Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2016

Veranstaltungen zur Statistik für BW/IM Sommersemester 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Vorlesung Statistik	Etschberger	Mi, 14.00-17.00	B2.14	16.03.2016
Vorlesung Statistik PLUS	Etschberger/Jansen	- Blocktermin -	?	?
Übung Statistik	Etschberger	Mi, 17.00-18.30	A1.10	30.03.2016
Übung Statistik	Jansen	Di, 11.30-13.00	W1.06	22.03.2016
Übung Statistik	Jansen	Di, 14.00-15.30	W2.14	22.03.2016
Übung Statistik	Jansen	Mi, 11.30-13.00	W2.11	30.03.2016
Übung Statistik	Jansen	Do, 14.00-15.30	W2.14	31.03.2016
Übung Statistik	Schneller	Do, 14.00-15.30	W3.03	31.03.2016
Übung Statistik	Schneller	Do, 15.30-17.00	W3.03	31.03.2016
Übung Statistik	Wins	Di, 14.00-15.30	J3.19	22.03.2016
Übung Statistik	Wins	Di, 15.30-17.00	J3.13	22.03.2016
Offener Statistikraum	Etschberger/Tutoren	?	?	?
Veranstaltungen für Teilnehmer der WiMa-Klausur im Juli 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Tutorium Mathematik	Burkart	Do 13.30-15.00	W1.06	07.04.2016
Tutorium Mathematik	Burkart	Do 15.00-16.15	W1.06	07.04.2016
Offener Matheraum	Jansen/Tutoren	?	?	?

	HSA Statistik SS 2016 Sessionlist		
2			
3	Datum	Statistik für IM/BW	Nr.
4	Mittwoch, 16. März 2016	Einführung, R Installation, Rstudio Einführung, Skalen	1
5	Mittwoch, 23. März 2016	univ. deskr. Stat., Quantile, Plots	2
6	Mittwoch, 30. März 2016	Streuung, Konzentrationsmaße	3
7	Mittwoch, 6. April 2016	Kontingenztabellen, Mosaikplots, Korrelation	4
8	Mittwoch, 13. April 2016	Preisindizes, lineare Regression	5
9	Mittwoch, 20. April 2016	Kombinatorik, Wahrscheinlichkeit	6
10	Mittwoch, 27. April 2016	Wahrscheinlichkeit, diskrete Zufallsvariablen	7
11	Mittwoch, 4. Mai 2016	Pyramid	
12	Mittwoch, 11. Mai 2016	Binomial-, Hypergeom.-, Poisson-Verteilung	8
13	Mittwoch, 18. Mai 2016	Stetige ZV, Gleichverteilung	9
14	Mittwoch, 25. Mai 2016	Normalverteilung, Verteilungsparameter	10
15	Mittwoch, 1. Juni 2016	Schätzfunktionen und Punktschätzer	11
16	Mittwoch, 8. Juni 2016	Konfidenzintervalle	12
17	Mittwoch, 15. Juni 2016	Tests	13
18	Mittwoch, 22. Juni 2016	Puffer, WH, Fragen zur Probekl.	14
19	Mittwoch, 29. Juni 2016	AW Prüfungswoche	

Prof. Dr. Stefan Etschberger
Hochschule Augsburg

Daten einlesen und Struktur anschauen


```r
# Arbeitsverzeichnis setzen (alternativ über Menü)
setwd("C:/ste/work/vorlesungen/2015SS_HSA_Statistik")

# Daten einlesen aus einer csv-Datei (Excel)
MyData = read.csv2(file="../_genericFiles/Daten/Umfrage_HSA_2015_03.csv", header=TRUE)
```

```r
# inspect structure of data
str(MyData)
```

```
## 'data.frame': 670 obs. of  18 variables:
##  $ Jahrgang       : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ Alter          : int  20 25 19 21 25 20 25 20 23 21 ...
##  $ Groesse        : int  174 157 163 185 178 170 165 175 180 161 ...
##  $ Geschlecht     : Factor w/ 2 levels "Frau","Mann": 1 1 1 2 2 1 1 2 2 1 ...
##  $ AlterV         : int  55 54 51 52 60 50 60 52 56 70 ...
##  $ AlterM         : int  53 61 49 50 63 55 60 49 50 55 ...
##  $ GroesseV       : int  187 185 178 183 170 183 185 175 175 180 ...
##  $ GroesseM       : int  169 160 168 165 160 160 170 169 170 165 ...
##  $ Geschwister    : num  3 1 1 4 2 2 4 1 1 2 ...
##  $ Farbe          : Factor w/ 6 levels "blau","gelb",..: 4 6 4 4 1 6 1 6 4 4 ...
##  $ AusgKomm       : num  240 119 270 40 550 ...
##  $ AnzSchuhe      : int  25 30 25 6 5 65 10 7 10 22 ...
##  $ AusgSchuhe     : int  450 300 100 100 80 250 150 400 150 300 ...
##  $ Essgewohnheiten: Factor w/ 5 levels "carnivor","fruktarisch",..: 1 1 1 1 1 1 5 1 1 1 ...
##  $ Raucher        : Factor w/ 2 levels "ja","nein": NA 2 2 2 1 2 2 2 2 1 ...
##  $ NoteMathe      : num  2.3 3.3 1.7 2 4 4 3.3 2.7 3.7 3.3 ...
##  $ MatheZufr      : Ord.factor w/ 4 levels "unzufrieden"<..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Studiengang    : Factor w/ 5 levels "BW","ET","IM",..: NA NA NA NA NA NA NA NA NA NA ...
```

Erste Zeilen der Datentabelle

```
# Erste Zeilen in Datentabelle
head(MyData, 6)
```

```
##   Jahrgang Alter Groesse Geschlecht AlterV AlterM GroesseV GroesseM Geschwister   Farbe AusgKomm
## 1     2015    20     174       Frau     55     53      187      169           3 schwarz    240.0
## 2     2015    25     157       Frau     54     61      185      160           1   weiss    119.4
## 3     2015    19     163       Frau     51     49      178      168           1 schwarz    270.0
## 4     2015    21     185       Mann     52     50      183      165           4 schwarz     40.0
## 5     2015    25     178       Mann     60     63      170      160           2    blau    550.0
## 6     2015    20     170       Frau     50     55      183      160           2   weiss    420.0
##   AnzSchuhe AusgSchuhe Essgewohnheiten Raucher NoteMathe MatheZufr Studiengang
## 1        25        450        carnivor    <NA>       2.3   geht so        <NA>
## 2        30        300        carnivor    nein       3.3   geht so        <NA>
## 3        25        100        carnivor    nein       1.7   geht so        <NA>
## 4         6        100        carnivor    nein       2.0   geht so        <NA>
## 5         5         80        carnivor      ja       4.0   geht so        <NA>
## 6        65        250        carnivor    nein       4.0   geht so        <NA>
```

```
# lege MyData als den "Standard"-Datensatz fest
attach(MyData)
```

```
# Wie Viele Objekte gibt's im Datensatz?
nrow(MyData)
```

```
## [1] 670
```

```
# Wie Viele Merkmale?
ncol(MyData)
```

```
## [1] 18
```

