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:

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, 1, 2)  
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 auch sen 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

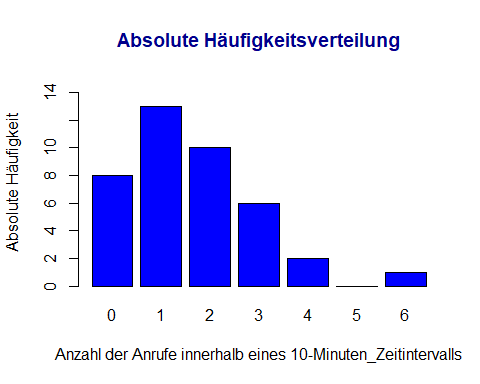
##### **b) Ermittle die absolute und relative Häufigkeitstabelle der eingehenden Anrufe und stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.**

##### **Absolute Häufigkeit:**

table(Liste.vec)

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

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", ylab = "Absolute Häufigkeit", main = "Absolute Häufigkeitsverteilung", col.main = "darkblue", col = "blue")

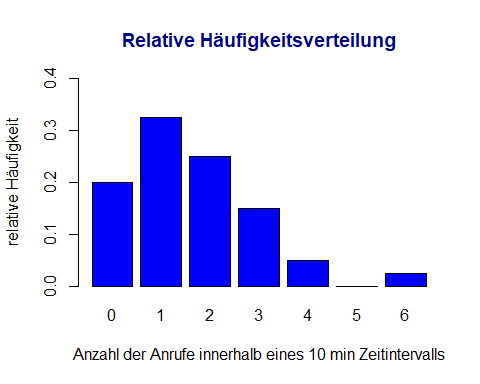


##### **Relative Häufigkeit:**

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

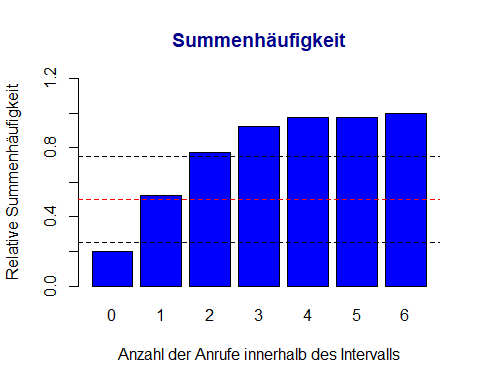
barplot(table(factor(Liste.vec,levels=c(0,1,2,3,4,5,6)))/sum(table(Liste.vec)), ylim = c(0,0.4), xlab = "Anzahl der Anrufe innerhalb eines 10 min Zeitintervalls",ylab = "relative Häufigkeit", main = "Relative Häufigkeitsverteilung", col.main = "darkblue", col = "blue")

 ##### **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)), ylim = c(0,1.2), xlab = "Anzahl der Anrufe innerhalb des Intervalls", ylab = "Relative Summenhäufigkeit", main = "Summenhäufigkeit", col.main = "darkblue", col = "blue")  
abline(0.25, 0, lty = "dashed")  
abline(0.5, 0, lty = "dashed", col = "red")  
abline(0.75, 0, lty = "dashed")



##### **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 arithmestische 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 Lageparameter (Zentralmaße und Streumaße)**

\_\_\*Zentralmaße:\*\*\_\_  
***Modus***

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

## [1] 1

**Median**

median(Liste.vec)

## [1] 1

**Arithmetisches Mittel**

mean(Liste.vec)

## [1] 1.625

***Streumaße:*** Minimum

Minimum <- min(Liste.vec)  
Minimum

## [1] 0

**Maximum**

Maximum <- max(Liste.vec)  
Maximum

## [1] 6

**Spannweite**

Spannweite <- range(Liste.vec)  
Spannweite

## [1] 0 6

**Quantile**

quantile(Liste.vec)

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

**Mittlere absolute Abweichung**

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

## [1] 1.05625

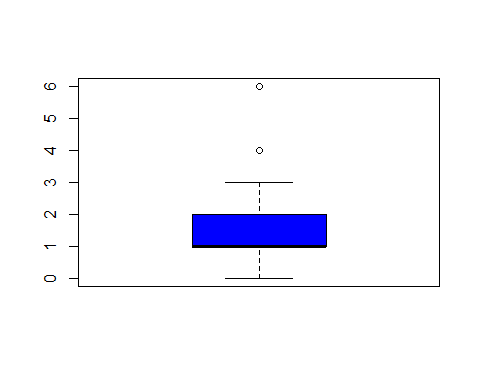
**Standardabweichung**

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]")  
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

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$Lebensdauer, Klassenmitte=c(1,3,5,7,9), Klassenbreite=rep(2,5), Anzahl\_Motoren=Lebensdauer.df$Anzahl\_Motoren)  
  
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

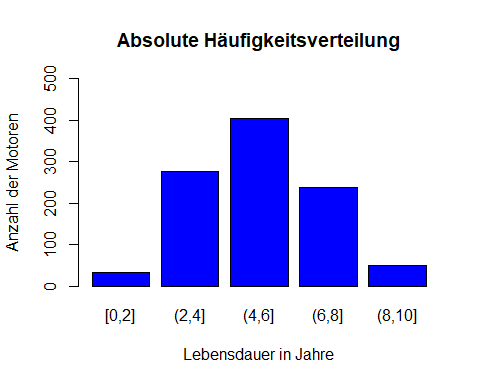
HfgkDichte.vec <- Lebensdauer.df$RelHfgk.vec / Lebensdauer.df$Klassenbreite  
   
Lebensdauer.df <- data.frame(Lebensdauer.df, Dichte=HfgkDichte.vec)  
  
Lebensdauer.df

## 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

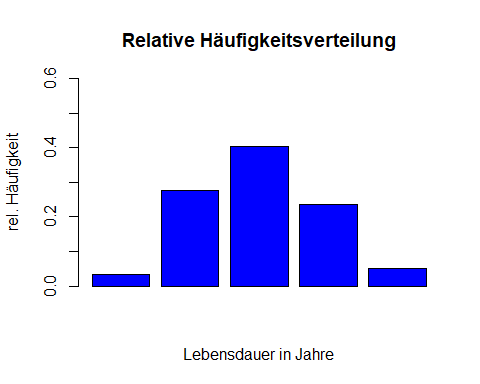
##### **a) Stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.**

**Absolute Häufigkeitsverteilung**

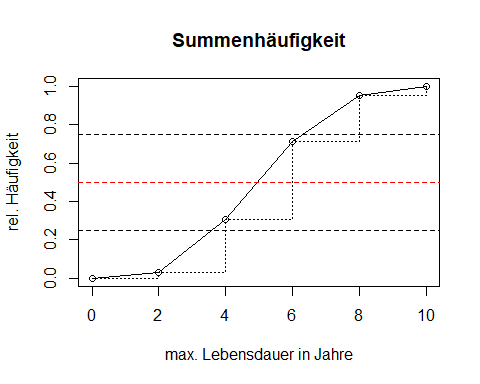
barplot(Lebensdauer.df$Anzahl\_Motoren, names=Lebensdauer.df$Lebensdauer, col = "blue", xlab = "Lebensdauer in Jahre", ylab = "Anzahl der Motoren", main = "Absolute Häufigkeitsverteilung", ylim = c(0,500))

 **Relative Häufigkeitsverteilung**

barplot(Lebensdauer.df$Anzahl\_Motoren / sum(Lebensdauer.df$Anzahl\_Motoren), ylim = c(0,0.6), xlab = "Lebensdauer in Jahre", ylab = "rel. Häufigkeit", col = "blue", main = "Relative Häufigkeitsverteilung")

  
**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 = "rel. Häufigkeit")  
points(X.vec, Y.vec)  
lines(X.vec, Y.vec, type = "s", lty = 3)  
abline(0.25, 0, lty = "dashed")  
abline(0.5, 0, lty = "dashed", col = "red")  
abline(0.75, 0, lty = "dashed")



##### **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 liegt bei rund 5. Beim Ablesen des arithmetischen Mittelwerts müssen links und rechts auf der x-Achse gleich viele Werte vorhanden sein.  
***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 4.

