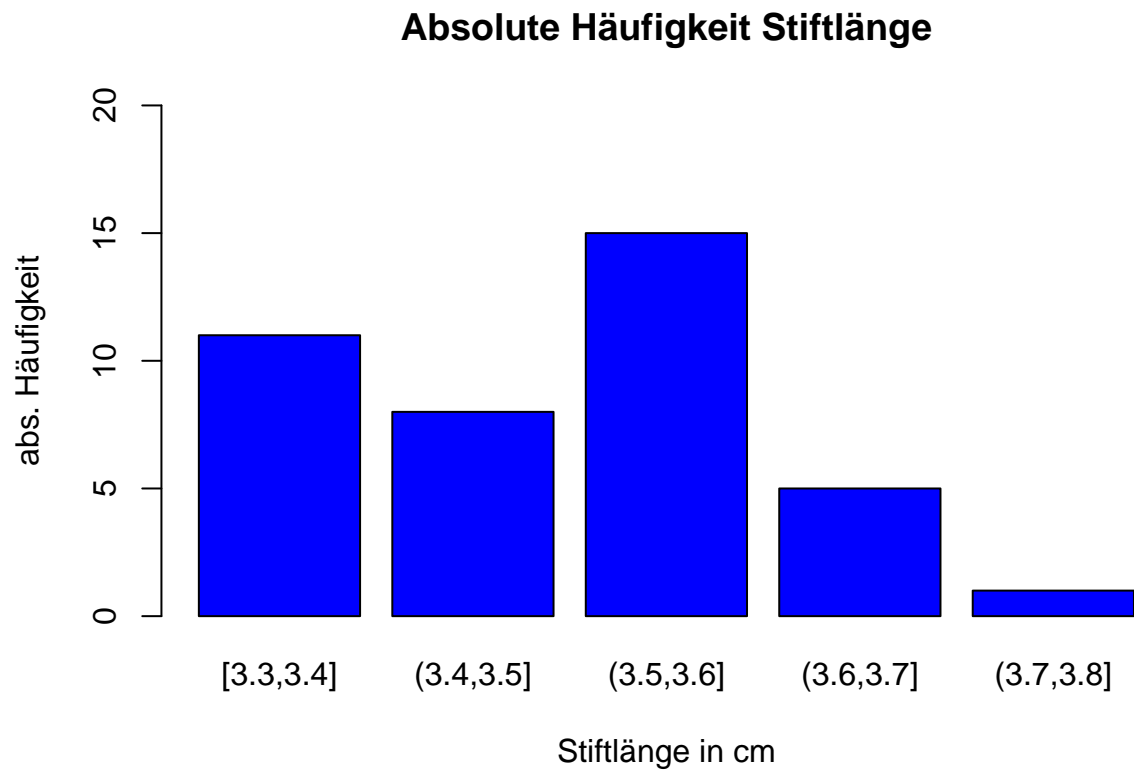
Stiftlänge.df <- data.frame(Stiftlänge.df, Summenhfk. = cumsum(Stiftlänge.df$rel.Häufigkeit), Dichte = 1/11)

Stiftlänge.df
```

##	Stiftlänge	Klassenbreite	Klassenmitte	abs.Häufigkeit	rel.Häufigkeit
## 1	[3.3,3.4]	0.1	3.35	11	0.275
## 2	(3.4,3.5]	0.1	3.45	8	0.200
## 3	(3.5,3.6]	0.1	3.55	15	0.375
## 4	(3.6,3.7]	0.1	3.65	5	0.125
## 5	(3.7,3.8]	0.1	3.75	1	0.025

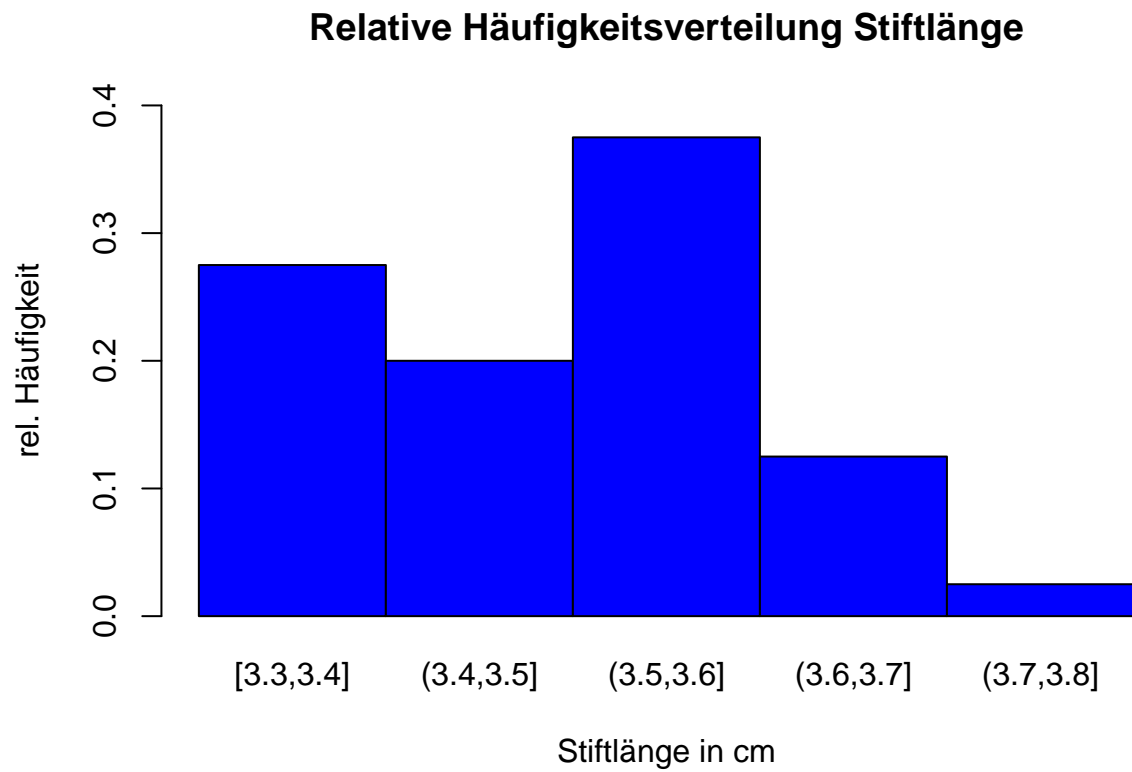
##	Summenhfk.	Dichte
## 1	0.275	2.75
## 2	0.475	2.00
## 3	0.850	3.75
## 4	0.975	1.25
## 5	1.000	0.25

