Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2017

HSA Statistik SS 2017 Sessionlist		
Datum	Statistik für BW/IM/I/Winf	Nr.
15.03.2017	Einführung Statistik	1
22.03.2017	Differentialrechnung, 2-dim Diff.Rechnung	2
29.03.2017	univ. deskr. Stat., Quantile, Plots	3
05.04.2017	Streuung, Konzentrationsmaße	4
12.04.2017	Kontingenztabellen, Mosaikplots, Korrelation	5
19.04.2017	Preisindizes, lineare Regression	6
26.04.2017	Wahrscheinlichkeitsbegriff	7
03.05.2017	Bedingte Wahrscheinlichkeit, Bayes	8
10.05.2017	diskrete Zufallsvariablen	9
17.05.2017	Stetige ZV, Gleichverteilung	10
24.05.2017	Pyramid	
31.05.2017	Normalverteilung, Verteilungsparameter	11
07.06.2017	Schätzfunktionen und Punktschätzer	12
14.06.2017	Konfidenzintervalle	13
21.06.2017	Wiederholung, Besprechung Probeklausur	14
28.06.2017	Prüfungswoche	15

Stefan Etschberger

Gliederung

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

	$h(a_j)$	$H(a_j) = \sum_{i=1}^{j} h(a_j)$	$f(a_j) = \frac{h(a_j)}{n}$	$F(a_j) = \sum_{i=1}^{j} f(a_i)$
2	1	1	0.0011	0.0011
17	1	2	0.0011	0.0021
18	50	52	0.0532	0.0554
19	163	215	0.1736	0.2290
20	156	371	0.1661	0.3951
21	137	508	0.1459	0.5410
22	108	616	0.1150	0.6560
23	93	709	0.0990	0.7551
24	65	774	0.0692	0.8243
25	39	813	0.0415	0.8658
26	34	847	0.0362	0.9020
27	23	870	0.0245	0.9265
28	22	892	0.0234	0.9499
29	13	905	0.0138	0.9638
30	8	913	0.0085	0.9723
31	7	920	0.0075	0.9798
32	8	928	0.0085	0.9883
33	2	930	0.0021	0.9904
34	3	933	0.0032	0.9936
35	2	935	0.0021	0.9957
36	2	937	0.0021	0.9979
37	2	939	0.0021	1.0000