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

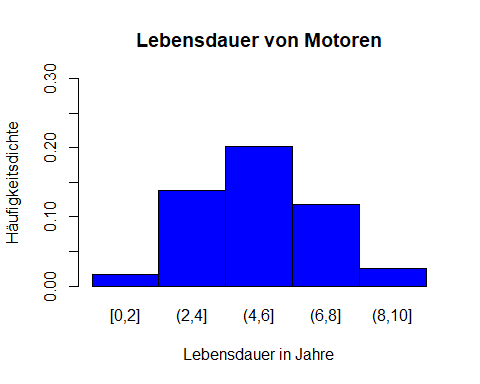
LangeLebensdauer.vec <- Lebensdauer.df$Anzahl\_Motoren[4] + Lebensdauer.df$Anzahl\_Motoren[5]  
  
LangeLebensdauer.vec

## [1] 287

Motoren mit Lebensdauern 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**

barplot(Lebensdauer.df$Dichte, names=Lebensdauer.df$Lebensdauer, main = "Lebensdauer von Motoren", xlab = "Lebensdauer in Jahre", ylab = "Häufigkeitsdichte", ylim = c(0,0.3), space = 0, col = "blue")

 ##### **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)))]  
}  
getmode(Lebensdauer.vec)

## [1] 5

**Median**

median(Lebensdauer.vec)

## [1] 5

**Mittelwert**

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

## [1] 4.99

***Streumaße***  
**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)  
  
Service.df

## Einsätze Tage  
## 1 [0,10] 16  
## 2 (10,20] 48  
## 3 (20,30] 27  
## 4 (30,80] 9

RelHfk.vec <- Service.df$Tage / sum(Service.df$Tage)  
SumHfk.vec <- cumsum(RelHfk.vec)  
  
Service.df <- data.frame(Service.df, relHfk = RelHfk.vec, sumHfk = SumHfk.vec)  
  
Service.df

## Einsätze Tage relHfk sumHfk  
## 1 [0,10] 16 0.16 0.16  
## 2 (10,20] 48 0.48 0.64  
## 3 (20,30] 27 0.27 0.91  
## 4 (30,80] 9 0.09 1.00

Service.df <- data.frame(Service.df, Klassenbreite=c(10,10,10,50), Klassenmitte=c(5,15,25,55))  
Service.df

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

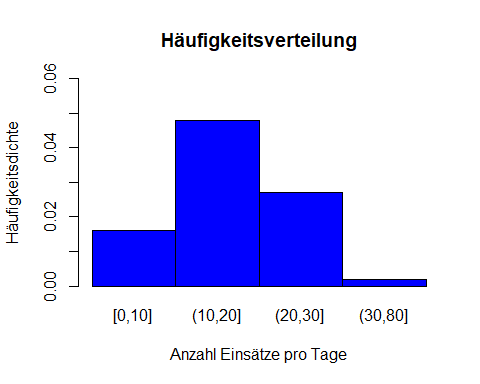
HfkDichte.vec <- Service.df$relHfk / Service.df$Klassenbreite  
  
Service.df <- data.frame(Service.df, Dichte=HfkDichte.vec)  
  
Service.df

## 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

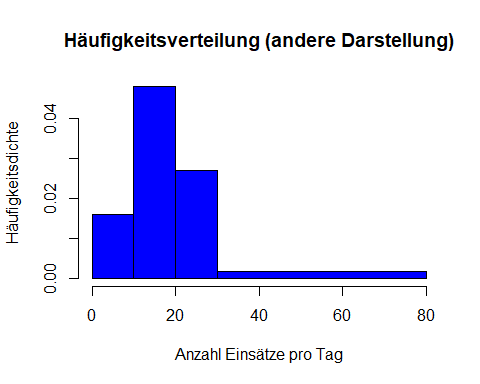
##### **a) Stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.**

**Häufigkeitsverteilung**

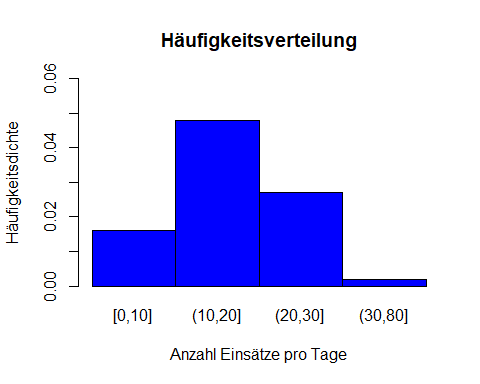
barplot(Service.df$Dichte, names=Service.df$Einsätze, col = "blue", main = "Häufigkeitsverteilung", xlab = "Anzahl Einsätze pro Tage", ylab = "Häufigkeitsdichte", space = 0, ylim = c(0,0.06))

  
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 Einsätzte pro Tag breit. 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.

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", ylab = "Häufigkeitsdichte", col = "blue")

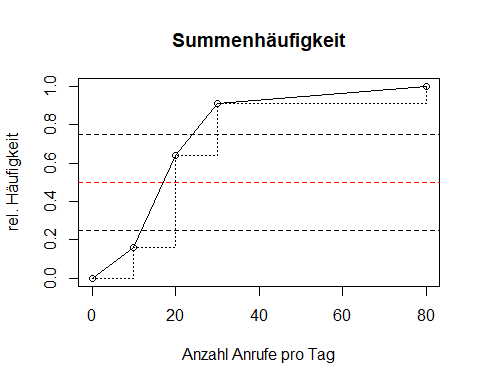


barplot(Service.df$Dichte, names=Service.df$Einsätze, col = "blue", main = "Häufigkeitsverteilung", xlab = "Anzahl Einsätze pro Tage", ylab = "Häufigkeitsdichte", space = 0, ylim = c(0,0.06))



**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", type = "l", lty = 1)  
points(Xa4.vec, Ya4.vec)  
lines(Xa4.vec, Ya4.vec, type = "s", lty = 3)  
abline(0.25, 0, lty = "dashed")  
abline(0.75, 0, lty = "dashed")  
abline(0.5, 0, lty = "dashed", col = "red")



##### **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 aritmethische Mittel wird auf 19 geschätzt. Links und rechts auf der x-Achse müssen gleich viele Werte liegen, da das arithmetische Mittel den Schwerpunkt der Verteilung symbolisiert.  
***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 Einätze 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)))