```
barplot(Stiftlänge.df$abs.Häufigkeit, names = Länge_Intervall.vec, main = "Absolute Häufigkeit Stiftlänge")
```



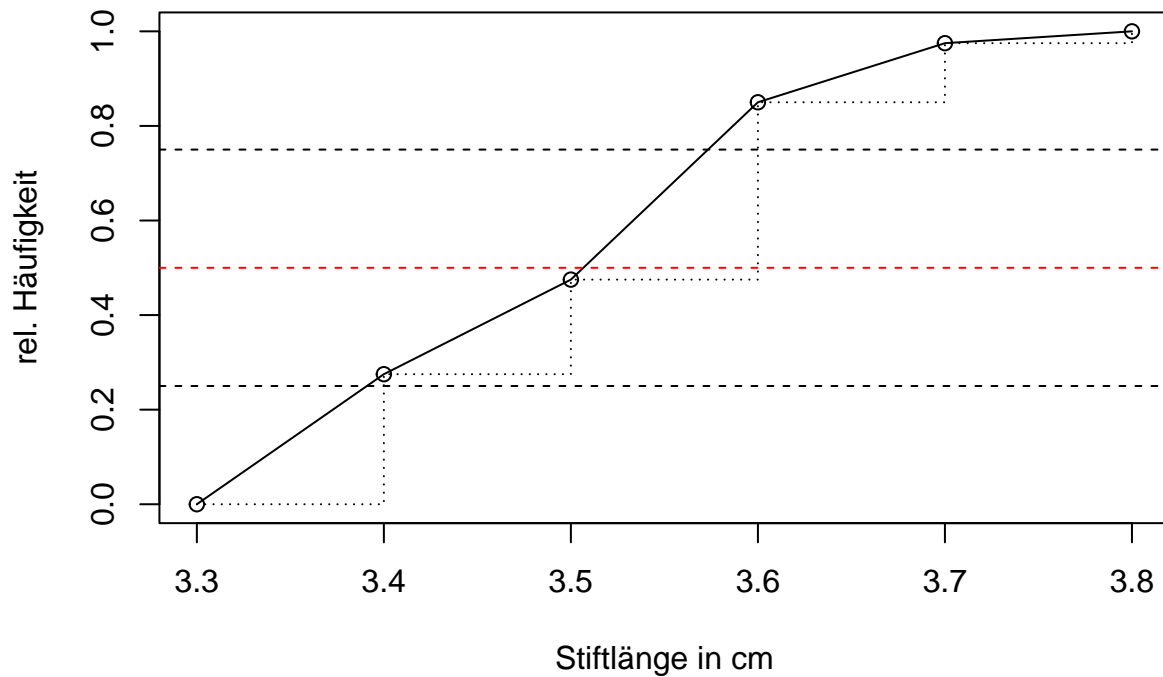
Aufgrund der unglücken Wahl der Maße in cm und der geringen Steuung sowie der folglichen geringen Klassenbreite, wird hier kein Diagramm der Häufigkeitsdichte, sondern ein Diagramm der rel. Häufigkeit erstellt.

```
barplot(Stiftlänge.df$rel.Häufigkeit, names = Länge_Intervall.vec, space = 0, col = "blue", ylim = c(0,
```



```
plot(c(3.3,3.4,3.5,3.6,3.7,3.8), c(0,Stiftlänge.df$Summenhfk.), main = "Summenhäufigkeit Stiftlänge", xlab = "Stiftlänge in cm", ylab = "rel. Häufigkeit", type = "b", col = "blue", lty = 1)
points(c(3.3,3.4,3.5,3.6,3.7,3.8), c(0,Stiftlänge.df$Summenhfk.))
lines(c(3.3,3.4,3.5,3.6,3.7,3.8), c(0,Stiftlänge.df$Summenhfk.), type = "s", lty = 3)
abline(0.25, 0, lty = "dashed")
abline(0.75, 0, lty = "dashed")
abline(0.5, 0, lty = "dashed", col = "red")
```

Summenhäufigkeit Stiftlänge



b) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

Würfel:

arithmetisches Mittel: 3,5

Median: 3

Wartezeit:

arithmetisches Mittel: liegt im Intervall $[0,3]$ und beträgt geschätzt 3

Median: 2,8

Telefongespräche:

arithmetisches Mittel: liegt im Intervall $(2,4]$ und beträgt geschätzt 3

Median: 3,5

Stiftlänge:

arithmetisches Mittel: liegt im Intervall $(3.5,3.6]$ und beträgt geschätzt 3,55

Median: 3,51

c) Bestimme für jede Messreihe jene Merkmalsausprägung, unter welcher die kleinsten 25 % zu finden sind.

Würfel

```
quantile(Verteilungsvergleich.df$Wuerfel, probs = c(0,0.25))
```

```
## 0% 25%  
## 1 2
```

Wartezeit

```
quantile(Verteilungsvergleich.df$Wartezeit_min, probs = c(0,0.25))
```

```
## 0% 25%  
## 0.102 0.743
```

Telefongespräche

```
quantile(Verteilungsvergleich.df$Telgesprae_min, probs = c(0,0.25))
```

```
## 0% 25%  
## 1.6200 2.8425
```

Stiftlänge

```
quantile(Verteilungsvergleich.df$Stiftlaenge_cm, probs = c(0,0.25))
```

```
## 0% 25%  
## 3.34 3.40
```

d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße). Versuche die Lage und die Größe der Lageparameter aufgrund der Eigenschaften der Verteilungen zu verstehen. (z.B.: Worauf deutet die verschiedenen Lage von Median und arithmetischem Mittel, wie verhält sich die Standardabweichung zu Streuparametern wie Spannweite oder Interquartilsabstand?)

Würfel - Zentralmaße:

Modus

```
Modus_Wuerfel.vec <- Verteilungsvergleich.df$Wuerfel  
getmode <- function(Modus_Wuerfel.vec) {  
  uniqv <- unique(Modus_Wuerfel.vec)  
  uniqv[which.max(tabulate(match(Modus_Wuerfel.vec, uniqv)))]  
}  
getmode(Modus_Wuerfel.vec)
```

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

Median

```
median(Verteilungsvergleich.df$Wuerfel)
```

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

Mittelwert

```
mean(Verteilungsvergleich.df$Wuerfel)
```

```
## [1] 3.5
```

Würfel - Streumaße

Minimum

```
min(Verteilungsvergleich.df$Wuerfel)
```

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

Maximum

```
max(Verteilungsvergleich.df$Wuerfel)
```

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

Spannweite

```
range(Verteilungsvergleich.df$Wuerfel)
```

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

Quantile

```
quantile(Verteilungsvergleich.df$Wuerfel)
```

```
##    0%   25%   50%   75%  100%
```

```
##     1     2     3     5     6
```

Standardabweichung

```
sd(Verteilungsvergleich.df$Wuerfel)
```

```
## [1] 1.679438
```

Mittlere absolute Abweichung

```
mean(abs(Verteilungsvergleich.df$Wuerfel-mean(Verteilungsvergleich.df$Wuerfel)))
```

```
## [1] 1.475
```

Wartezeit - Zentralmaße:

Modus

```
Modus_Wartezeit.vec <- Verteilungsvergleich.df$Wartezeit_min  
getmode <- function(Modus_Wartezeit.vec) {  
  uniqv <- unique(Modus_Wartezeit.vec)  
  uniqv[which.max(tabulate(match(Modus_Wartezeit.vec, uniqv)))]  
}  
getmode(Modus_Wartezeit.vec)
```