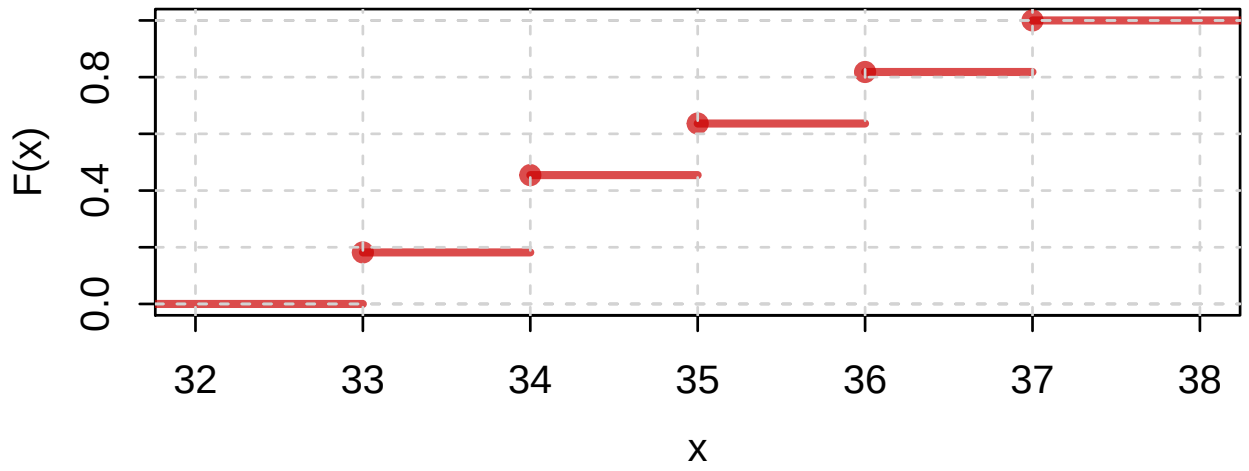
Empirische Verteilungsfunktion

- 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 37 37

# 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
```



```
Studenten.ueber.32 = sort(MyData$Alter[MyData$Alter > 32])
Studenten.ueber.32
##  j  1  2  3  4  5  6  7  8  9 10 11
## aj 33 33 34 34 34 35 35 36 36 37 37
```

n = 11

Empirische Verteilungsfunktion:

$$F(x) = \sum_{a_j \le x} f(a_j)$$

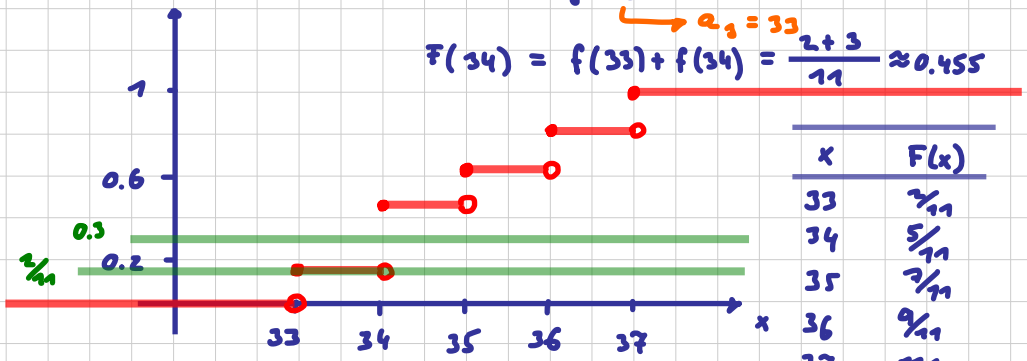
$$F(5) = \sum_{a_j \le 5} f(a_j) = 0$$ → gibt's nicht

$$F(32.5) = \sum_{a_j \le 32.5} f(a_j)$$

$F(32.999\ldots) = 0$, $F(33) = \sum_{a_j \le 33} f(a_j) = f(33) = \frac{2}{11} \approx 0.18$

$a_j = 33$

$F(34) = f(33) + f(34) = \frac{2+3}{11} \approx 0.455$



x	F(x)
33	2/11
34	5/11
35	7/11
36	9/11
37	11/11 = 1

Empirisches Quantil

Beispiel:

```
##  [1] 33 33 34 34 34 35 35 36 36 37 37
n = length(Studenten.ueber.32)
p = c(0.05, 2/n, 0.3, 0.5, 0.75, 0.9)
```

4. Wert

gesucht: empirischer 30% - Quantil