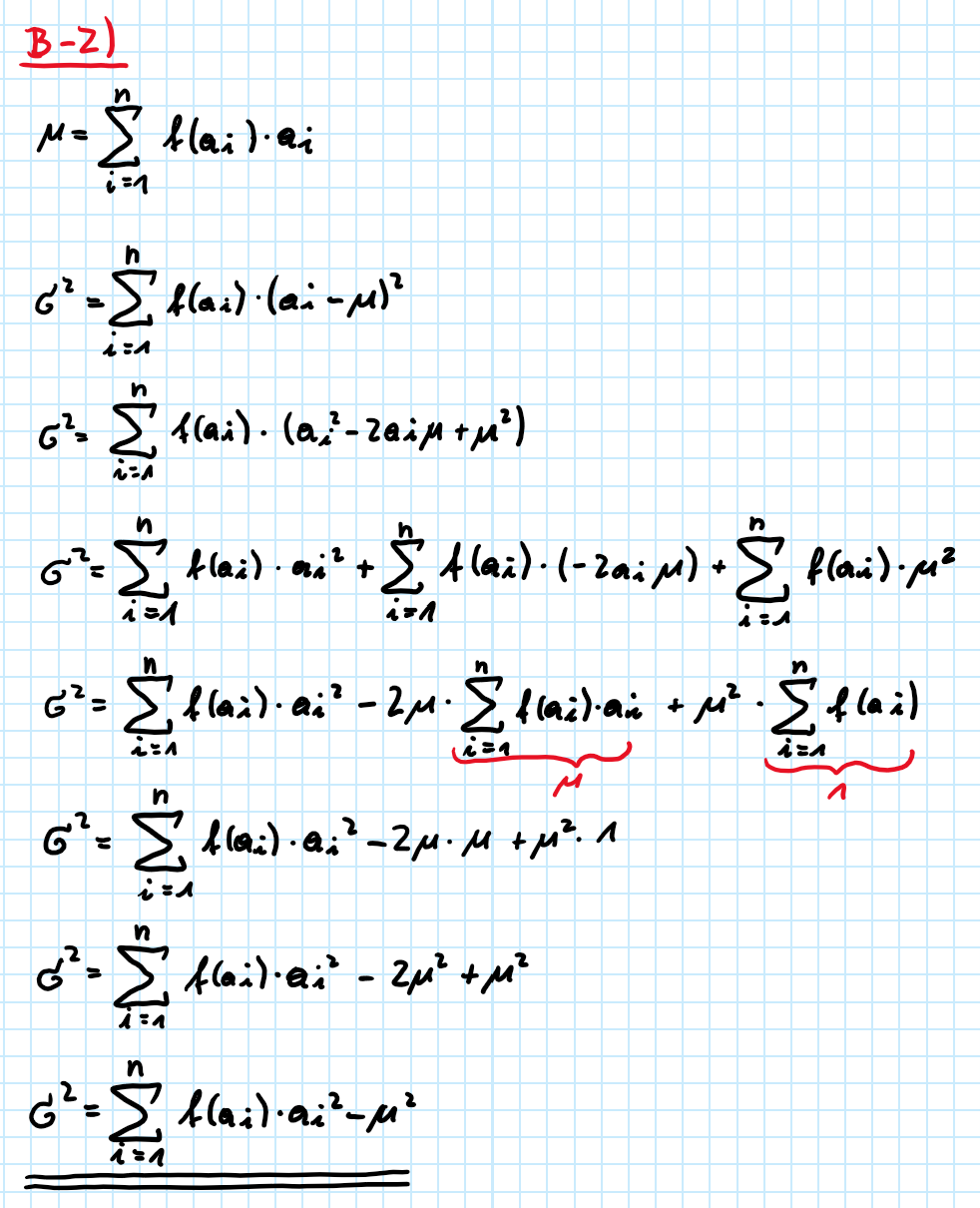
## [1] 9.216

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-2)** Verschiebungssatz zur Berechnung der Standardabweichung



Herleitung des Verschiebungssatz

### **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.

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

##### **a) Welche Skalierung hat dieses Merkmal?**

Die Skalierung dieses Merkmals entspricht einem normalskalierten dichotomem Merkmal.

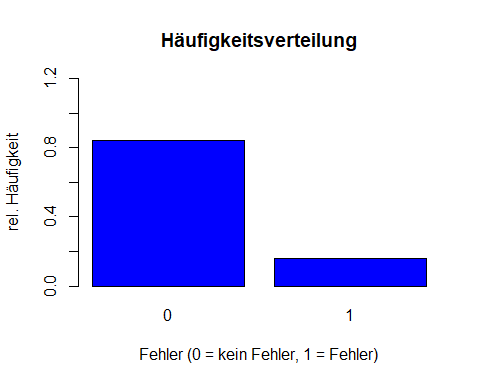
##### **b) Stelle die Messergebnisse in einer Häufigkeitstabelle und grafisch dar.**

***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)", ylab = "rel. Häufigkeit", ylim = c(0,1.2), col = "blue")



##### **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 ? Leite eine Formel her und zeige, wie man sie in diesem Fall einfach aus dem arithmetischen Mittel errechnen kann.**

Standardabweigung:

sd(Fehlerquote.df$fehlerhaft)

## [1] 0.370328

Standardabweichung über arithmetisches Mittel:

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

### **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)

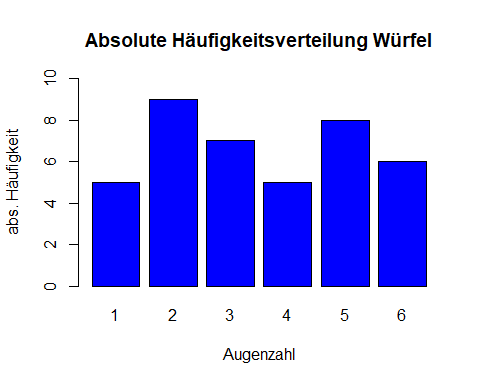
##### **a) Erstelle für jede Messreihe eine Häufigkeitstabelle sowie eine grafische Darstellung der Häufigkeitsverteilung und der Summenhäufigkeit.**

***Würfel***

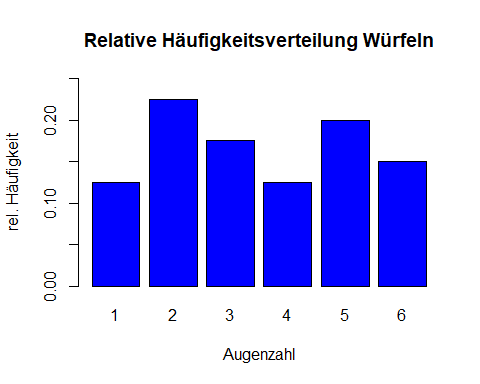
Würfel.df <- data.frame(Augenzahl = (1:6), abs.Häufigkeit = tabulate(Verteilungsvergleich.df$Wuerfel), rel.Häufigkeit = tabulate(Verteilungsvergleich.df$Wuerfel)/length(Verteilungsvergleich.df$Wuerfel))  
  
Würfel.df <- data.frame(Würfel.df, Summenhfk. = cumsum(Würfel.df$rel.Häufigkeit))  
Würfel.df