```
## [1] 3.138
```

Median

```
median(Verteilungsvergleich.df$Wartezeit_min)
```

```
## [1] 2.745
```

Mittelwert

```
mean(Verteilungsvergleich.df$Wartezeit_min)
```

```
## [1] 3.5
```

Wartezeit - Streumaße

Minimum

```
min(Verteilungsvergleich.df$Wartezeit_min)
```

```
## [1] 0.102
```

Maximum

```
max(Verteilungsvergleich.df$Wartezeit_min)
```

```
## [1] 15.788
```

Spannweite

```
range(Verteilungsvergleich.df$Wartezeit_min)
```

```
## [1] 0.102 15.788
```

Quantile

```
quantile(Verteilungsvergleich.df$Wartezeit_min)
```

```
##      0%      25%      50%      75%     100%  
## 0.102 0.743 2.745 5.652 15.788
```

Standardabweichung

```
sd(Verteilungsvergleich.df$Wartezeit_min)
```

```
## [1] 3.493427
```

Mittlere absolute Abweichung

```
mean(abs(Verteilungsvergleich.df$Wartezeit_min-mean(Verteilungsvergleich.df$Wartezeit_min)))
```

```
## [1] 2.66165
```

Telefongespräch - Zentralmaße:

Modus

```
Modus_Telefongespräch.vec <- Verteilungsvergleich.df$Telgesprae_min  
getmode <- function(Modus_Telefongespräch.vec) {  
  uniqv <- unique(Modus_Telefongespräch.vec)  
  uniqv[which.max(tabulate(match(Modus_Telefongespräch.vec, uniqv)))]  
}  
getmode(Modus_Telefongespräch.vec)
```

```
## [1] 3.56
```

Median

```
median(Verteilungsvergleich.df$Telgesprae_min)
```

```
## [1] 3.295
```

Mittelwert

```
mean(Verteilungsvergleich.df$Telgesprae_min)
```

```
## [1] 3.49975
```

Telefongespräche - Streumaße:

Minimum


```
min(Verteilungsvergleich.df$Telgesprae_min)
```

```
## [1] 1.62
```

Maximum

```
max(Verteilungsvergleich.df$Telgesprae_min)
```

```
## [1] 14.21
```

Spannweite

```
range(Verteilungsvergleich.df$Telgesprae_min)
```

```
## [1] 1.62 14.21
```

Quantile

```
quantile(Verteilungsvergleich.df$Telgesprae_min)
```

```
##      0%      25%      50%      75%     100%  
## 1.6200 2.8425 3.2950 3.7925 14.2100
```

Standardabweichung

```
sd(Verteilungsvergleich.df$Telgesprae_min)
```

```
## [1] 1.882077
```

Mittlere absolute Abweichung

```
mean(abs(Verteilungsvergleich.df$Telgesprae_min-mean(Verteilungsvergleich.df$Telgesprae_min)))
```

```
## [1] 0.8622
```

Stiftlänge - Zentralmaße:

Modus

```
Modus_Stiftlänge.vec <- Verteilungsvergleich.df$Stiftlaenge_cm  
getmode <- function(Modus_Stiftlänge.vec) {  
  uniqv <- unique(Modus_Stiftlänge.vec)  
  uniqv[which.max(tabulate(match(Modus_Stiftlänge.vec, uniqv)))]  
}  
getmode(Modus_Stiftlänge.vec)
```

```
## [1] 3.4
```

Median

```
median(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## [1] 3.51
```

Mittelwert

```
mean(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## [1] 3.5
```

Stiftlänge - Streumaße:

Minimum

```
min(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## [1] 3.34
```

Maximum

```
max(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## [1] 3.75
```

Spannweite

```
range(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## [1] 3.34 3.75
```

Quantile

```
quantile(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## 0% 25% 50% 75% 100%
```

```
## 3.34 3.40 3.51 3.57 3.75
```

Standardabweichung

```
sd(Verteilungsvergleich.df$Stiftlaenge_cm)
```

```
## [1] 0.1059511
```

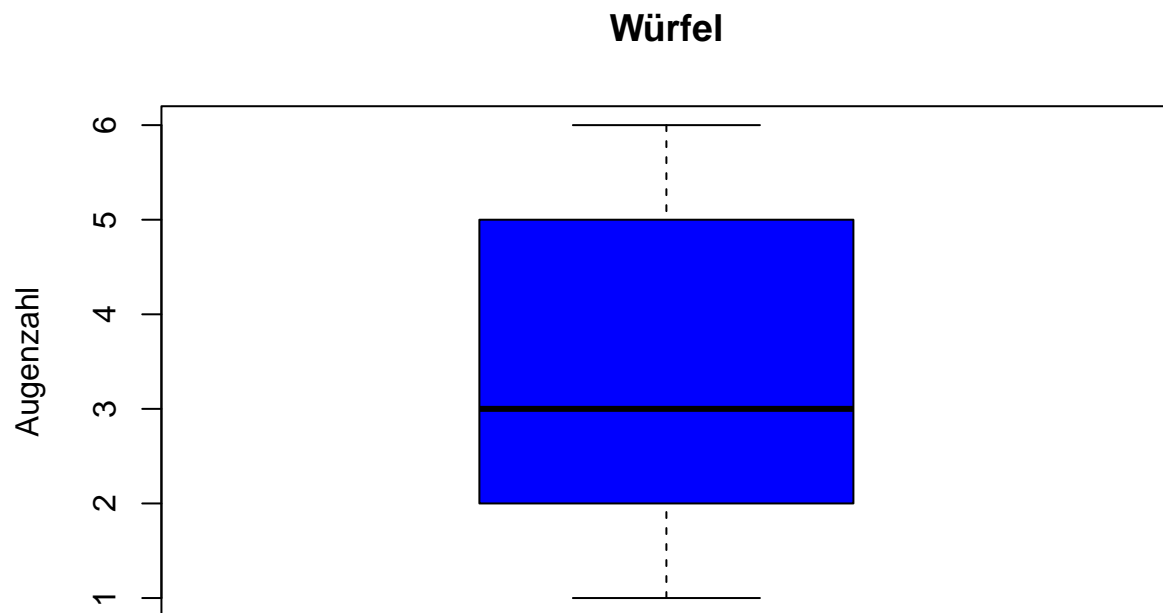
Mittlere absolute Abweichung