Daten kennenlernen


```
# Auswahl spezieller Objekte und Merkmale über [Zeile, Spalte]
MyData[1:3, 2:5]
```

```
##   Alter Groesse Geschlecht AlterV
## 1    20     174       Frau     55
## 2    25     157       Frau     54
## 3    19     163       Frau     51
```

```
# Auswahl von Objekten über logische Ausdrücke
Auswahl = (MyData$Geschlecht=="Mann" & MyData$Alter < 19)
# zeige die ersten Einträge
head(Auswahl, 30)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [17] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
# Ausgabe der Auswahl: Alter, Alter des Vaters und der Mutter
MyData[Auswahl,                       # Objektauswahl
      c("Alter", "AlterM", "AlterV")]  # Welche Merkmale?
```

```
##     Alter AlterM AlterV
## 23     18     44     48
## 268    18     46     52
## 424    17     46     50
## 456    18     52     55
## 460    18     50     57
## 464    18     40     44
## 479    18     52     44
## 501    18     51     55
## 566    18     52     57
## 620    18     49     58
```

Daten kennenlernen

```
# Zeige die Männer, die mehr als 1300 Euro für Schuhe
# und Mobilfunk zusammen ausgegeben haben
MyData.Auswahl = MyData[MyData$Geschlecht=="Mann" &
                        MyData$AusgSchuhe + MyData$AusgKomm > 1300,
                        c("Alter", "Geschwister", "Farbe",
                          "AusgSchuhe", "AusgKomm")]
```


Daten kennenlernen

```
# ohne NAs
MyData.Auswahl = na.exclude(MyData.Auswahl)
MyData.Auswahl
```

```
##     Alter Geschwister   Farbe AusgSchuhe AusgKomm
## 42     24         1.0 schwarz       1000      600
## 81     25         2.0  silber        200     1900
## 121    22         0.0  silber        300     1100
## 142    20         2.0 schwarz        290     1570
## 161    19         1.0 schwarz        600      800
## 227    20         1.0 schwarz        200     1250
## 249    20         1.0    blau       1000      350
## 256    25         0.0 schwarz        280     1200
## 315    21         1.0   weiss        200     1300
## 353    20         0.0 schwarz        400      950
## 415    26         1.0    blau        600     1850
## 419    21         0.0 schwarz        200     1500
## 492    23         2.0   weiss        160     1800
## 493    26         2.0 schwarz        300     2000
## 494    20         2.0 schwarz        250     1500
## 535    20         2.0   weiss       2500     1500
## 548    26         2.0 schwarz        240     1200
## 562    24         1.0 schwarz         70     4668
## 573    21         1.0 schwarz        300     1200
## 581    19         2.0  silber        500      950
## 582    20         1.0 schwarz        500     1000
## 604    24         1.0 schwarz        150     1340
## 605    21         1.0  silber        600      800
## 615    25         4.5 schwarz       1200      600
## 646    22         1.0     rot        200     2500
## 647    23         1.0 schwarz        200     2000
```


Daten kennenlernen

```
# Neue Spalte Gesamtausgaben:
MyData.Auswahl$AusgGesamt = MyData.Auswahl$AusgKomm + MyData.Auswahl$AusgSchuhe
# sortiert nach Gesamtausgaben
MyData.Auswahl[order(MyData.Auswahl$AusgGesamt), ]
```

```
##     Alter Geschwister   Farbe AusgSchuhe AusgKomm AusgGesamt
## 249    20         1.0    blau       1000      350       1350
## 353    20         0.0 schwarz        400      950       1350
## 121    22         0.0  silber        300     1100       1400
## 161    19         1.0 schwarz        600      800       1400
## 605    21         1.0  silber        600      800       1400
## 548    26         2.0 schwarz        240     1200       1440
## 227    20         1.0 schwarz        200     1250       1450
## 581    19         2.0  silber        500      950       1450
## 256    25         0.0 schwarz        280     1200       1480
## 604    24         1.0 schwarz        150     1340       1490
## 315    21         1.0   weiss        200     1300       1500
## 573    21         1.0 schwarz        300     1200       1500
## 582    20         1.0 schwarz        500     1000       1500
## 42     24         1.0 schwarz       1000      600       1600
## 653    27         2.0 schwarz        700      950       1650
## 419    21         0.0 schwarz        200     1500       1700
## 494    20         2.0 schwarz        250     1500       1750
## 615    25         4.5 schwarz       1200      600       1800
## 142    20         2.0 schwarz        290     1570       1860
## 492    23         2.0   weiss        160     1800       1960
## 663    27         2.0 schwarz        200     1800       2000
## 81     25         2.0  silber        200     1900       2100
## 647    23         1.0 schwarz        200     2000       2200
## 493    26         2.0 schwarz        300     2000       2300
## 415    26         1.0    blau        600     1850       2450
```


Statistik: Table of Contents

Auswertungsmethoden für eindimensionales Datenmaterial

- Merkmal X wird an n Merkmalsträgern beobachtet ➠
  Urliste $(x_1, \ldots, x_n)$
  Im Beispiel: $x_1 = 4,\ x_2 = 11, \ldots, x_{12} = 6$
- Urlisten sind oft unübersichtlich, z.B.:

```
##  [1] 4 5 4 1 5 4 3 4 5 6 6 5 5 4 7 4 6 5 6 4 5 4 7 5 5 6 7 3
## [29] 7 6 6 7 4 5 4 7 7 5 5 5 5 6 6 4 5 2 5 4 7 5
```

- Dann zweckmäßig: Häufigkeitsverteilungen

Ausprägung (sortiert)	a_j	1	2	3	4	5	6	7	$\sum$
absolute Häufigkeit	$h(a_j) = h_j$	1	1	2	12	17	9	8	50
kumulierte abs. H.	$H(a_j) = \sum_{i=1}^{j} h(a_i)$	1	2	4	16	33	42	50	—
relative Häufigkeit	$f(a_j) = h(a_j)/n$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{12}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$	1
kumulierte rel. H.	$F(a_j) = \sum_{i=1}^{j} f(a_i)$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{4}{50}$	$\frac{16}{50}$	$\frac{33}{50}$	$\frac{42}{50}$	1	—