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

"Gaußklammer" „ceiling" aufrunden hier: $\lceil 3.3 \rceil = 4$

$\tilde{x}_{0.3} = x_{\lceil 11 \cdot 0.3 \rceil} = x_{\lceil 3.3 \rceil} = x_4 = 34$

4. Wert

$p = 0.3$, $n = 11$
$\Rightarrow n \cdot p = 3.3$ (ist keine natürliche Zahl, $\notin \mathbb{N}_0$)

$\tilde{x}_{2/11} = \frac{1}{2}(x_2 + x_{2+1}) = \frac{1}{2}(x_2 + x_3) = \frac{1}{2}(33+34)$
$= 33.5$

$n \cdot p = 11 \cdot \frac{2}{11} = 2 \in \mathbb{N}$

Interpretation:

$\tilde{x}_{0.3} = 34$: mind. 30% der Leute sind höchstens 34 Jahre alt

$\tilde{x}_{2/11} \approx \tilde{x}_{0.18} = 33.5$: mind 18% der Leute sind höchstens 33.5 J. alt

Empirische Quantile

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

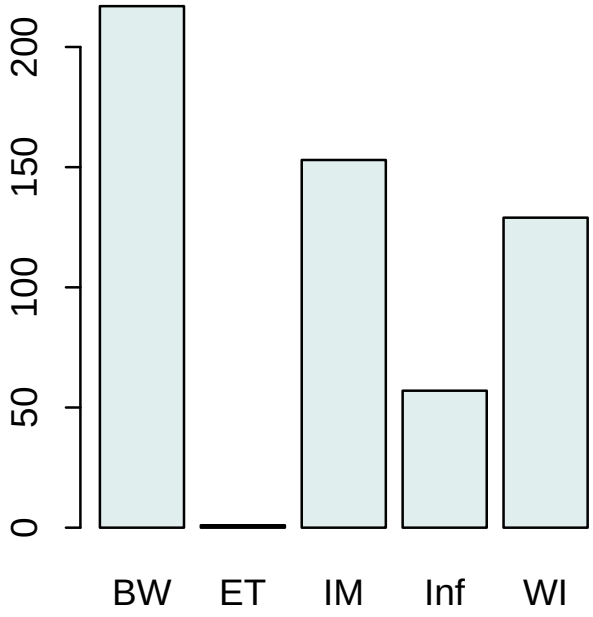
$$\tilde{x}_p = \begin{cases} \frac{1}{2}(x_{n\cdot p} + x_{n\cdot p+1}), & \text{wenn } n \cdot p \in \mathbb{N}_0 \\ x_{\lceil n\cdot p \rceil}, & \text{sonst} \end{cases}$$

Beispiel