## 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

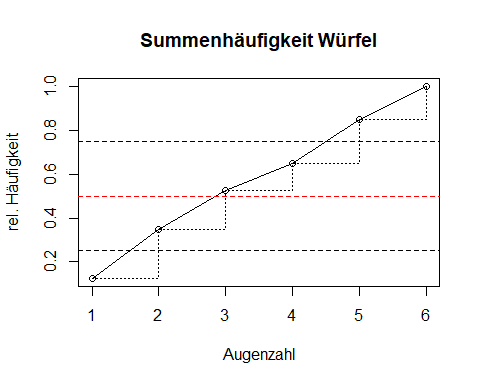
barplot(Würfel.df$abs.Häufigkeit, names = Würfel.df$Augenzahl, ylim = c(0,10), main = "Absolute Häufigkeitsverteilung Würfel", xlab = "Augenzahl", ylab = "abs. Häufigkeit", col = "blue")



barplot(Würfel.df$rel.Häufigkeit, names = Würfel.df$Augenzahl, col = "blue", main = "Relative Häufigkeitsverteilung Würfeln", xlab = "Augenzahl", ylab = "rel. Häufigkeit", ylim = c(0,0.25))



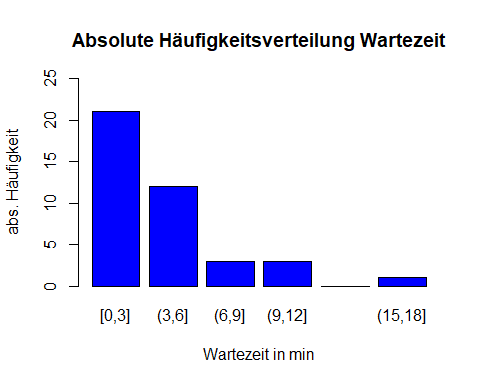
plot(Würfel.df$Augenzahl, Würfel.df$Summenhfk., main = "Summenhäufigkeit Würfel", xlab = "Augenzahl", ylab = "rel. Häufigkeit", type = "l", lty = 1)  
points(Würfel.df$Augenzahl, Würfel.df$Summenhfk.)  
lines(Würfel.df$Augenzahl, Würfel.df$Summenhfk., type = "s", lty = 3)  
abline(0.25, 0, lty = "dashed")  
abline(0.75, 0, lty = "dashed")  
abline(0.5, 0, lty = "dashed", col = "red")

 ***Wartezeit***

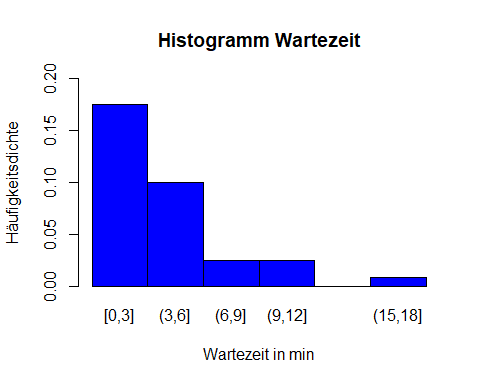
Zeit.vec <- c(0,3,6,9,12,15,18)  
Zeitintervalle.vec <- c("[0,3]", "(3,6]", "(6,9]", "(9,12]", "(12,15]", "(15,18]")  
Klasse.vec <- tabulate(cut(Verteilungsvergleich.df$Wartezeit\_min, breaks = Zeit.vec))  
Wartezeit.df <- data.frame(Zeit = Zeitintervalle.vec, Klassenbreite = rep(3,6), Klassenmitte = c(1.5,4.5,7.5,10.5,13.5,16.5), abs.Häufigkeit = Klasse.vec, rel.Häufigkeit = Klasse.vec/length(Verteilungsvergleich.df$Wartezeit\_min))  
  
Wartezeit.df <- data.frame(Wartezeit.df, Summenhfk. = cumsum(Wartezeit.df$rel.Häufigkeit), Dichte = Wartezeit.df$rel.Häufigkeit/Wartezeit.df$Klassenbreite)  
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

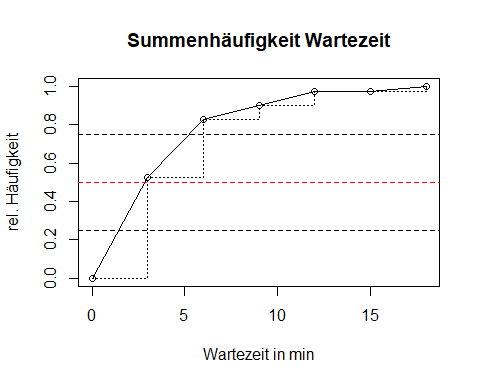
barplot(Wartezeit.df$abs.Häufigkeit, names = Wartezeit.df$Zeit, main = "Absolute Häufigkeitsverteilung Wartezeit", xlab = "Wartezeit in min", ylab = "abs. Häufigkeit", col = "blue", ylim = c(0,25))



barplot(Wartezeit.df$Dichte, names = Wartezeit.df$Zeit, main = "Histogramm Wartezeit", xlab = "Wartezeit in min", ylab = "Häufigkeitsdichte", space = 0, col = "blue", ylim = c(0,0.2))



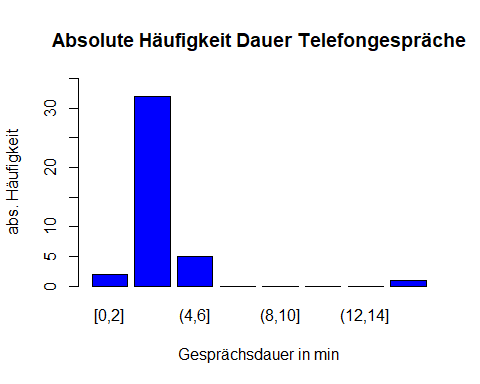
plot(c(0,3,6,9,12,15,18), c(0,Wartezeit.df$Summenhfk.), main = "Summenhäufigkeit Wartezeit", xlab = "Wartezeit in min", ylab = "rel. Häufigkeit", type = "l", lty = 1)  
points(c(0,3,6,9,12,15,18), c(0,Wartezeit.df$Summenhfk.))  
lines(c(0,3,6,9,12,15,18), c(0,Wartezeit.df$Summenhfk.), type = "s", lty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

  
***Telegespräche***

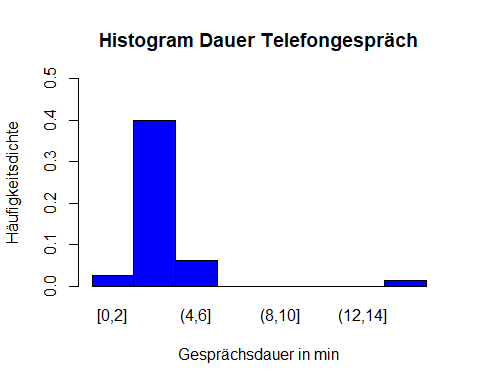
Dauer.vec <- c(0,2,4,6,8,10,12,14,16)  
Dauer\_Intervall.vec <- c("[0,2]", "(2,4]", "(4,6]", "(6,8]", "(8,10]", "(10,12]", "(12,14]", "(14,16]")  
Dauer\_Klasse.vec <- tabulate(cut(Verteilungsvergleich.df$Telgesprae\_min, breaks = Dauer.vec))  
Gesprächsdauer.df <- data.frame(Gesprächsdauer = Dauer\_Intervall.vec, Klassenbreite = rep(2,8), Klassenmitte = c(1,3,5,7,9,11,13,15), abs.Häufigkeit = Dauer\_Klasse.vec, rel.Häufigkeit = Dauer\_Klasse.vec/length(Verteilungsvergleich.df$Telgesprae\_min))  
  