Beispiel: Alter von Studierenden in Statistik Vorlesung

	$h(a_j)$	$H(a_j) = \sum_{i=1}^{j} h(a_j)$	$f(a_j) = \frac{h(a_j)}{n}$	$F(a_j) = \sum_{i=1}^{j} f(a_i)$
17	1	1	0.0015	0.0015
18	37	38	0.0552	0.0567
19	113	151	0.1687	0.2254
20	113	264	0.1687	0.3940
21	94	358	0.1403	0.5343
22	71	429	0.1060	0.6403
23	67	496	0.1000	0.7403
24	49	545	0.0731	0.8134
25	24	569	0.0358	0.8493
26	25	594	0.0373	0.8866
27	19	613	0.0284	0.9149
28	20	633	0.0299	0.9448
29	11	644	0.0164	0.9612
30	5	649	0.0075	0.9687
31	5	654	0.0075	0.9761
32	7	661	0.0104	0.9866
33	2	663	0.0030	0.9896
34	3	666	0.0045	0.9940
35	2	668	0.0030	0.9970
36	2	670	0.0030	1.0000

Empirische Verteilungsfunktion

- für metrische Merkmale
- Anteil der Ausprägungen, die höchstens so hoch sind wie x.
- Exakt:

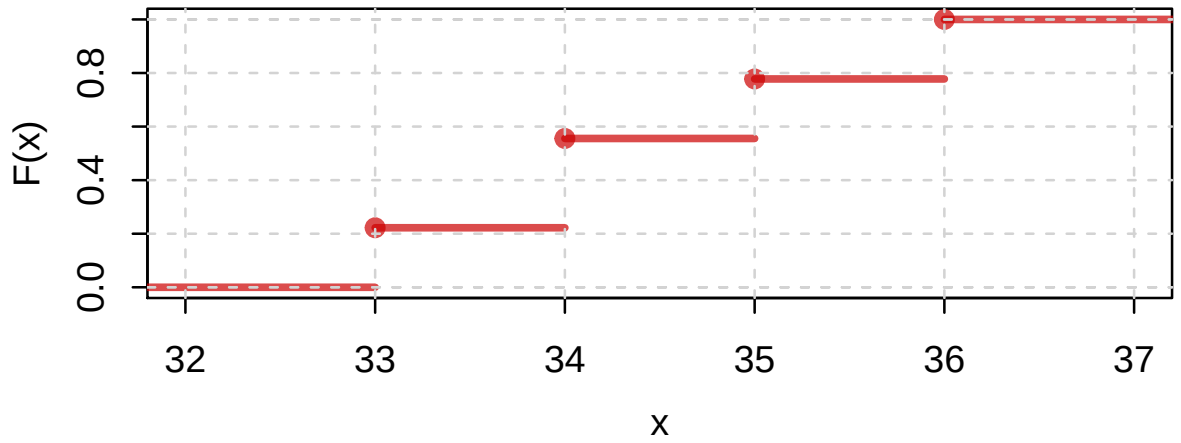
$$F(x) = \sum_{a_i \leqslant x} f(a_i)$$

Beispiel

```
Studenten.ueber.32 = sort(MyData$Alter[MyData$Alter > 32])
Studenten.ueber.32

## [1] 33 33 34 34 34 35 35 36 36

# empirical cumulative distribution function (ecdf)
Studenten.F = ecdf(Studenten.ueber.32)
plot(Studenten.F, col=rgb(0.8,0,0,.7), lwd=3, main="", xlab="x", ylab="F(x)")
grid(lty=2) # Gitternetz
```



Beispiel:

```
## [1] 33 33 34 34 34 35 35 36 36
```

(Urliste)

emp. Vtlgs. fkt:

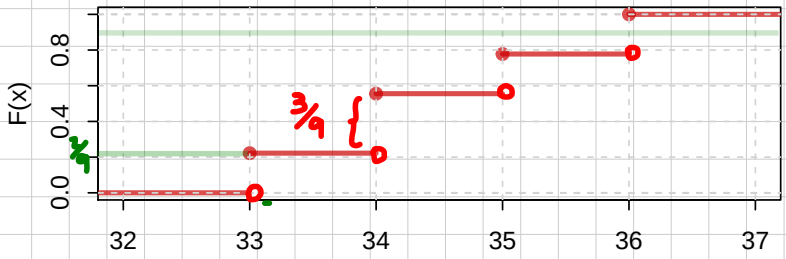
a_j	33	34	35	36
f_j	2/9	3/9	2/9	2/9

x	F(x)
-10	0
-9	0
:	0
0	0
0.2	0
32.999	0
33	2/9
33.5	2/9
33.99	2/9
34	5/9
35	7/9
35.9	7/9
36	1
36.2	1
100	1

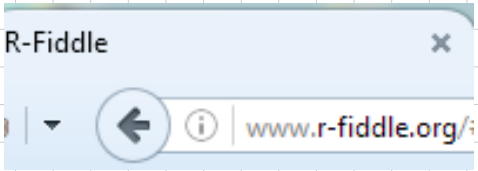
$\sum_{a_j \le -10} f(a_j)$

$\sum_{a_j \le 33} f(a_j) = f(33) = 2/9$

$\sum_{a_j \le 34} f(a_j) = f(33) + f(34) = 2/9 + 3/9 = 5/9$



R online (für iOs, android & Co.)



- für metrische Merkmale; Voraussetzung: sortierte Urliste
- Umkehrung der Verteilungsfunktion
- Anteil p gegeben, gesucht: $F^{-1}(p)$, falls vorhanden.
- Definition p-Quantil:

$$\tilde{x}_p = \begin{cases} \frac{1}{2}(x_{n\cdot p} + x_{n\cdot p+1}), & \text{wenn } n \cdot p \in \mathbb{N}_0 \\ x_{\lceil n\cdot p \rceil}, & \text{sonst} \end{cases}$$

$\lceil x \rceil$ obere Gaußklammer / Aufrundungsfunktion / ceiling

Beispiel

$\tilde{x}_{0.4} = x_{\lceil 9\cdot 0.4 \rceil} = x_{\lceil 3.6 \rceil} = x_4 = 34$ → 4. Wert der sortierten Urliste

↳ p : $n \cdot p = 9 \cdot 0.4 = 3.6 \notin \mathbb{N}_0$

$\tilde{x}_{3/9} = \frac{1}{2}\cdot(x_2 + x_3) = \frac{1}{2}(33+34) = 33.5$ ($n\cdot p+1 = 3$)