```
mean(abs(Verteilungsvergleich.df$Stiftlaenge_cm-mean(Verteilungsvergleich.df$Stiftlaenge_cm)))
```

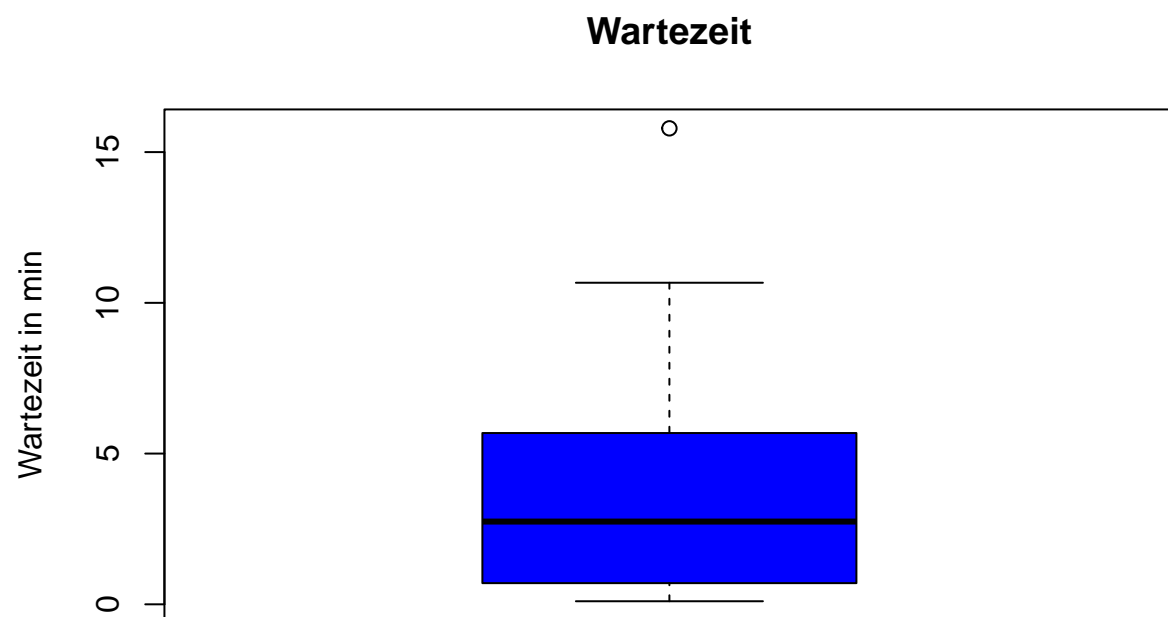
```
## [1] 0.0875
```

e) Zeichne für jede Messreihe einen Boxplot

```
boxplot(Verteilungsvergleich.df$Wuerfel, col = "blue", main = "Würfel", ylab = "Augenzahl")
```

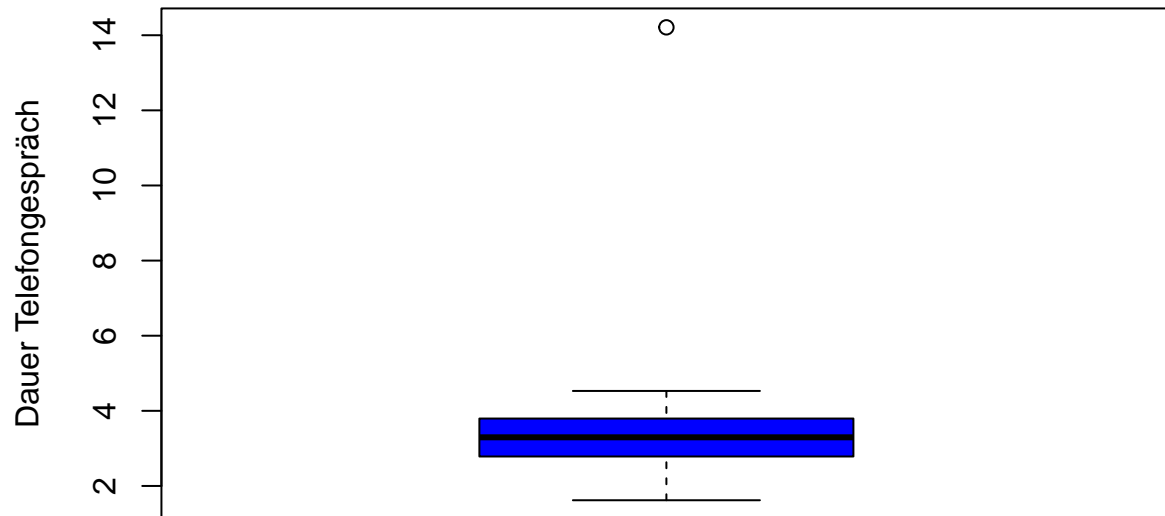


```
boxplot(Verteilungsvergleich.df$Wartezeit_min, col = "blue", ylab = "Wartezeit in min", main = "Wartezeit")
```

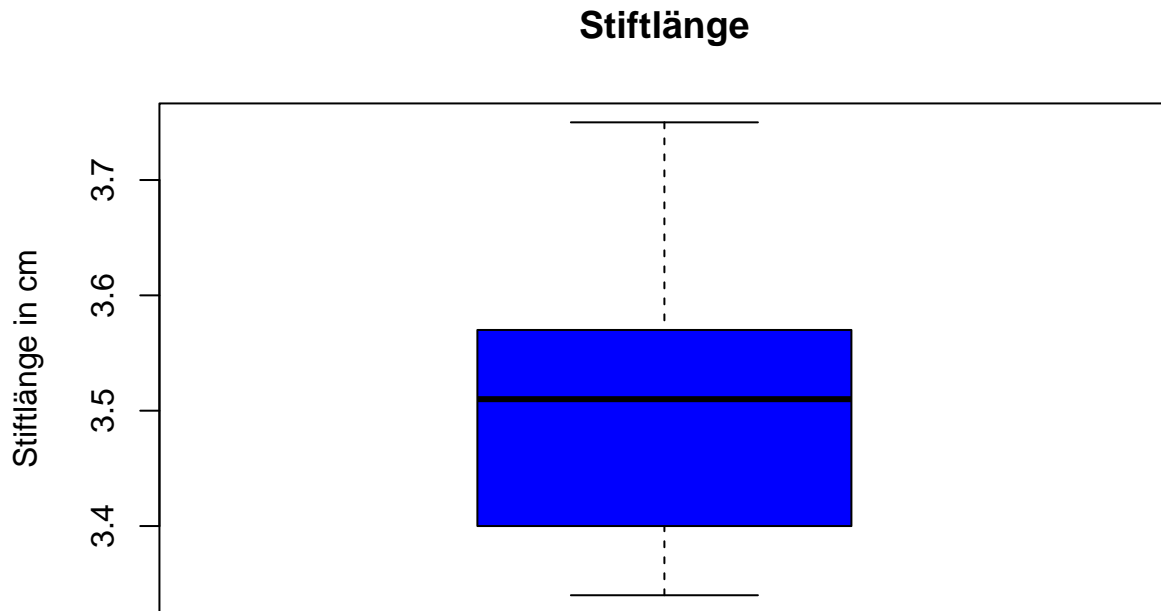


```
boxplot(Verteilungsvergleich.df$Telgesprae_min, main = "Telefongespräch", ylab = "Dauer Telefongespräch")
```

Telefongespräch