Gesprächsdauer.df <- data.frame(Gesprächsdauer.df, Summenhfk. = cumsum(Gesprächsdauer.df$rel.Häufigkeit), Dichte = Gesprächsdauer.df$rel.Häufigkeit/Gesprächsdauer.df$Klassenbreite)  
  
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

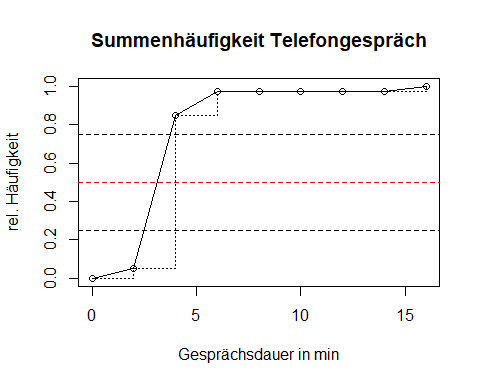
barplot(Gesprächsdauer.df$abs.Häufigkeit, names = Gesprächsdauer.df$Gesprächsdauer, ylim = c(0,35), col = "blue", main = "Absolute Häufigkeit Dauer Telefongespräche", ylab = "abs. Häufigkeit", xlab = "Gesprächsdauer in min")



barplot(Gesprächsdauer.df$Dichte, names = Gesprächsdauer.df$Gesprächsdauer, ylim = c(0,0.5), col = "blue", main = "Histogram Dauer Telefongespräch", ylab = "Häufigkeitsdichte", xlab = "Gesprächsdauer in min", space = 0)



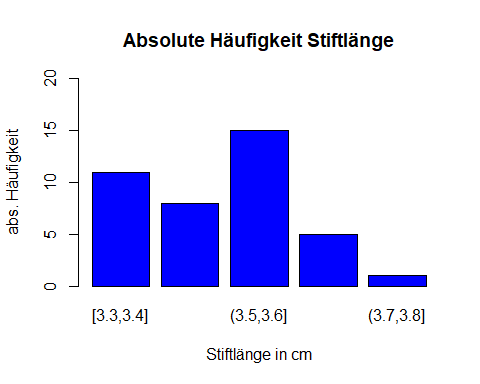
plot(c(0,2,4,6,8,10,12,14,16), c(0,Gesprächsdauer.df$Summenhfk.), main = "Summenhäufigkeit Telefongespräch", xlab = "Gesprächsdauer in min", ylab = "rel. Häufigkeit", type = "l", lty = 1)  
points(c(0,2,4,6,8,10,12,14,16), c(0,Gesprächsdauer.df$Summenhfk.))  
lines(c(0,2,4,6,8,10,12,14,16), c(0,Gesprächsdauer.df$Summenhfk.), type = "s", lty = 3)  
abline(0.25, 0, lty = "dashed")  
abline(0.75, 0, lty = "dashed")  
abline(0.5, 0, lty = "dashed", col = "red")

 ***Stiftlänge***

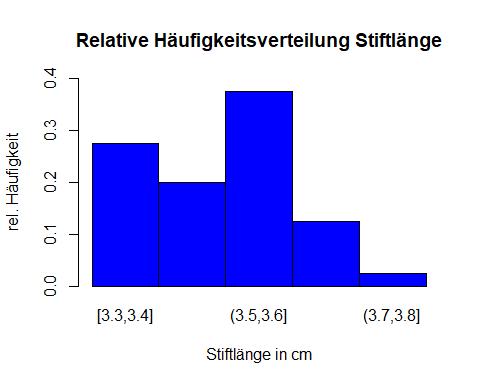
Länge.vec <- c(3.3,3.4,3.5,3.6,3.7,3.8)  
Länge\_Intervall.vec <- c("[3.3,3.4]", "(3.4,3.5]", "(3.5,3.6]", "(3.6,3.7]", "(3.7,3.8]")  
Länge\_Klasse.vec <- tabulate(cut(Verteilungsvergleich.df$Stiftlaenge\_cm, breaks = Länge.vec))  
Stiftlänge.df <- data.frame(Stiftlänge = Länge\_Intervall.vec, Klassenbreite = rep(0.1,5), Klassenmitte = c(3.35,3.45,3.55,3.65,3.75), abs.Häufigkeit = Länge\_Klasse.vec, rel.Häufigkeit = Länge\_Klasse.vec/length(Verteilungsvergleich.df$Stiftlaenge\_cm))  
  
Stiftlänge.df <- data.frame(Stiftlänge.df, Summenhfk. = cumsum(Stiftlänge.df$rel.Häufigkeit), Dichte = Stiftlänge.df$rel.Häufigkeit/Stiftlänge.df$Klassenbreite)  
  
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

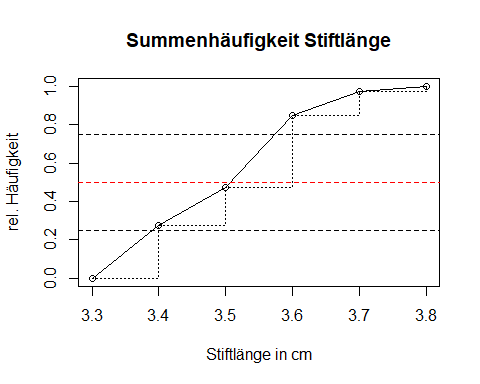
barplot(Stiftlänge.df$abs.Häufigkeit, names = Länge\_Intervall.vec, main = "Absolute Häufigkeit Stiftlänge", ylab = "abs. Häufigkeit", ylim = c(0,20), xlab = "Stiftlänge in cm", col = "blue")

 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(0,0.4), xlab = "Stiftlänge in cm", ylab = "rel. Häufigkeit", main = "Relative Häufigkeitsverteilung Stiftlänge")



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 = "l", 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")



##### **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.**

##### **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

***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

***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

***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

***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

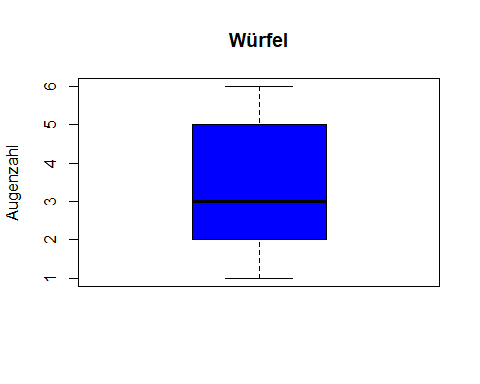
Mittlere absolute Abweichung:

mean(abs(Verteilungsvergleich.df$Stiftlaenge\_cm-mean(Verteilungsvergleich.df$Stiftlaenge\_cm)))