↳ $n \cdot p = 9 \cdot \frac{3}{9} = 2 \in \mathbb{N}_0$

```
## [1] 33 33 34 34 34 35 35 36 36

n = length(Studenten.ueber.32)
p = c(0.05, 2/n, 0.3, 0.5, 0.75, 0.9)

quantile(Studenten.ueber.32, probs=p, type=2)

##        5% 22.22222%       30%       50%       75%       90%
##      33.0       33.5      34.0      34.0      35.0      36.0
```

$\tilde{x}_p \hat{=}$ „Mind. p der Objekte hat eine m. Ausp. von höchstens $\tilde{x}_p$"

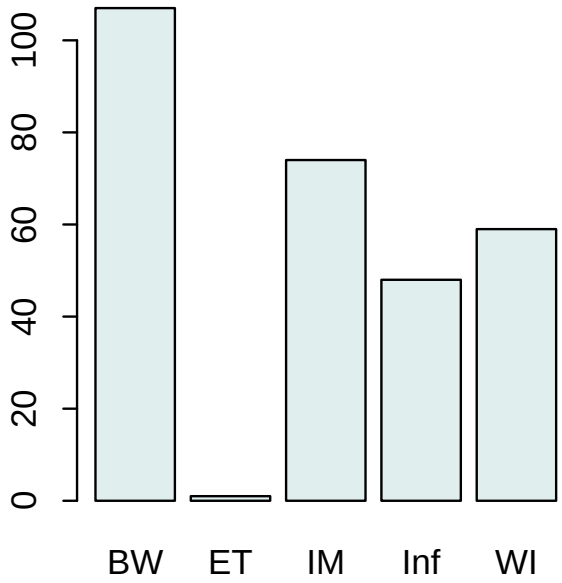
p: Anteil

❶ Balkendiagramm

```
M.t = table(MyData$Studiengang)
M.t

##
##  BW  ET  IM Inf  WI
## 107   1  74  48  59

barplot(M.t, col="azure2")
```



(Höhe proportional zu Häufigkeit)

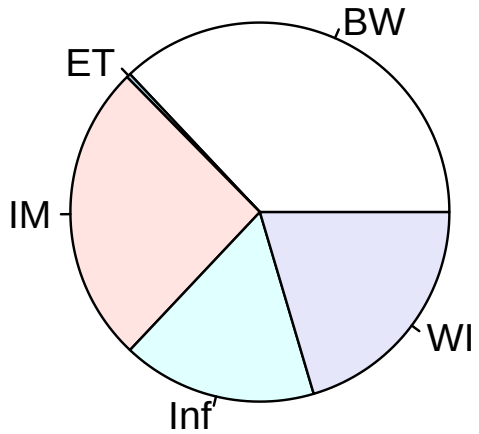
❷ Kreissektorendiagramm

Winkel: $w_j = 360° \cdot f(a_j)$

z.B. $w_{BW} = 360° \cdot \frac{107}{289} \approx 133.2°$

z.B. $w_{IM} = 360° \cdot \frac{74}{289} \approx 93.6°$

```
pie(M.t)
```

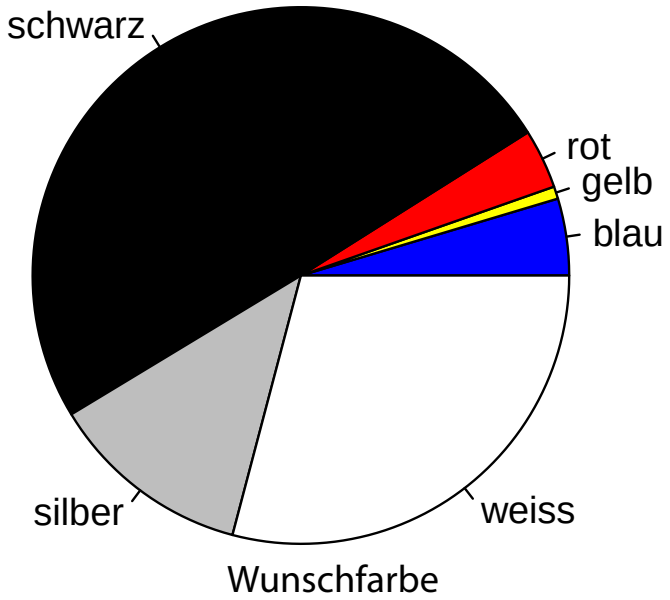


(Fläche proportional zu Häufigkeit)

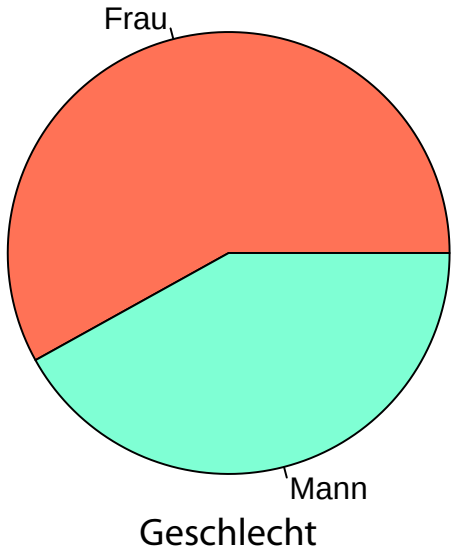
Umfrage: Kuchen

Kreissektorendiagramm

```
pie(table(MyData$Farbe),
    col=c("blue", "yellow", "red",
          "black", "grey", "white"))
```

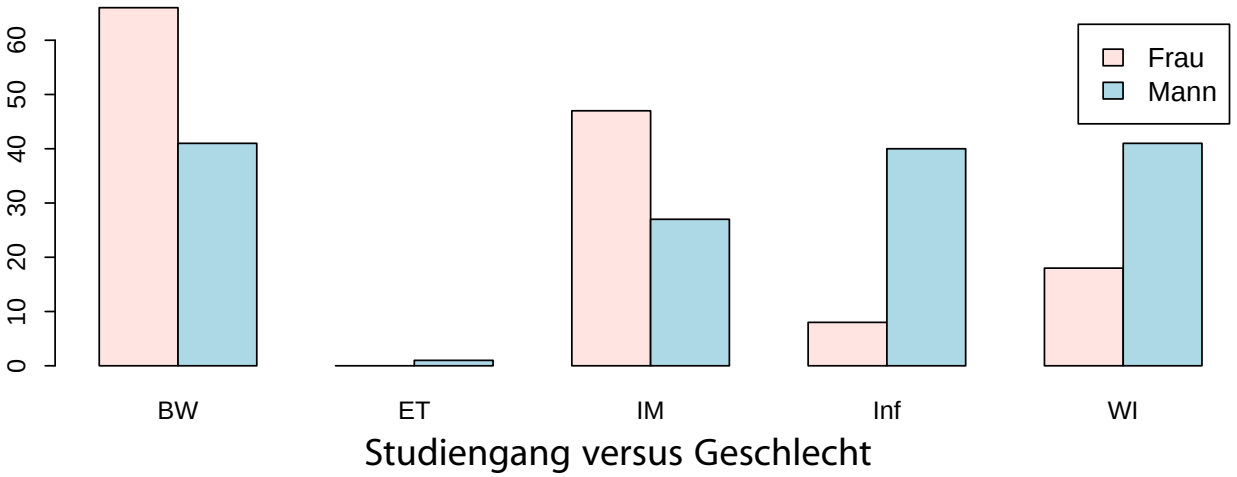



```
pie(table(MyData$Geschlecht),
    col=c("coral1", "aquamarine"))
```