```
##  [1] 33 33 34 34 34 35 35 36 36 37 37

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% 18.18182%       30%       50%       75%       90%
##      33.0       33.5      34.0      35.0      36.0      37.0
```

❶ Balkendiagramm

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

##
##  BW  ET  IM Inf  WI
## 217   1 153  57 129

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



(Höhe proportional zu Häufigkeit)

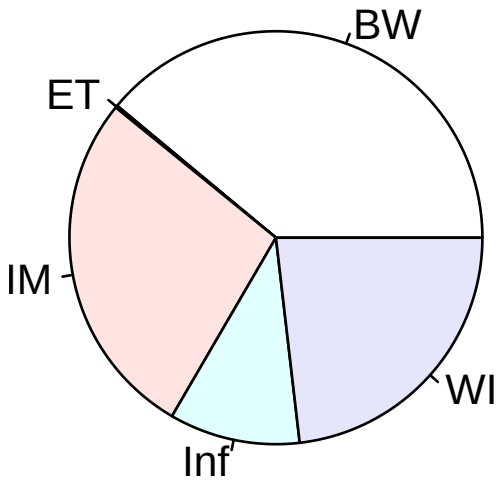
❷ Kreissektorendiagramm

Winkel: $w_j = 360° \cdot f(a_j)$

z.B. $w_{BW} = 360° \cdot \frac{217}{557} \approx 140.4°$

z.B. $w_{IM} = 360° \cdot \frac{153}{557} \approx 97.2°$

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

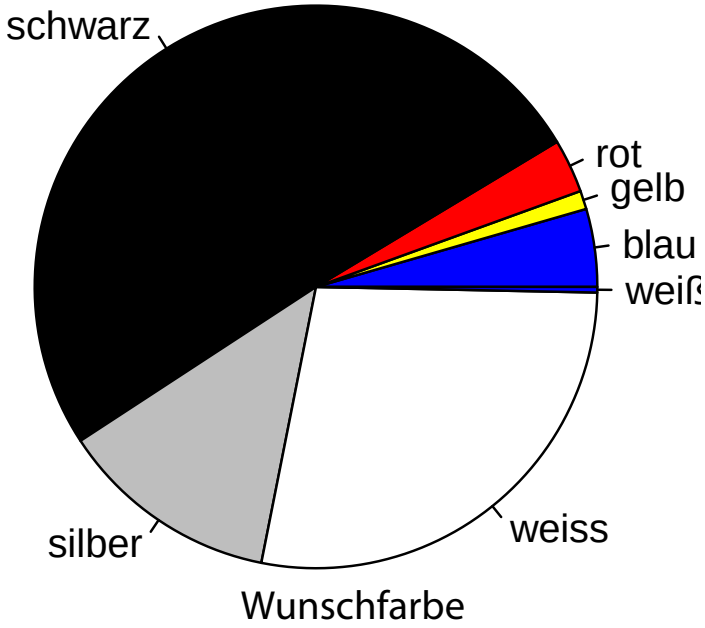


(Fläche proportional zu Häufigkeit)

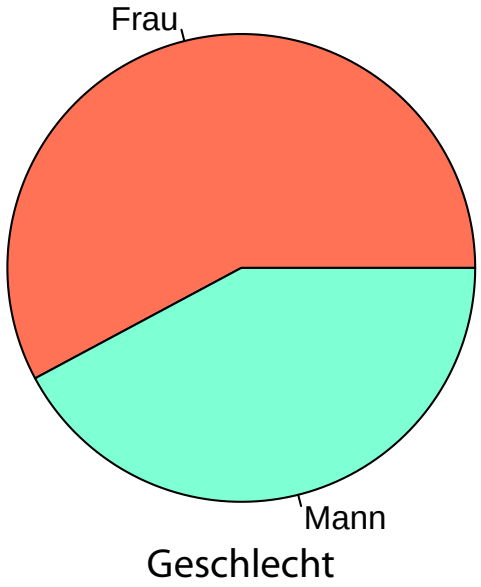
Umfrage: Kuchen

Kreissektorendiagramm

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

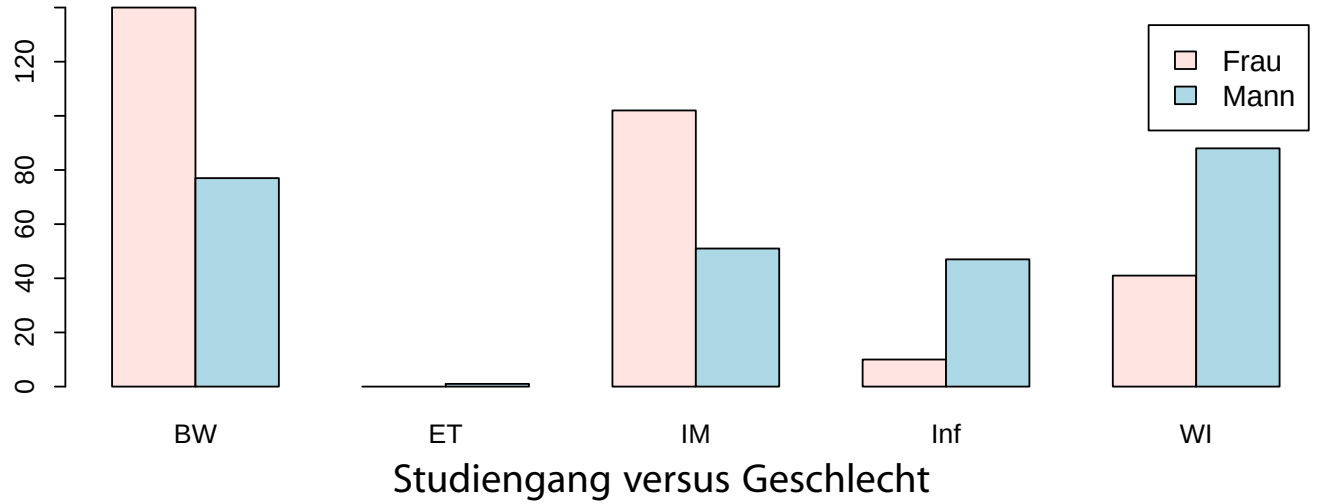