```
boxplot(Verteilungsvergleich.df$Stiftlaenge_cm, main = "Stiftlänge", ylab = "Stiftlänge in cm", col = "blue")
```



B-5) Lageparameter als Sicherheitskennzahlen. Eine Fluggesellschaft wirbt damit, dass pro 489 Millionen Passagierkilometer lediglich 1 Todesfall zu beklagen war. (Das klingt sehr gut, wenn man nur 800 km fliegen will). Mit dieser Statistik, so die Fluggesellschaft, ist die Reise mit ihr 10-mal sicherer als eine Autofahrt. (m.a.W.: Im Autoverkehr gibt es 10 Tote auf 48 Millionen Passagierkilometer, oder 1 Toten auf 4,8 Mio. Passagierkilometer.) Allerdings fliegt das Flugzeug im Durchschnitt auch 10-mal schneller als ein Auto fährt. Welche Kennzahlen würde eine Pro-Auto-Initiative dieser Werbung entgegenstellen? (Beachte die effektiven Reisezeiten.)

Wie sich aus der Statistik ableiten lässt, ist laut Fluggesellschaft das Fliegen rund 10 Mal sicherer als eine Autofahrt. Gleichzeitig wird aber auch angegeben, dass ein Flugzeug im Durchschnitt auch 10 Mal so schnell fliegt, als ein Auto fährt. Innerhalb eines gleichen Zeitabschnitts legt ein Flugzeug 48 Mio. Kilometer zurück, ein Auto hingegen nur 4,8 Mio. Kilometer. Aus der Angabe geht hervor, dass für beide Strecken jeweils ein Todesopfer zu beklagen ist. Das Sterberisiko ist zwar pro Kilometer Flugstrecke geringer als pro Kilometer Autofahrt, Für eine bestimmte Reisezeit ist das Risiko zu Sterben allerdings gleich groß. Die Pro-Auto-Initiative könnte also angeben, dass pro Reisetunde das Sterberisiko bei einer Autofahrt gleich groß ist wie bei der Reise mit einem Flugzeug.

B-6) Will Rogers Phänomen.

Das Will Rogers Phänomen beschreibt eine Gegebenheit, bei der durch geschicktes Verschieben von Elementen zwischen Gruppen der Mittelwert in beiden Gruppen verändert werden kann. Je nach Ausgangssituation kann dabei der Mittelwert in beiden Gruppen gesteigert oder wenn gewünscht auch verringert werden.

Die bei dieser Untersuchung erhobenen Daten sind alle korrekt. Durch das Zusammenlegen der Senioren und der Kinder ergeben sich für den Hersteller allerdings bessere Lebensdauerergebnisse. Aufgrund der unterschiedlichen Lebenserwartung von Kindern und Jugendlichen dürfen diese nicht in derselben Gruppe ausgewertet werden.

B-7) Simpsons Paradoxon.

Beim Simpson Paradoxon wird offensichtlich, dass Ergebnisse von Studien, je nach dem wie sie aufbereitet werden, völlig unterschiedliche Ergebnisse liefern, auch wenn die beiden Verfahren korrekt sind und an den Daten keine Manipulation vorgenommen wurde. So ist in diesem Beispiel ersichtlich, dass die Ergebnisse des Versuchs mit dem neuen Fertigungsverfahren einzeln betrachtet deutlich schlechter abschneidet, als bei gemeinsamer Betrachtung.

So werden bei den zusammengefassten Ergebnissen einzelne Teile und Einzelheiten des Produktionsablaufs und Gründe für das schlechtere Abschneiden des neuen Verfahrens nicht mit berücksichtigt. Die Daten über die Durchführung bei der Studie zum neuen Produktionsverfahren werden beim Zusammenfassen also unter den Tisch fallen gelassen.

Ein möglicher Grund für das schlechtere Abschneiden des neuen Fertigungsprozesses könnte in diesem Beispiel sein, dass die Ausschüsse nicht gleichverteilt waren, sondern am Anfang aufgrund der Umstellung viel höher ausfielen. Aus den vorliegenden Daten kann allerdings kein Rückschluss auf diese Behauptung gezogen werden. Ein weiterer Grund könnte sein, dass die Unterschiede in den beiden Werken darauf zurückgeführt werden könne, dass in den Unterschiedlichen Produktionsstätten unterschiedliche Maschinen zum Einsatz kommen. Ältere Maschinen in einem der Werke könnte z.B. dazu führen, dass der Ausschuss größer ist, da sie mit dem neuen Fertigungsverfahren nicht so gut arbeiten können.

Zusammengefasst beschreibt das Simpson Paradoxon, dass Studienergebnisse oftmals durch “für die Studienautoren günstige Auslegung” so präsentiert werden könne, dass das Ergebnis ein anderes ist, als würde man die einzelnen Studien für sich betrachten.

B-8) Studienplatz an einer Hochschule

Bei dieser Untersuchung kommt wieder das Simpson Paradoxon zum Einsatz. Betrachtet man die Zahlen der ersten Tabelle genauer, erkennt man, dass sich deutlich mehr Studenten für einen Studienplatz an der Hochschule beworben haben als Studentinnen. Diese Tatsache wirkt sich schlussendlich netagiv auf das Gesamtergebnis aus. In diesem Studiengang werden jeweils 10 Bewerberinnen und Bewerber aufgenommen. Die Erfolgsquote liegt bei den Bewerberinnen höher, da sich Absolut betrachtet mehr Bewerber für diesen Studienplatz gefunden haben. Auch muss berücksichtigt werden, dass der Studiengang STG-4 mit Abstand am meisten Bewerberinnen und Bewerber hat, in diesem Studiengang allerdings die wenigsten Studienplätze vergeben werden. Schlüsselt man die Daten für diesen Studiengang genauer auf, wird deutlich, dass knapp 74 % aller Studentinnen sich für diesen Studiengang beworben haben. Bei den Bewerbern liegt dieser Anteil nur bei nicht einmal 53 %. Da dieser Unterschied bei der Gesamtbetrachtung vernachlässigt wird, hat es den Anschein, dass Bewerberinnen die schlechteren Chancen haben, einen Studienplatz an dieser Hochschule zu bekommen als Bewerber. Betrachtet man jedoch alle Einflussfaktoren, wie es die QM der Hochschule durch die Aufschlüsselung der Erfolgsquote auf die einzelnen Studiengänge gemacht hat, kommt man zu dem Ergebnis, dass die Chancen von Bewerberinnen höher ist als jene von Bewerbern.

C) Offene Untersuchung

C-1) Das komma-separierte File „Unternehmensumsaetze.csv“ enthält Daten zu den 97 weltweit größten und börsennotierten Konzernen. In dieser Tabelle sind die Umsätze und Gewinne in Mrd. \$ angegeben. Untersuche die Daten mit den bekannten Methoden. Beantworte damit Fragen, wie z.B.: „Wie verteilen sich die Unternehmen auf Länder und Branchen?“, „Wie verteilen sich Gewinne, Umsätze und Mitarbeiter?“, „Welche Branchen generieren besonders viele Umsätze oder Gewinne pro Mitarbeiter?“, und andere mehr. Verwende dazu geeignete Häufigkeitsdarstellungen und Lageparameter. Lege eine passende Regressionsgerade durch die Merkmale „Mitarbeiter“ und „Gewinn“. Wo gibt es Ausreißer. Erkundige Dich nach der Lorenzkurve und wende sie auf die Werte der Merkmale „Umsätze“ und „Mitarbeiter“ an.

Verteilung der Unternehmen nach den Branchen