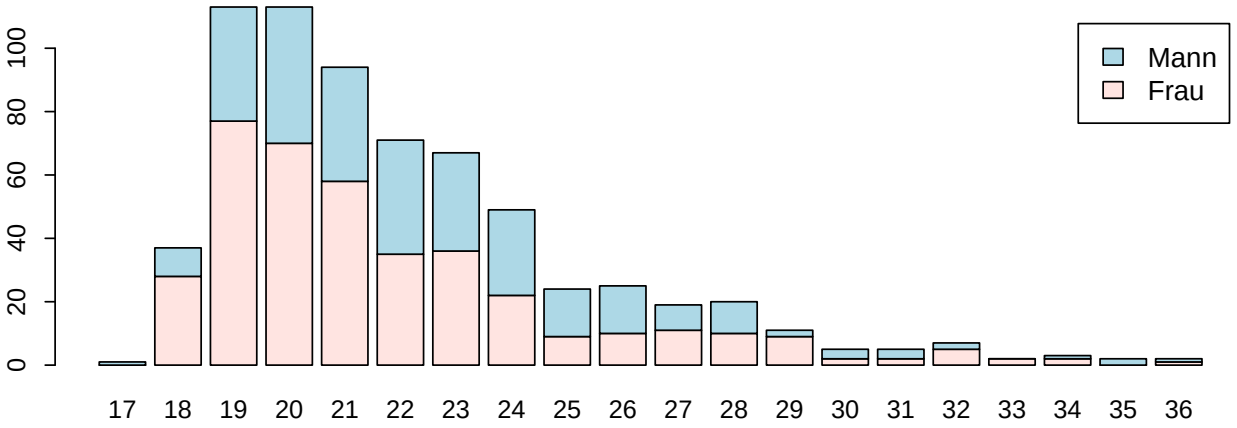
Balkendiagramm, Klassen getrennt oder gestapelt

```
barplot(xtabs(~ Geschlecht + Studiengang),
        legend=TRUE, beside=TRUE, col=c("mistyrose", "lightblue"))
```



Studiengang versus Geschlecht

```
barplot(xtabs(~ Geschlecht + Alter),
        legend=TRUE, beside=FALSE, col=c("mistyrose", "lightblue"))
```



Alter versus Geschlecht

❸ Histogramm

- für klassierte Daten
- Fläche proportional zu Häufigkeit:

frei wählbar, z.B. $c = 1$, $c = \frac{1}{n}$

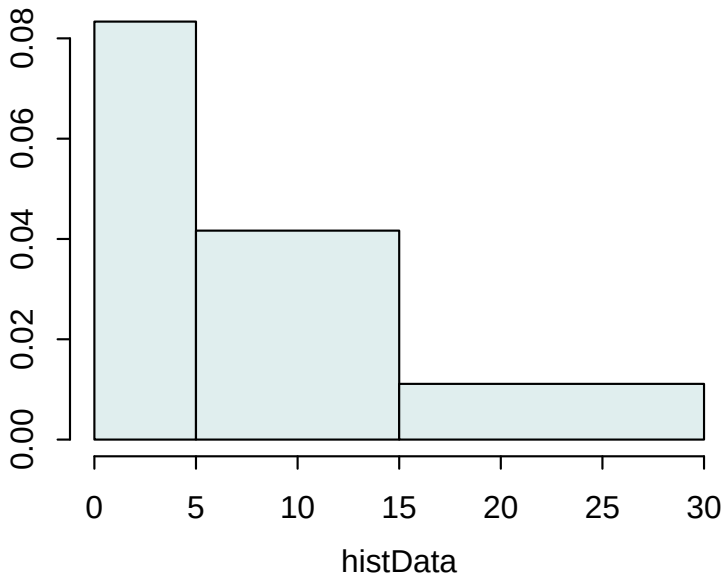
$$\text{Höhe}_j \cdot \text{Breite}_j = c \cdot h(a_j)$$

$$\Rightarrow \quad \text{Höhe}_j = c \cdot \frac{h(a_j)}{\text{Breite}_j}$$

- Im Beispiel mit $c = \frac{1}{12}$:

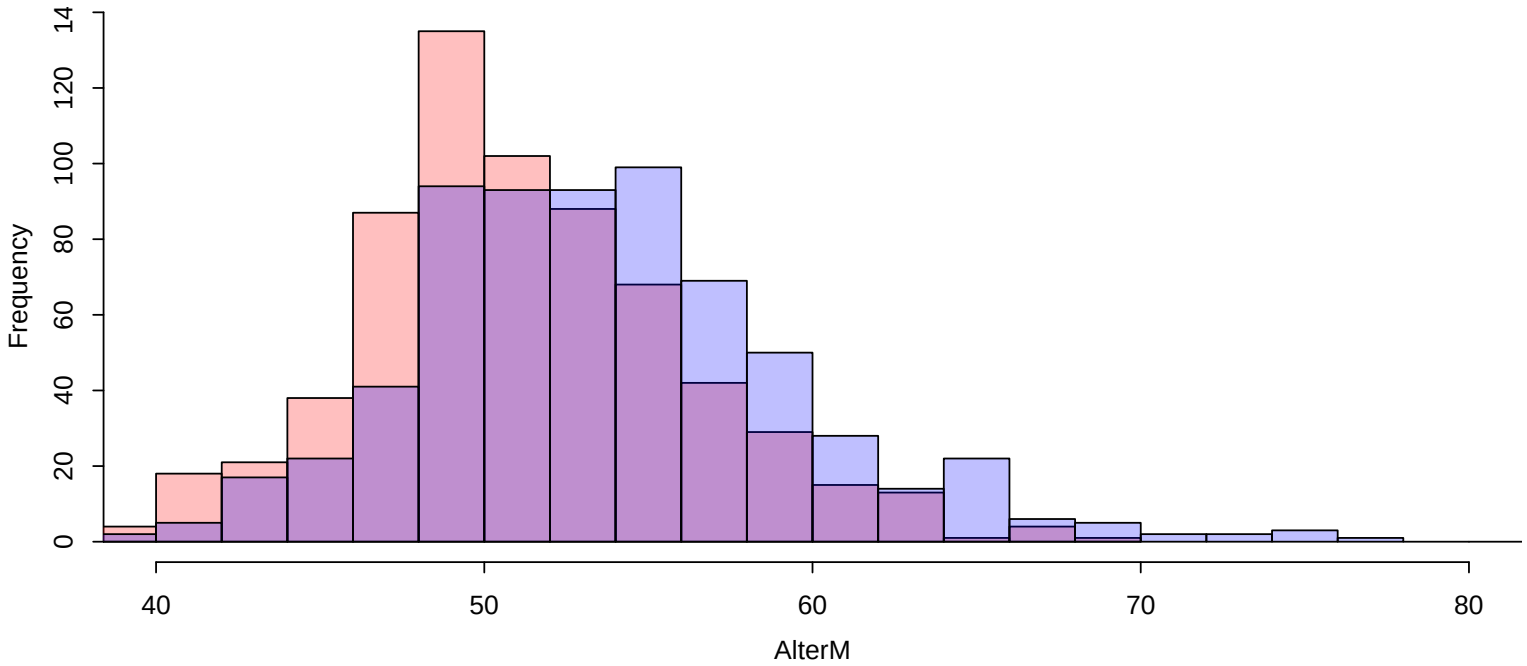
Klasse	$[0;5)$	$[5;15)$	$[15;30]$
$h(a_j)$	5	5	2
Breite_j	5	10	15
Höhe_j	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{90}$

```
require(MASS)
histData <- c(0,1,2,3,4,
              5,6,7,10,14,
              15,30)
truehist(histData,
         breaks=c(0, 4.999, 14.999, 30),
         col="azure2", ylab='')
```



Histogramm

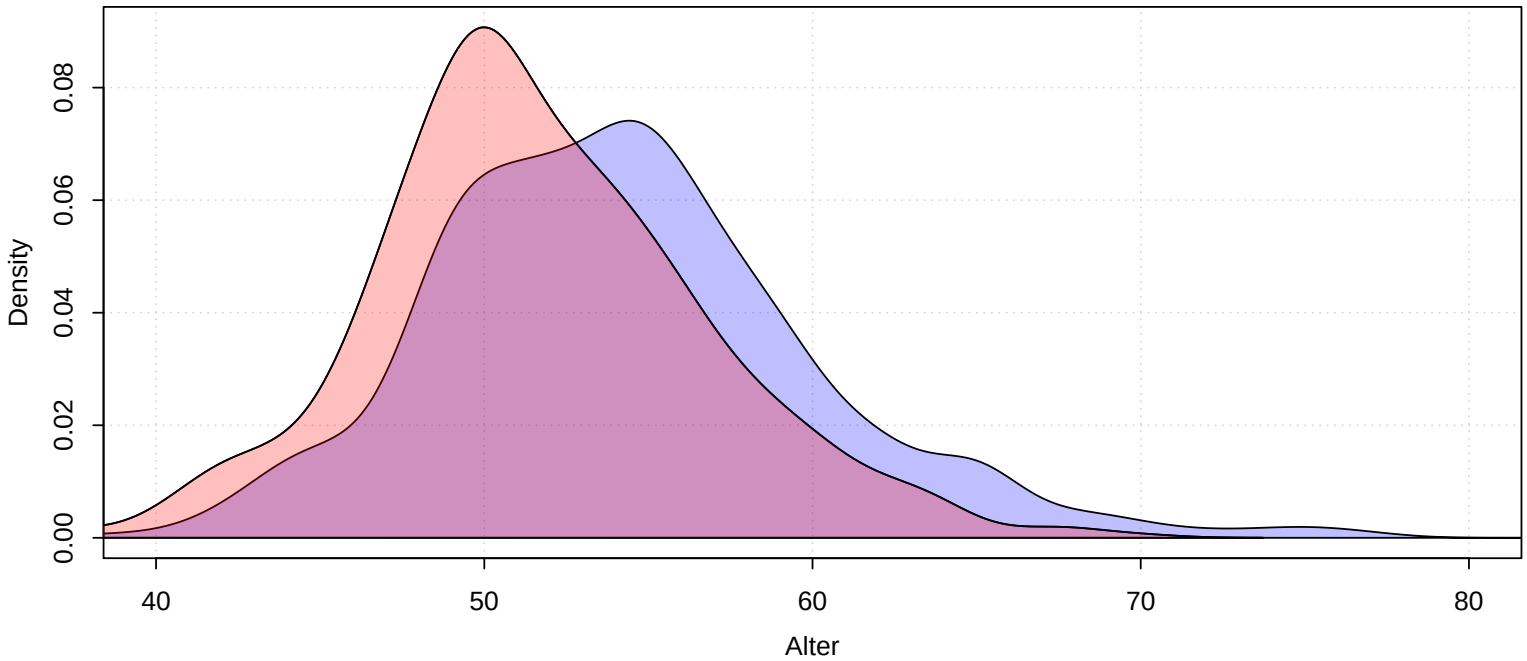
```
plot(hist(AlterM, plot=F, breaks=20),
          col=rgb(1,0,0,1/4), # make red transparent
          main="",
          xlim=c(40,80)) # draw from 40 to 80
plot(hist(AlterV, plot=F, breaks=20),
          col=rgb(0,0,1,1/4),
          add=TRUE)
```



Histogramm: Alter der Väter (blau) und Mütter (rosa)

Dichteplot

```
densMutter = density(na.exclude(AlterM))
densVater = density(na.exclude(AlterV))
plot(densMutter, main="", xlab="Alter",
xlim=c(40,80), # draw from 40 to 80
panel.first=grid()) # draw a grid
polygon(densVater, density=-1, col=rgb(0,0,1,1/4))
polygon(densMutter, density=-1, col=rgb(1,0,0,1/4))
```



Dichteplot: Alter der Väter (blau) und Mütter (rosa)

"Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?"

Modalwert

Modus x_{Mod}: häufigster Wert

Beispiel:

a_j	1	2	4
$h(a_j)$	4	3	1

$\Rightarrow x_{Mod} = 1$

Sinnvoll bei allen Skalenniveaus.