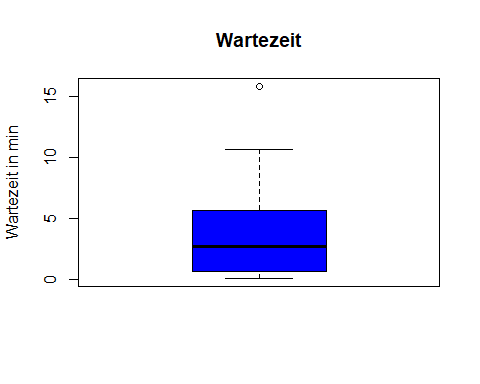
## [1] 0.0875

##### **e) Zeichne für jede Messreihe einen Boxplot**

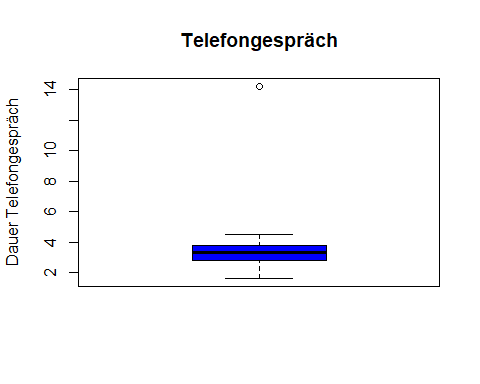
boxplot(Verteilungsvergleich.df$Wuerfel, col = "blue", main = "Würfel", ylab = "Augenzahl")



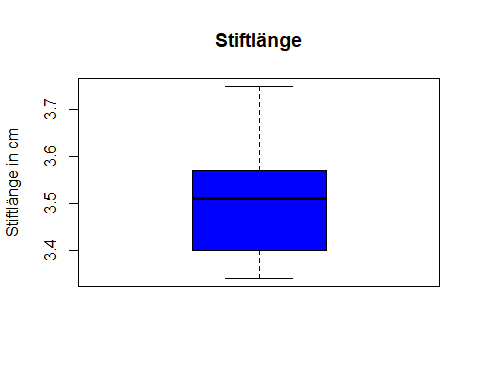
boxplot(Verteilungsvergleich.df$Wartezeit\_min, col = "blue", ylab = "Wartezeit in min", main = "Wartezeit")



boxplot(Verteilungsvergleich.df$Telgesprae\_min, main = "Telefongespräch", ylab = "Dauer Telefongespräch", col = "blue")



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 Reisestunde das Sterberisiko bei einer Autofahrt gleich groß ist wie bei der Reise mit einem Flugzeug.