```
pie(table(MyData$Geschlecht),
    col=c("coral1", "aquamarine"))
```



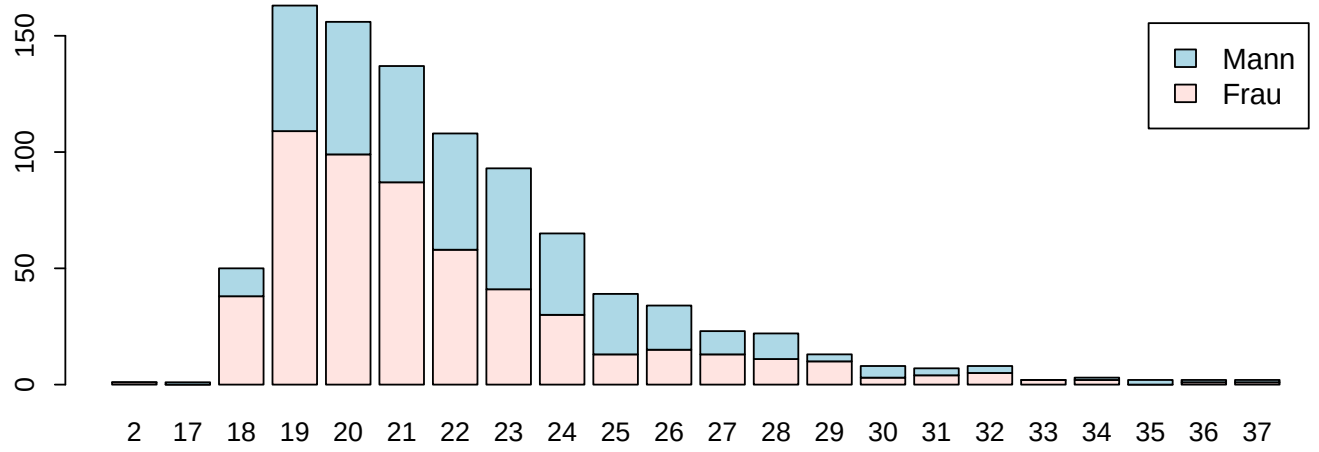
Balkendiagramm, Klassen getrennt oder gestapelt

```
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: oft $c=1$ oder $c=1/n$

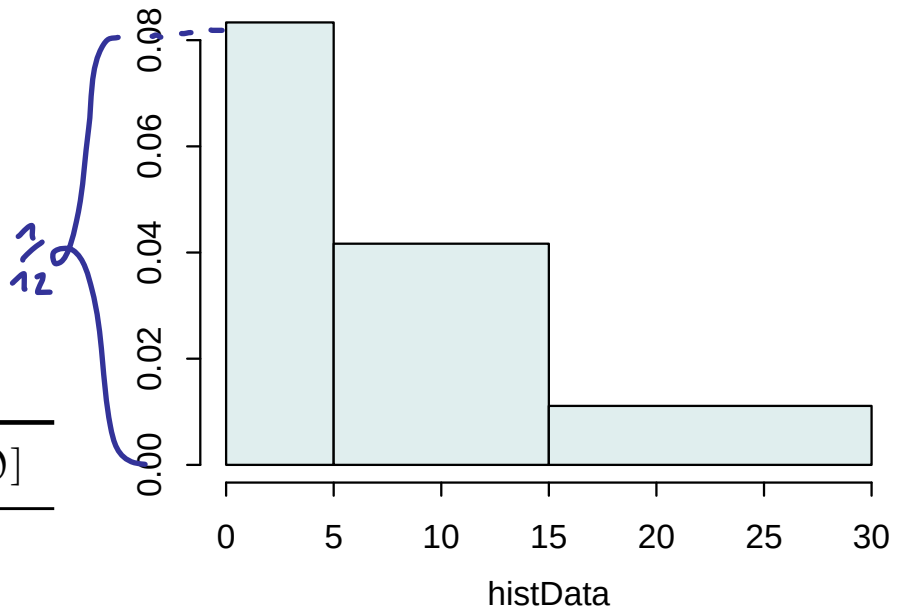
$$\text{Höhe}_j \cdot \text{Breite}_j = c \cdot h(a_j)$$