$= \tilde{x}_{0.5}$

Median x_{Med}: ‚mittlerer Wert', d.h.

1. Urliste aufsteigend sortieren: $x_1 \leqq x_2 \leqq \cdots \leqq x_n$
2. Dann

$$x_{Med} \begin{cases} = x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade} \\ \in [x_{\frac{n}{2}}; x_{\frac{n}{2}+1}], & \text{falls } n \text{ gerade (meist } x_{Med} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)) \end{cases}$$

Im Beispiel oben:
$1,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 4 \Rightarrow x_{Med} \in [1;2]$, z.B. $x_{Med} = 1{,}5$

Sinnvoll ab ordinalem Skalenniveau.

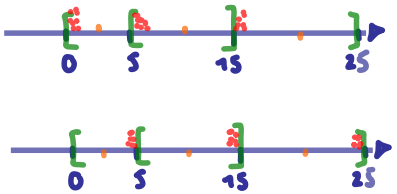
- Arithmetisches Mittel $\bar{x}$: Durchschnitt, d.h.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} a_j \cdot h(a_j) = \sum_{j=1}^{k} a_j \cdot f_j$$

Im Beispiel:

$$\bar{x} = \frac{1}{8} \cdot (\underbrace{1+1+1+1}_{1\cdot 4} + \underbrace{2+2+2}_{2\cdot 3} + \underbrace{4}_{4\cdot 1}) = 1{,}75$$

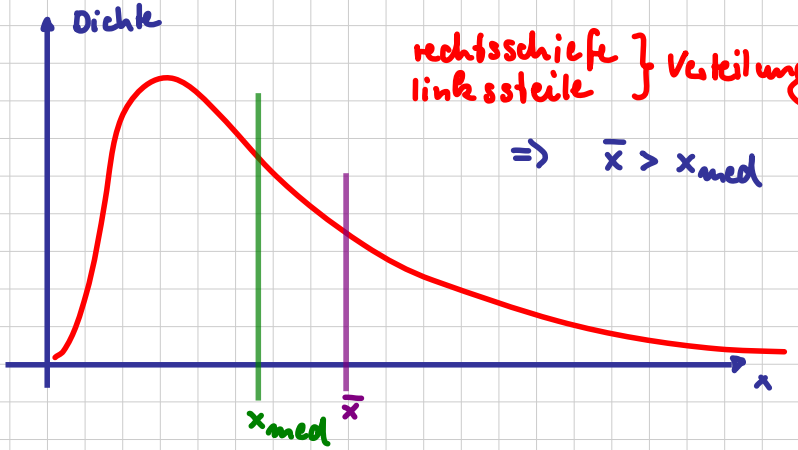
Sinnvoll nur bei kardinalem Skalenniveau.
Bei klassierten Daten: (nur Schätzwert)



$$\bar{x}^* = \frac{1}{n}\sum \text{Klassenmitte} \cdot \text{Klassenhäufigkeit}$$

Im Beispiel:

$$\bar{x}^* = \frac{1}{12} \cdot (2{,}5 \cdot 5 + 10 \cdot 5 + 22{,}5 \cdot 2) = 8{,}96 \neq 7{,}5 = \bar{x}$$



rechtsschiefe linkssteile } Verteilung

$\Rightarrow \quad \bar{x} > x_{med}$

Lageparameter

Ausgaben für Schuhe

```
median(na.exclude(AusgSchuhe))

## [1] 200

mean(na.exclude(AusgSchuhe))

## [1] 270.4529
```

Alter

```
median(Alter)

## [1] 21

mean(Alter)

## [1] 22.12537
```

~~Lieblingsfarbe~~

```
summary(Geschlecht)

## Frau Mann
##  389  281
```

↳ Modus

Alter der Mutter

```
median(na.exclude(AlterM))

## [1] 51

mean(na.exclude(AlterM))

## [1] 51.63677
```

Streuungsparameter

- Voraussetzung: kardinale Werte $x_1, \ldots, x_n$
- **Beispiel:**

a)	x_i	1950	2000	2050	} je $\bar{x} = 2000$
b)	x_i	0	0	6000	}

- Spannweite: $SP = \max_i x_i - \min_i x_i$

  Im Beispiel:

a) $SP = 2050 - 1950 = 100$
b) $SP = 6000 - 0 = 6000$

- Mittlere quadratische Abweichung:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}_{\text{Verschiebungssatz}}$$

$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum(x_i) - \frac{1}{n}\sum_{i=1}^{n}\bar{x} = \bar{x} - \frac{1}{n}\cdot n\cdot\bar{x} = 0$

$= \frac{1}{n}\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$

$= \frac{1}{n}\sum x_i^2 - \frac{1}{n}\cdot 2\cdot\bar{x}\sum x_i + \frac{1}{n}\sum\bar{x}^2 = \frac{1}{n}\sum x_i^2 - 2\bar{x}\cdot\bar{x} + \frac{1}{n}\cdot n\cdot\bar{x}^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$

Streuungsparameter (2)

- **Mittlere quadratische Abweichung** im Beispiel:

  a) $s^2 = \frac{1}{3} \cdot (50^2 + 0^2 + 50^2)$

  $= \frac{1}{3} \cdot (1950^2 + 2000^2 + 2050^2) - 2000^2 = 1666{,}67$

  b) $s^2 = \frac{1}{3} \cdot (2000^2 + 2000^2 + 4000^2)$

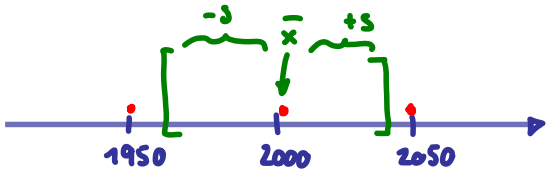
  $= \frac{1}{3} \cdot (0^2 + 0^2 + 6000^2) - 2000^2 \qquad = 8000000$

- **Standardabweichung:** $s = \sqrt{s^2}$

  Im Beispiel:

  a) $s = \sqrt{1666{,}67} = 40{,}82$

  b) $s = \sqrt{8000000} = 2828{,}43$



- **Variationskoeffizient:** $V = \frac{s}{\bar{x}}$ (maßstabsunabhängig)

  Im Beispiel:

  a) $V = \frac{40{,}82}{2000} = 0{,}02\ (\widehat{=} 2\,\%)$

  b) $V = \frac{2828{,}43}{2000} = 1{,}41\ (\widehat{=} 141\,\%)$

Univariate Statistik mit TR

Mode → STAT → 1-Var

... Dateneingabe ...