```
Unternehmensumsaetze.df <- read.csv("Unternehmensumsaetze.csv", sep = ";", dec = ",", header = TRUE)
```

Verteilung der Unternehmen nach den jeweiligen Länder in dem das Unternehmen seinen Hauptwohnsitz hat

```
Land.df <- data.frame(table(Unternehmensumsaetze.df$Land))
names(Land.df) <- c("Staat" , "Anzahl")
Land.df[order(Land.df$Anzahl, decreasing = TRUE),]
```

##	Staat	Anzahl
## 18	USA	31
## 20	Volksrepublik China	13
## 2	Deutschland	10
## 3	Frankreich	9
## 7	Japan	8
## 6	Italien	4
## 4	Grossbritannien	3
## 12	Russland	3
## 5	Indien	2
## 10	Niederlande	2
## 13	Schweiz	2
## 15	Suedkorea	2
## 1	Brasilien	1
## 8	Malaysia	1
## 9	Mexiko	1
## 11	Norwegen	1
## 14	Spanien	1
## 16	Taiwan	1
## 17	Thailand	1
## 19	Venezuela	1

Verteilung der Unternehmen nach Branchen

```
Branche.df <- data.frame(table(Unternehmensumsätze.df$Branche))
names(Branche.df) <- c("Branche", "Anzahl")
Branche.df[order(Branche.df$Anzahl, decreasing = TRUE),]
```

##	Branche	Anzahl
## 16	Öel und Gas	22
## 3	Banken	15
## 1	Automobile	9
## 6	Einzelhandel	6
## 20	Technologie	6
## 21	Telekommunikation	5
## 22	Versicherungen	5
## 23	Versorger	5
## 14	Mischkonzern	3
## 17	Pharmahandel	3
## 7	Eisenbahnbau	2
## 8	Finanzdienstleister	2
## 12	Konsumgüter	2
## 15	Nahrungsmittel	2
## 19	Rohstoffhandel	2
## 2	Automotive Telecom, IT & Tourismus	1
## 4	Bauhauptgewerbe	1
## 5	Chemie	1
## 9	Flugzeugbau	1
## 10	Grosshandel	1
## 11	Informationstechnik	1
## 13	Logistik, Bankwesen, Versicherungswesen	1
## 18	Pharmazie	1

Verteilung der Unternehmen nach Hauptsitz

```
Hauptsitz.df <- data.frame(table(Unternehmensumsätze.df$Hauptsitz))
names(Hauptsitz.df) <- c("Hauptsitz", "Anzahl")
Hauptsitz.df[order(Hauptsitz.df$Anzahl, decreasing = TRUE),]
```

##	Hauptsitz	Anzahl
## 41	Peking	12
## 40	Paris	8
## 52	Tokio	6
## 34	Muenchen	4
## 33	Moskau	3
## 37	New York	3
## 12	Cincinnati	2
## 20	Duesseldorf	2
## 28	London	2
## 43	Rom	2
## 44	San Antonio	2

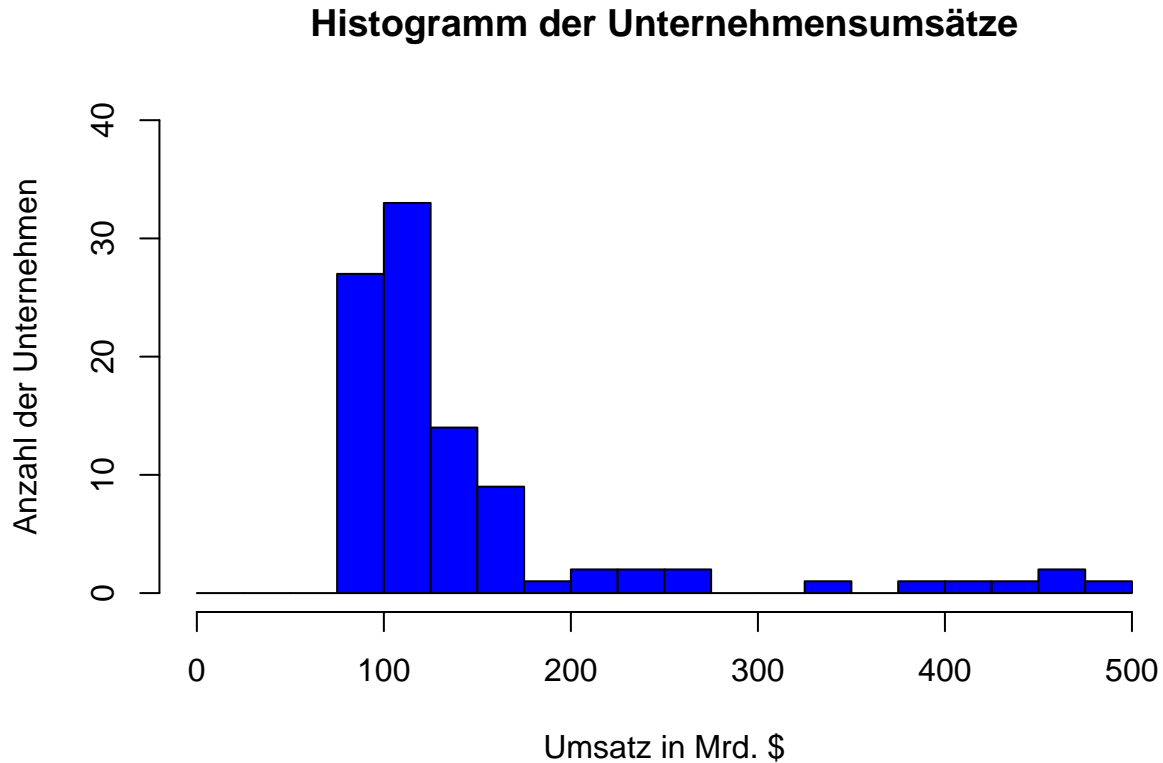
## 45	San Francisco	2
## 47	Seoul	2
## 1	Amsterdam	1
## 2	Armonk	1
## 3	Baar ZG	1
## 4	Bangkok	1
## 5	Bentonville	1
## 6	Bonn	1
## 7	Caracas	1
## 8	Charlotte	1
## 9	Cheshunt	1
## 10	Chesterbrook	1
## 11	Chicago	1
## 13	Courbevoie	1
## 14	Cupertino	1
## 15	Dearborn	1
## 16	Decatur	1
## 17	Den Haag	1
## 18	Detroit	1
## 19	Dublin	1
## 21	Fairfield	1
## 22	Hongkong	1
## 23	Houston	1
## 24	Irving	1
## 25	Issaquah	1
## 26	Kasumigaseki, Chiyoda, Tokio	1
## 27	Kuala Lumpur	1
## 29	Ludwigshafen	1
## 30	Madrid	1
## 31	Mexiko-Stadt	1
## 32	Minnetonka	1
## 35	Mumbai	1
## 36	Neu-Delhi	1
## 38	Omaha	1
## 39	Palo Alto	1
## 42	Rio de Janeiro	1
## 46	San Ramon	1
## 48	St. Louis (Missouri)	1
## 49	Stavanger	1
## 50	Stuttgart	1
## 51	Taipeh	1
## 53	Toyota	1
## 54	Triest	1
## 55	Turin	1
## 56	Tysons Corner	1
## 57	Vevey	1
## 58	Washington, D.C.	1
## 59	Wolfsburg	1
## 60	Woonsocket	1

Aufteilung nach Unternehmensumsätze