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

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

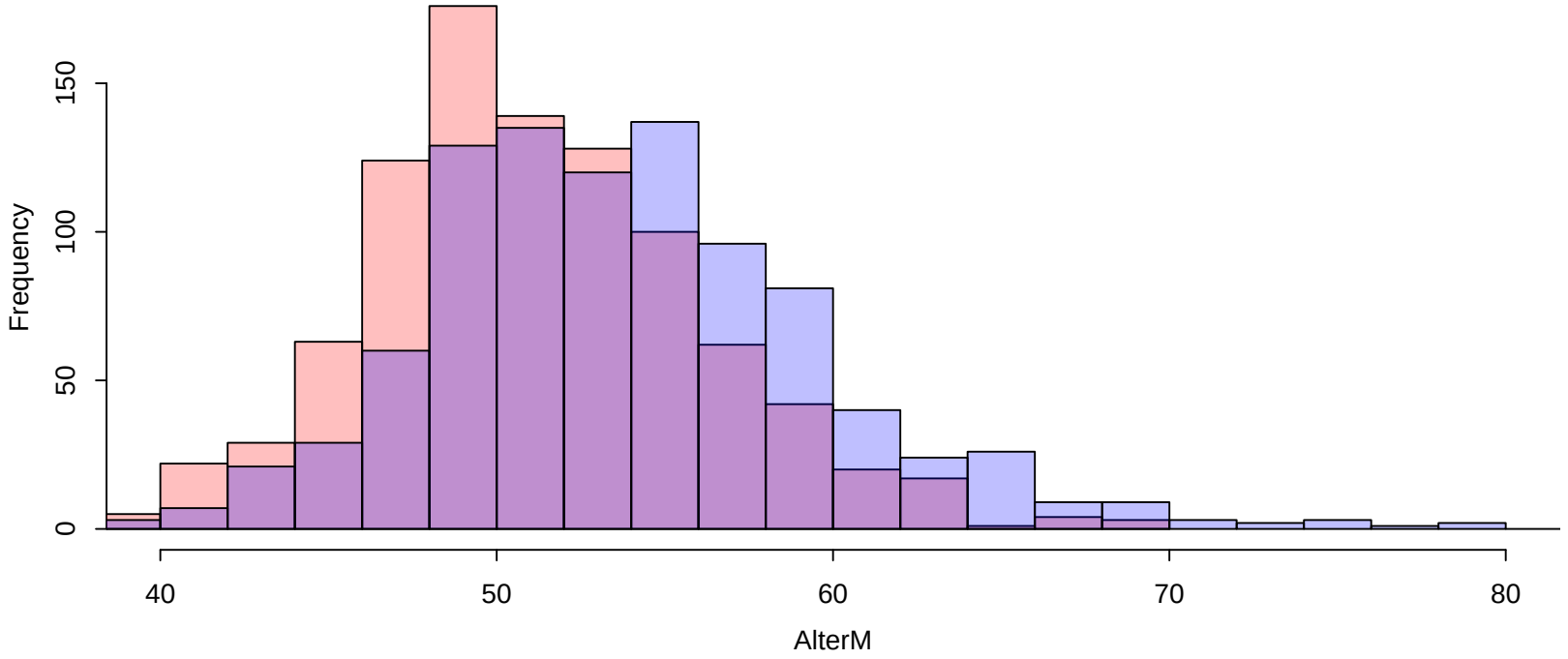
Klasse	$[0;5)$	$[5;15)$	$[15;30]$
$h(a_j)$	5	5	2
Breite_j	5	10	15
Höhe_j	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{90}$

```
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

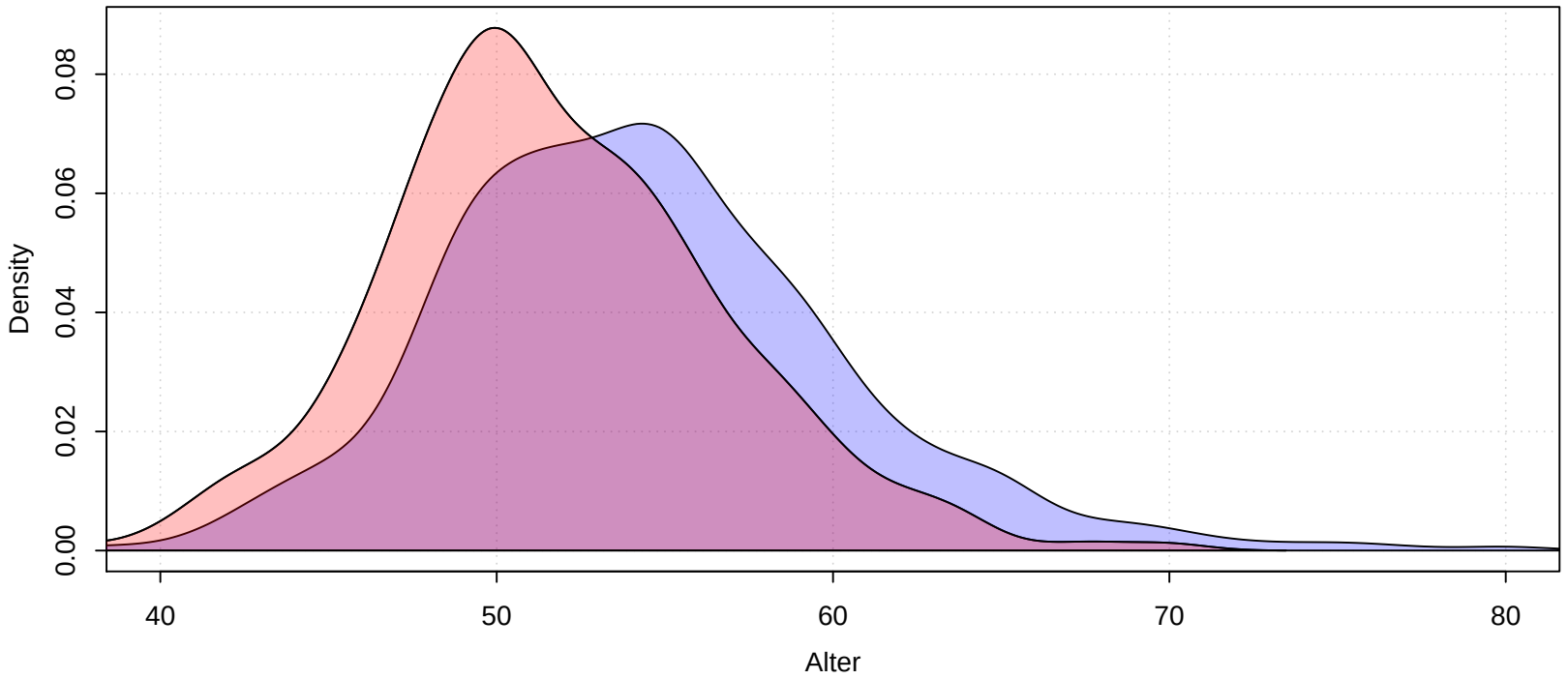
```
plot(hist(AlterM, plot=F, breaks=20),
          col=rgb(1,0,0,1/4), # make red transparent
          main="",
          xlim=c(40,80)) # draw from 40 to 80
plot(hist(AlterV, plot=F, breaks=20),
          col=rgb(0,0,1,1/4),
          add=TRUE)
```



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?"

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.

Median $\tilde{x}_{0.5}$ (empirisches 50 %-Quantil) 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.

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

Im Beispiel:

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

Sinnvoll nur bei kardinalem Skalenniveau.
Bei klassierten Daten:

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

Im Beispiel:

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

Beispiel: Urliste: $\underbrace{2000, \ldots, 2000}_{20 \text{ mal}}, 86000$ (x_1, x_{20}, x_{21})

$$x_{med} = \tilde{x}_{0.5} = x_{\lceil 10.5 \rceil} = x_{11} = 2000$$

↳ $n \cdot p = 21 \cdot 0.5 \notin \mathbb{N}$
$= 10.5$

Lageparameter

Ausgaben für Schuhe

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

## [1] 200

mean(na.exclude(AusgSchuhe))

## [1] 278.6876
```

Alter

```
median(Alter)

## [1] 21

mean(Alter)

## [1] 22
```

Lieblingsfarbe

```
summary(Geschlecht)

## Frau Mann
##  543  396
```

Alter der Mutter

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

## [1] 51

mean(na.exclude(AlterM))

## [1] 51.68763
```