1950
2000
2050

AC

Shift → STAT → Var → $\bar{x}$ („2000")
σ_x („40.82")
(σ_{x_n})

Min Max → Median
$\tilde{x}_{0,75}$
$\tilde{x}_{0.25}$

Lage und Streuung: Überblick

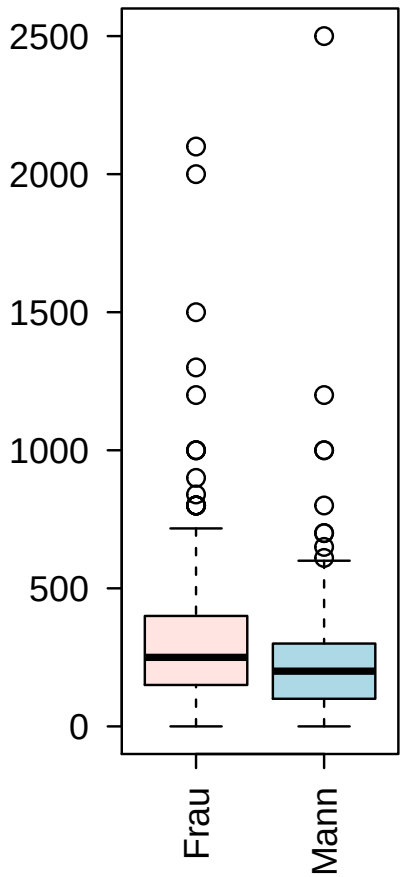

```
LageStreuung = function(x) {
  x=na.omit(x)  # ignoriere fehlende Werte
  n = length(x) # Anzahl nicht fehlender Werte
  popV = var(x)*(n-1)/n # var() ist nicht mittl. qu. Abweichung
  return(list(mean=mean(x),
              median=median(x),
              Variance=popV,
              StdDev=sqrt(popV),
              VarCoeff=sqrt(popV)/mean(x)))
}
mat1 = sapply(MyData[c("Alter","AlterV","AlterM", # sapply: pro Spalte anwenden
                       "Geschwister", "AnzSchuhe", "AusgSchuhe")],
              LageStreuung)
```

	Alter	AlterV	AlterM	Geschwister	AnzSchuhe	AusgSchuhe
mean	22.13	54.28	51.64	1.51	21.22	270.45
median	21.00	54.00	51.00	1.00	16.00	200.00
Variance	11.36	35.35	25.74	1.18	415.51	56333.39
StdDev	3.37	5.95	5.07	1.08	20.38	237.35
VarCoeff	0.15	0.11	0.10	0.72	0.96	0.88

Lage und Streuung als Grafik: Boxplot

- Graphische Darstellung von Lage und Streuung
- Box: Oberer/Unterer Rand: 3. bzw. 1. Quartil ($\tilde{x}_{0,75}$ bzw. $\tilde{x}_{0,25}$),
- Linie in Mitte: Median
- Whiskers: Länge: Max./Min Wert, aber beschränkt durch das 1,5-fache des Quartilsabstands (falls größter/kleinster Wert größeren/kleineren Abstand von Box: Länge Whiskers durch größten/kleinsten Wert innerhalb dieser Schranken)
- Ausreißer: Alle Objekte außerhalb der Whisker-Grenzen

```
boxplot(AusgSchuhe ~ Geschlecht,
        col=c("mistyrose", "lightblue"),
        data=MyData, main="", las=2)
```



Ausgaben für Schuhe


```
summary(MyData)
```

```
##     Jahrgang        Alter          Groesse      Geschlecht     AlterV          AlterM
##  Min.   :2014   Min.   :17.00   Min.   :150.0   Frau:389   Min.   :38.00   Min.   :37.00
##  1st Qu.:2014   1st Qu.:20.00   1st Qu.:166.0   Mann:281   1st Qu.:50.00   1st Qu.:48.00
##  Median :2015   Median :21.00   Median :172.0              Median :54.00   Median :51.00
##  Mean   :2015   Mean   :22.13   Mean   :173.1              Mean   :54.28   Mean   :51.64
##  3rd Qu.:2016   3rd Qu.:24.00   3rd Qu.:180.0              3rd Qu.:57.00   3rd Qu.:55.00
##  Max.   :2016   Max.   :36.00   Max.   :198.0              Max.   :87.00   Max.   :70.00
##                                                            NA's   :1       NA's   :1
##     GroesseV        GroesseM      Geschwister       Farbe        AusgKomm        AnzSchuhe
##  Min.   :160.0   Min.   : 76.0   Min.   :0.000   blau   : 31   Min.   :   0.0   Min.   :  2.00
##  1st Qu.:175.0   1st Qu.:162.0   1st Qu.:1.000   gelb   :  5   1st Qu.: 207.5   1st Qu.:  8.00
##  Median :180.0   Median :165.0   Median :1.000   rot    : 24   Median : 360.0   Median : 16.00
##  Mean   :179.1   Mean   :166.2   Mean   :1.509   schwarz:333   Mean   : 458.1   Mean   : 21.22
##  3rd Qu.:183.0   3rd Qu.:170.0   3rd Qu.:2.000   silber : 82   3rd Qu.: 600.0   3rd Qu.: 30.00
##  Max.   :204.0   Max.   :192.0   Max.   :9.000   weiss  :195   Max.   :4668.0   Max.   :275.00
##  NA's   :11      NA's   :8                                     NA's   :2
##    AusgSchuhe         Essgewohnheiten Raucher      NoteMathe              MatheZufr   Studiengang
##  Min.   :   0.0   carnivor    :420    ja  : 81   Min.   :1.000   unzufrieden   :185   BW  :107
##  1st Qu.: 100.0   fruktarisch :  1    nein:381   1st Qu.:2.650   geht so       :151   ET  :  1
##  Median : 200.0   pescetarisch: 26    NA's:208   Median :3.300   zufrieden     :114   IM  : 74
##  Mean   : 270.5   vegan       :  3               Mean   :3.233   sehr zufrieden: 74   Inf : 48
##  3rd Qu.: 350.0   vegetarisch : 15               3rd Qu.:4.000   NA's          :146   WI  : 59
##  Max.   :2500.0   NA's        :205               Max.   :5.000                        NA's:381
##  NA's   :1                                       NA's   :162
```

Boxplots

```
for(attribute in c("Alter", "AlterV", "AlterM", "Geschwister",
                   "AusgSchuhe", "AusgKomm")) {
  data=MyData[, attribute]
  boxplot(data, # all rows, column of attribute
          col="lightblue",     # fill color
          lwd=3,               # line width
          cex=2,                # character size
          oma=c(1,1,2,1)
          )
  text(0.7,max(data), attribute, srt=90, adj=1)
}
```