```
summary(Unternehmensumsätze.df$Umsatz)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   79.83   98.53  114.30  143.97  146.90  476.29
```

```
hist(Unternehmensumsätze.df$Umsatz, c(0,25,50,75,100,125,150,175,200,225,250,275,300,325,350,375,400,425,450,475,500))
```



Aufteilung nach Unternehmensgewinnen

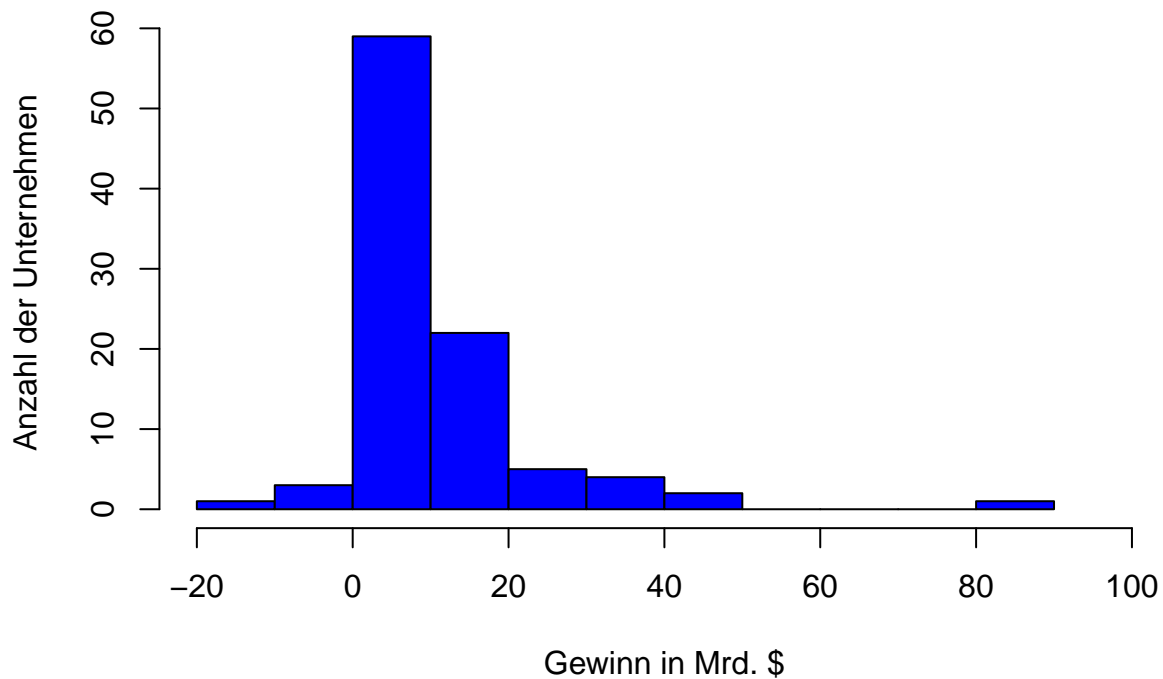
```
summary(Unternehmensumsätze.df$Gewinn)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -12.300   2.700   6.076  10.049  13.700  83.900
```

```
Breaks.vec <- c(-20,-10,0,10,20,30,40,50,60,70,80,90)
```

```
hist(Unternehmensumsätze.df$Gewinn, Breaks.vec, main = "Histogramm der Unternehmensgewinne", xlab = "Gewinn in Mrd. $")
```

Histogramm der Unternehmensgewinne



Sortierung nach Umsatz je Mitarbeiter in \$

```
Mitarbeit_Bran.vec <- aggregate(Unternehmensumsätze.df$Mitarbeiter, by = list(Branche = Unternehmensumsätze.df$Branche), FUN = sum)
Umsatz_Bran.vec <- aggregate(Unternehmensumsätze.df$Umsatz, by = list(Branche = Unternehmensumsätze.df$Branche), FUN = sum)
Umsatz_Mitarbeit_Bran.df <- data.frame(Bran = Umsatz_Bran.vec$Branche, Ums_Mit = 1000000000*Umsatz_Bran.vec$Umsatz)
names(Umsatz_Mitarbeit_Bran.df) <- c("Branche", "Umsatz_je_Mitarbeiter")
Umsatz_Mitarbeit_Bran.df[order(Umsatz_Mitarbeit_Bran.df$Umsatz_je_Mitarbeiter, decreasing = TRUE),]
```

##	Branche	Umsatz_je_Mitarbeiter
## 19	Rohstoffhandel	4591305.6
## 17	Pharmahandel	4130774.1
## 18	Pharmazie	3462518.6
## 13	Logistik, Bankwesen, Versicherungswesen	1382954.5
## 22	Versicherungen	1034492.8
## 16	Öl und Gas	997392.1
## 5	Chemie	883562.3
## 8	Finanzdienstleister	702052.8
## 14	Mischkonzern	683145.9
## 1	Automobile	612787.8
## 4	Bauhauptgewerbe	587860.8
## 12	Konsumgueter	570462.8
## 15	Nahrungsmittel	527599.7

## 9	Flugzeugbau	496691.5
## 21	Telekommunikation	487441.5
## 3	Banken	470430.1
## 23	Versorger	385077.2
## 7	Eisenbahnbau	370829.4
## 10	Grosshandel	345464.9
## 20	Technologie	313889.9
## 6	Einzelhandel	277279.8
## 11	Informationstechnik	231324.7
## 2	Automotive Telecom, IT & Tourismus	156818.2

Im Rohstoffsektor und im Pharmaunternehmen sind die vergleichsweise größten Gewinne je Mitarbeiter zu erzielen. Im Rohstoffhandel betragen die Umsätze rund 4,6 Mio. \$ je Mitarbeiter.

Sortierung nach Gewinn je Mitarbeiter nach Branche in \$

```
Gewinn_Bran.vec <- aggregate(Unternehmensumsätze.df$Gewinn, by = list(Branche = Unternehmensumsätze.df$Branche), FUN = sum)

Gewinn_Mitarbeit_Bran.df <- data.frame(Bran = Gewinn_Bran.vec$Branche, Gew_Mit = 100000000*Gewinn_Bran.vec$Gewinn)

names(Gewinn_Mitarbeit_Bran.df) <- c("Branche", "Gewinn_je_Mitarbeiter")