### **B-6)** Will Rogers Phänomen.

### **B-7)** Simpsons Paradoxon.

### **B-8)** Studienplatz an einer Hochschule

### **C-1)** Offene Untersuchung: 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**

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

## Rang Name Hauptsitz  
## 1 1 Walmart Bentonville  
## 2 2 Royal Dutch Shell Den Haag  
## 3 3 Sinopec Peking  
## 4 4 China National Petroleum Peking  
## 5 5 ExxonMobil Irving  
## 6 6 BP London  
## 7 7 State Grid Peking  
## 8 8 Volkswagen Wolfsburg  
## 9 9 Toyota Motor Toyota  
## 10 10 Glencore Baar ZG  
## 11 11 Total Courbevoie  
## 12 12 Chevron San Ramon  
## 13 13 Samsung Electronics Seoul  
## 14 14 Berkshire Hathaway Omaha  
## 15 15 Apple Cupertino  
## 16 16 AXA Paris  
## 17 17 Gazprom Moskau  
## 18 18 E.ON Duesseldorf  
## 19 19 ConocoPhillips Houston  
## 20 20 Daimler Stuttgart  
## 21 21 General Motors Detroit  
## 22 22 Eni Rom  
## 23 23 Nippon Yuesei Kasumigaseki, Chiyoda, Tokio  
## 24 24 Exor Turin  
## 25 25 ICBC Peking  
## 26 26 Ford Motor Dearborn  
## 27 27 General Electric Fairfield  
## 28 28 Petrobras Rio de Janeiro  
## 29 29 McKesson San Francisco  
## 30 30 Valero Energy San Antonio  
## 31 31 Allianz Muenchen  
## 32 32 Hon Hai Precision Industry (Foxconn) Taipeh  
## 33 33 Societe Generale Paris  
## 34 34 AT&T San Antonio  
## 35 35 CVS Caremark Woonsocket  
## 36 36 PEMEX Mexiko-Stadt  
## 37 37 Fannie Mae Washington, D.C.  
## 38 38 China Construction Bank Peking  
## 39 39 UnitedHealth Minnetonka  
## 40 40 BNP Paribas Paris  
## 41 41 PDVSA Caracas  
## 42 42 Verizon New York  
## 43 43 Lukoil Moskau  
## 44 44 GDF Suez Paris  
## 45 45 Honda Motor Tokio  
## 46 46 Rosneft Oil Moskau  
## 47 47 Agricultural Bank of China Peking  
## 48 48 Assicurazioni Generali Triest  
## 49 49 ING Groep Amsterdam  
## 50 50 Hewlett-Packard Palo Alto  
## 51 51 JX Holdings Tokio  
## 52 52 China State Construction Engineering Peking  
## 53 53 NTT Tokio  
## 54 54 Statoil Stavanger  
## 55 55 China Mobile Peking  
## 56 56 Enel Rom  
## 57 57 JPMorgan Chase New York  
## 58 58 Tata Mumbai  
## 59 59 Siemens Muenchen  
## 60 60 Bank of China Peking  
## 61 61 Costco Wholesale Issaquah  
## 62 62 Nissan Motor Tokio  
## 63 63 Express Scripts Holding St. Louis (Missouri)  
## 64 64 Tesco Cheshunt  
## 65 65 SK Holdings Seoul  
## 66 66 Carrefour Paris  
## 67 67 Bank of America Charlotte  
## 68 68 Cardinal Health Dublin  
## 69 69 BMW Group Muenchen  
## 70 70 Petronas Kuala Lumpur  
## 71 71 Electricite de France Paris  
## 72 72 International Business Machines Armonk  
## 73 73 Nestle Vevey  
## 74 74 Banco Santander Madrid  
## 75 75 Kroger Cincinnati  
## 76 76 BASF Ludwigshafen  
## 77 77 Noble Group Hongkong  
## 78 78 HSBC London  
## 79 79 Hitachi Tokio  
## 80 80 China National Offshore Oil Peking  
## 81 81 China Railway Construction Peking  
## 82 82 Archer Daniels Midland Decatur  
## 83 83 Citigroup New York  
## 84 84 Credit Agricole Paris  
## 85 85 PTT Public Company Bangkok  
## 86 86 PSA Peugeot Citroen Paris  
## 87 87 China Railway Group Peking  
## 88 88 Sony Tokio  
## 89 89 AmerisourceBergen Chesterbrook  
## 90 90 Wells Fargo San Francisco  
## 91 91 Boeing Chicago  
## 92 92 Metro Duesseldorf  
## 93 93 Procter & Gamble Cincinnati  
## 94 94 Munich Re Muenchen  
## 95 95 Indian Oil Corporation Neu-Delhi  
## 96 96 Freddie Mac Tysons Corner  
## 97 97 Deutsche Telekom Bonn  
## Land Umsatz Gewinn Mitarbeiter  
## 1 USA 476.294 16.000 2200000  
## 2 Niederlande 459.599 16.371 90000  
## 3 Volksrepublik China 457.201 8.932 1021979  
## 4 Volksrepublik China 432.007 16.317 1668072  
## 5 USA 407.666 32.580 99100  
## 6 Grossbritannien 396.217 23.451 83400  
## 7 Volksrepublik China 333.387 7.982 1583000  
## 8 Deutschland 261.539 12.072 626715  
## 9 Japan 256.454 18.180 325905  
## 10 Schweiz 232.694 -7.402 58000  
## 11 Frankreich 227.882 11.204 96104  
## 12 USA 220.356 21.423 61189  
## 13 Suedkorea 208.838 27.245 307000  
## 14 USA 182.150 19.476 270858  
## 15 USA 170.910 37.037 63300  
## 16 Frankreich 98.534 5.617 166000  
## 17 Russland 165.016 35.769 401000  
## 18 Deutschland 162.560 2.843 78889  
## 19 USA 161.175 3.726 29800  
## 20 Deutschland 156.628 9.083 271370  
## 21 USA 155.427 5.346 207000  
## 22 Italien 154.108 6.850 78686  
## 23 Japan 152.125 4.782 110000  
## 24 Italien 150.996 2.768 273460  
## 25 Volksrepublik China 148.800 42.700 408859  
## 26 USA 146.900 7.200 164000  
## 27 USA 146.231 13.057 301000  
## 28 Brasilien 141.462 11.094 81918  
## 29 USA 138.000 1.300 37700  
## 30 USA 137.800 2.700 21942  
## 31 Deutschland 134.600 7.900 141938  
## 32 Taiwan 133.100 3.600 1232000  
## 33 Frankreich 132.700 2.800 159616  
## 34 USA 128.700 18.300 256420  
## 35 USA 126.700 4.600 163000  
## 36 Mexiko 125.344 -7.358 150884  
## 37 USA 125.700 83.900 7000  
## 38 Volksrepublik China 125.400 34.900 329438  
## 39 USA 122.500 5.600 99000  
## 40 Frankreich 121.900 6.400 198423  
## 41 Venezuela 121.000 12.900 91949  
## 42 USA 120.500 11.400 193900  
## 43 Russland 119.100 7.800 150000  
## 44 Frankreich 118.500 -12.300 240303  
## 45 Japan 118.200 5.700 187094  
## 46 Russland 117.080 17.100 158884  
## 47 Volksrepublik China 115.400 27.000 447401  
## 48 Italien 115.200 2.540 81997  
## 49 Niederlande 114.295 6.076 104419  
## 50 USA 112.300 5.100 349600  
## 51 Japan 111.000 1.100 24236  
## 52 Volksrepublik China 110.800 1.800 188480  
## 53 Japan 109.054 5.800 224239  
## 54 Norwegen 108.459 6.800 31715  
## 55 Volksrepublik China 107.600 9.200 216677  
## 56 Italien 106.900 4.300 75360  
## 57 USA 106.300 17.900 260157  
## 58 Indien 103.500 8.900 660000  
## 59 Deutschland 106.100 5.600 360000  
## 60 Volksrepublik China 118.200 5.700 289951  
## 61 USA 105.200 2.000 128000  
## 62 Japan 104.600 3.900 157000  
## 63 USA 104.620 17.100 30215  
## 64 Grossbritannien 103.300 1.530 406088  
## 65 Suedkorea 102.100 0.260 58287  
## 66 Frankreich 101.800 1.670 412443  
## 67 USA 101.697 11.431 281791  
## 68 USA 101.090 0.334 31900  
## 69 Deutschland 100.900 7.050 110351  
## 70 Malaysia 100.700 17.180 43860  
## 71 Frankreich 100.300 4.670 156168  
## 72 USA 99.750 16.500 431212  
## 73 Schweiz 99.450 10.800 328000  
## 74 Spanien 98.500 5.800 193349  
## 75 USA 98.370 1.500 339000  
## 76 Deutschland 98.200 6.400 111141  
## 77 Volksrepublik China 97.880 0.244 14000  
## 78 Grossbritannien 97.530 16.200 305984  
## 79 Japan 95.990 2.600 323540  
## 80 Volksrepublik China 95.970 7.700 5377  
## 81 Volksrepublik China 95.746 0.986 220000  
## 82 USA 89.800 1.300 30700  
## 83 USA 93.600 13.700 266000  
## 84 Frankreich 93.600 3.300 87451  
## 85 Thailand 92.500 3.080 18240  
## 86 Frankreich 83.305 0.818 209019  
## 87 Volksrepublik China 91.152 1.524 284000  
## 88 Japan 82.237 -5.784 162700  
## 89 USA 89.100 0.434 9850  
## 90 USA 88.100 21.900 264200  
## 91 USA 86.623 4.585 174400  
## 92 Deutschland 86.350 0.504 249953  
## 93 USA 84.167 11.300 129000  
## 94 Deutschland 83.800 4.400 47206  
## 95 Indien 81.300 1.170 36198  
## 96 USA 81.200 48.700 4890  
## 97 Deutschland 79.829 1.234 228248  
## Branche Geschaeftsfuehrer  
## 1 Einzelhandel Doug McMillon  
## 2 Oel und Gas Ben van Beurden  
## 3 Oel und Gas Fu Chengyu  
## 4 Oel und Gas Zhou Jiping  
## 5 Oel und Gas Rex Tillerson  
## 6 Oel und Gas Robert Dudley  
## 7 Versorger Liu Zhenya  
## 8 Automobile Matthias MÈller  
## 9 Automobile Akio Toyoda  
## 10 Rohstoffhandel Ivan Glasenberg  
## 11 Oel und Gas Patrick Pouyann\_  
## 12 Oel und Gas John S. Watson  
## 13 Technologie Kwon Oh-hyun  
## 14 Mischkonzern Warren Buffett  
## 15 Technologie Tim Cook  
## 16 Versicherungen Henri de Castries  
## 17 Oel und Gas Alexei Miller  
## 18 Versorger Johannes Teyssen  
## 19 Oel und Gas James Mulva  
## 20 Automobile Dieter Zetsche  
## 21 Automobile Daniel Akerson  
## 22 Oel und Gas Paolo Scaroni  
## 23 Logistik, Bankwesen, Versicherungswesen Jiro Saito  
## 24 Finanzdienstleister John Elkann  
## 25 Banken Yang Kaisheng  
## 26 Automobile Mark Fields  
## 27 Mischkonzern Jeffrey R. Immelt  
## 28 Oel und Gas Maria das Gracas Silva Foster  
## 29 Pharmahandel John Hammergren  
## 30 Oel und Gas William Klesse  
## 31 Versicherungen Oliver Baete  
## 32 Technologie Terry Gou  
## 33 Banken Frederic Oudea  
## 34 Telekommunikation Randall Stephenson  
## 35 Einzelhandel Larry J. Merlo  
## 36 Oel und Gas Juan Jose Suarez Coppel  
## 37 Banken Timothy Mayopoulos  
## 38 Banken Wang Hongzhang  
## 39 Versicherungen Stephen Hemsley  
## 40 Banken Jean-Laurent Bonnafe  
## 41 Oel und Gas Rafael Ramirez Carreno  
## 42 Telekommunikation Ivan Seidenberg  
## 43 Oel und Gas Wagit Alekperow  
## 44 Versorger Gerard Mestrallet  
## 45 Automobile Takanobu Ito  
## 46 Oel und Gas Igor Setschin  
## 47 Banken Jiang Chaoliang  
## 48 Versicherungen Mario Greco  
## 49 Finanzdienstleister Jan Hommen  
## 50 Technologie Meg Whitman  
## 51 Oel und Gas Yasushi Kimura  
## 52 Bauhauptgewerbe Yi Jun  
## 53 Telekommunikation Satoshi Miura  
## 54 Oel und Gas Helge Lund  
## 55 Telekommunikation Li Yue  
## 56 Versorger Fulvio Conti  
## 57 Banken James Dimon  
## 58 Automotive Telecom, IT & Tourismus Cyrus Mistry  
## 59 Technologie Joe Kaeser  
## 60 Banken Li Lihui  
## 61 Einzelhandel W. Craig Jelinek  
## 62 Automobile Carlos Ghosn  
## 63 Pharmazie George Paz  
## 64 Einzelhandel Philip Clarke  
## 65 Mischkonzern Chey Tae-won  
## 66 Einzelhandel Georges Plassat  
## 67 Banken Brian Moynihan  
## 68 Pharmahandel George S. Barrett  
## 69 Automobile Harald Krueger  
## 70 Oel und Gas Shamsul Azhar Abbas  
## 71 Versorger Henri Proglio  
## 72 Informationstechnik Virginia Rometty  
## 73 Nahrungsmittel Paul Bulcke  
## 74 Banken Alfredo Saenz  
## 75 Einzelhandel David Dillon  
## 76 Chemie Kurt Bock  
## 77 Rohstoffhandel Yusuf Alireza  
## 78 Banken Stuart Gulliver  
## 79 Technologie Hiroaki Nakanishi  
## 80 Oel und Gas -  
## 81 Eisenbahnbau Li Guorui  
## 82 Nahrungsmittel Patricia Woertz  
## 83 Banken Vikram Pandit  
## 84 Banken Jean-Paul Chifflet  
## 85 Oel und Gas Pailin Chuchottaworn  
## 86 Automobile Philippe Varin  
## 87 Eisenbahnbau -  
## 88 Konsumgueter Kazuo Hirai  
## 89 Pharmahandel Steven Collis  
## 90 Banken John Stumpf  
## 91 Flugzeugbau -  
## 92 Grosshandel Olaf Koch  
## 93 Konsumgueter Robert A. McDonald  
## 94 Versicherungen Nikolaus von Bomhard  
## 95 Oel und Gas Ranbir Singh Butola  
## 96 Banken Donald Layton  
## 97 Telekommunikation Timotheus Hoettges