Gewinn_Mitarbeit_Bran.df[order(Gewinn_Mitarbeit_Bran.df$Gewinn_je_Mitarbeiter, decreasing = TRUE),]
```

##	Branche	Gewinn_je_Mitarbeiter
## 18	Pharmazie	565944.068
## 3	Banken	97682.986
## 16	Öl und Gas	58023.869
## 5	Chemie	57584.510
## 14	Mischkonzern	52040.403
## 22	Versicherungen	48601.021
## 13	Logistik, Bankwesen, Versicherungswesen	43472.727
## 21	Telekommunikation	41031.404
## 11	Informationstechnik	38264.241
## 15	Nahrungsmittel	33732.924
## 20	Technologie	30803.964
## 1	Automobile	30706.404
## 9	Flugzeugbau	26290.138
## 17	Pharmahandel	26028.949
## 8	Finanzdienstleister	23404.317
## 12	Konsumgüter	18909.839
## 2	Automotive Telecom, IT & Tourismus	13484.848
## 4	Bauhauptgewerbe	9550.085
## 6	Einzelhandel	7482.463
## 7	Eisenbahnbau	4980.159
## 23	Versorger	3512.645
## 10	Grosshandel	2016.379
## 19	Rohstoffhandel	-99416.667

Die höchsten Gewinne je Mitarbeiter können mit rund 570.00 \$ je Mitarbeiter die Unternehmen in der Pharmazie aufweisen.

Sortierung nach Gewinn je Mitarbeiter nach Unternehmen in \$

```

Mitarbeit_Firma.vec <- aggregate(Unternehmensumsätze.df$Mitarbeiter, by = list(Firma = Unternehmensumsätze.df$Firma), FUN = sum)
Gewinn_Firma.vec <- aggregate(Unternehmensumsätze.df$Gewinn, by = list(Firma = Unternehmensumsätze.df$Firma), FUN = sum)
Gewinn_Mitarbeit_Firma.df <- data.frame(Firma = Gewinn_Firma.vec$Firma, Gew_Mit = 1000000000*Gewinn_Firma.vec$Gewinn)
names(Gewinn_Mitarbeit_Firma.df) <- c("Firma", "Gewinn_je_Mitarbeiter")
Gewinn_Mitarbeit_Firma.df[order(Gewinn_Mitarbeit_Firma.df$Gewinn_je_Mitarbeiter, decreasing = TRUE),]

```

	Firma	Gewinn_je_Mitarbeiter
## 42	Fannie Mae	11985714.286
## 44	Freddie Mac	9959100.204
## 23	China National Offshore Oil	1432025.293
## 4	Apple	585102.686
## 40	Express Scripts Holding	565944.068
## 74	Petronas	391700.866
## 20	Chevron	350111.948
## 41	ExxonMobil	328758.829
## 17	BP	281187.050
## 87	Statoil	214409.585
## 79	Royal Dutch Shell	181900.000
## 77	PTT Public Company	168859.649
## 71	PDVSA	140295.164
## 73	Petrobras	135428.111
## 29	ConocoPhillips	125033.557
## 93	Valero Energy	123051.682
## 90	Total	116582.036
## 78	Rosneft Oil	107625.689
## 21	China Construction Bank	105937.991
## 55	ICBC	104436.982
## 65	Munich Re	93208.490
## 45	Gazprom	89199.501
## 80	Samsung Electronics	88745.928
## 75	Procter & Gamble	87596.899
## 38	Eni	87054.876
## 97	Wells Fargo	82891.749
## 13	Berkshire Hathaway	71904.836
## 7	AT&T	71367.288
## 59	JPMorgan Chase	68804.606
## 14	BMW Group	63887.051
## 1	Agricultural Bank of China	60348.546
## 94	Verizon	58793.192
## 57	ING Groep	58188.644
## 12	BASF	57584.510
## 37	Enel	57059.448
## 92	UnitedHealth	56565.657
## 91	Toyota Motor	55783.127
## 2	Allianz	55658.104
## 54	HSBC	52943.945
## 62	Lukoil	52000.000
## 28	Citigroup	51503.759
## 60	JX Holdings	45387.028
## 3	AmerisourceBergen	44060.914

## 43	Ford Motor	43902.439
## 67	Nippon Yuese	43472.727
## 47	General Electric	43378.738
## 22	China Mobile	42459.513
## 5	Archer Daniels Midland	42345.277
## 10	Bank of America	40565.526
## 58	International Business Machines	38264.241
## 31	Credit Agricole	37735.418
## 35	E.ON	36037.977
## 63	McKesson	34482.759
## 8	AXA	33837.349
## 33	Daimler	33470.907
## 66	Nestle	32926.829
## 56	Indian Oil Corporation	32322.228
## 15	BNP Paribas	32254.325
## 6	Assicurazioni Generali	30976.743
## 53	Honda Motor	30465.969
## 9	Banco Santander	29997.569
## 36	Electricite de France	29903.693
## 32	CVS Caremark	28220.859
## 16	Boeing	26290.138
## 70	NTT	25865.260
## 48	General Motors	25826.087
## 68	Nissan Motor	24840.764
## 11	Bank of China	19658.494
## 95	Volkswagen	19262.344
## 84	Societe Generale	17542.101
## 69	Noble Group	17428.571
## 30	Costco Wholesale	15625.000
## 81	Siemens	15555.556
## 50	Hewlett-Packard	14588.101
## 88	Tata	13484.848
## 18	Cardinal Health	10470.219
## 39	Exor	10122.139
## 24	China National Petroleum	9781.952
## 27	China State Construction Engineering	9550.085
## 82	Sinopec	8739.906
## 51	Hitachi	8036.101
## 96	Walmart	7272.727
## 34	Deutsche Telekom	5406.400
## 26	China Railway Group	5366.197
## 86	State Grid	5042.325
## 25	China Railway Construction	4481.818
## 83	SK Holdings	4460.686
## 61	Kroger	4424.779
## 19	Carrefour	4049.044
## 76	PSA Peugeot Citroen	3913.520
## 89	Tesco	3767.656
## 52	Hon Hai Precision Industry (Foxconn)	2922.078
## 64	Metro	2016.379
## 85	Sony	-35550.092
## 72	PEMEX	-48765.939
## 46	GDF Suez	-51185.378
## 49	Glencore	-127620.690

Regressionsgerade durch die Merkmale “Mitarbeiter” und “Gewinn”

```
plot(Unternehmensumsätze.df$Mitarbeiter/1000000, Unternehmensumsätze.df$Gewinn, main = "Regression Mita  
abline(lm(Unternehmensumsätze.df$Gewinn~Unternehmensumsätze.df$Mitarbeiter), col = "red")
```



Den größten Ausreißer, also das Unternehmen, das den größten Gewinn je Mitarbeiter erzielt, gibt es bei der Firma Fanni Mae mit fast 12 Mrd. \$ Gewinn je Mitarbeiter. Berechnet wird dieser Ausreißer in der vorherigen Berechnung bei “Sortierung nach Gewinn je Mitarbeiter nach Unternehmen in \$”

Lorenzkurve angewandt auf die Merkmale “Umsätze” und “Mitarbeiter”

```
"ineq" %in% installed.packages()
```

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

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

```
plot(Lc(Unternehmensumsätze.df$Mitarbeiter, Unternehmensumsätze.df$Umsatz), main = "Umsätze je Mitarbei
```