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

Land.df <- data.frame(table(Unternehmensumsätze.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 Oel 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 Konsumgueter 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),]

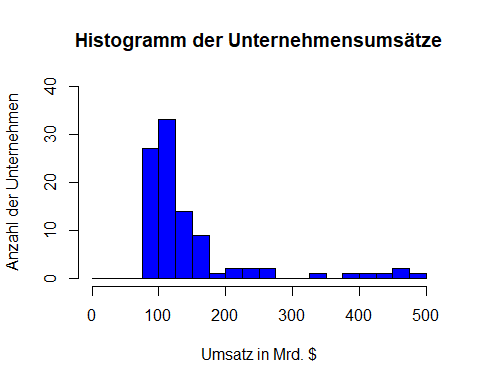
## Hauptsitz Anzahl  
## 41 Peking 12  
## 40 Paris 8  
## 52 Tokio 6  
## 34 Muenchen 4  
## 33 Moskau 3  
## 37 New York 3  
## 12 Cincinnati 2  
## 20 Duesseldorf 2  
## 28 London 2  
## 43 Rom 2  
## 44 San Antonio 2  
## 45 San Francisco 2  
## 47 Seoul 2  
## 1 Amsterdam 1  
## 2 Armonk 1  
## 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

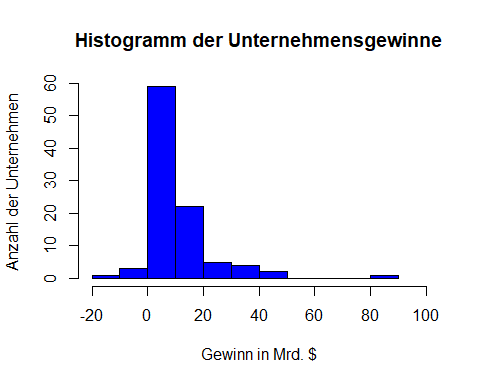
hist(Unternehmensumsätze.df$Umsatz, c(0,25,50,75,100,125,150,175,200,225,250,275,300,325,350,375,400,425,450,475,500), col = "blue", main = "Histogramm der Unternehmensumsätze", xlab = "Umsatz in Mrd. $", ylab = "Anzahl der Unternehmen", ylim = c(0,40))

  
**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. $", ylab = "Anzahl der Unternehmen", col = "blue", xlim = c(-20,100))

  
**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$Branche, Ums\_Mit = 1000000000\*Umsatz\_Bran.vec$x/Mitarbeit\_Bran.vec$x)  
  
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 Oel 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$Branche, Gew\_Mit = 1000000000\*Gewinn\_Bran.vec$x/Mitarbeit\_Bran.vec$x)  
  
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 Oel 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 Konsumgueter 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

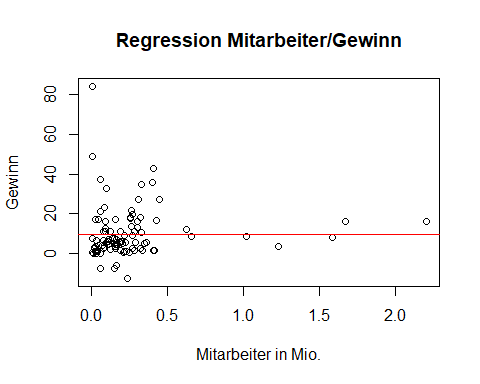
Die höchsten Gewinne je Mitarbeiter können mit rund 570.00 $ je Mitarbeiter die Unternehmen in der Pharmazie aufweisen.  
**Sortierung nach Gewinn je Mitarbeiter nach Unternehmen in $**

Mitarbeit\_Firma.vec <- aggregate(Unternehmensumsätze.df$Mitarbeiter, by = list(Firma = Unternehmensumsätze.df$Name), FUN = sum)  
  
Gewinn\_Firma.vec <- aggregate(Unternehmensumsätze.df$Gewinn, by = list(Firma = Unternehmensumsätze.df$Name), FUN = sum)  
  
Gewinn\_Mitarbeit\_Firma.df <- data.frame(Firma = Gewinn\_Firma.vec$Firma, Gew\_Mit = 1000000000\*Gewinn\_Firma.vec$x/Mitarbeit\_Firma.vec$x)  
  
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 Yuesei 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 Mitarbeiter/Gewinn", xlab = "Mitarbeiter in Mio.", ylab = "Gewinn")  
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)  
  
par(pty = "s")  
plot(Lc(Unternehmensumsätze.df$Mitarbeiter, Unternehmensumsätze.df$Umsatz), main = "Umsätze je Mitarbeiter", xlab = "Mitarbeiter", ylab = "Umsätze")

