Statistik Workshop

Mini-Einführung und Auffrischung zu einigen Teilen der angewandten Statistik

12. und 14. Januar 2015

Prof. Dr. Stefan Etschberger
HSA

Outline

Material zur Vorlesung

Kursmaterial:

- Handout der Folien
- Alle Folien inklusive Anmerkungen (am Abend)
- Beispieldaten
- Alle Auswertungen als **R**-Datei

Literatur:

Bamberg, Günter, Franz Baur und Michael Krapp (2011). **Statistik**. 16. Aufl. München: Oldenbourg Verlag. ISBN: 3486702580.

Dalgaard, Peter (2002). **Introductory Statistics with R**. New York: Springer.

Fahrmeir, Ludwig, Rita Künstler, Iris Pigeot und Gerhard Tutz (2009). **Statistik: Der Weg zur Datenanalyse**. 7. Aufl. Berlin, Heidelberg: Springer. ISBN: 3642019382.

Fragebogen

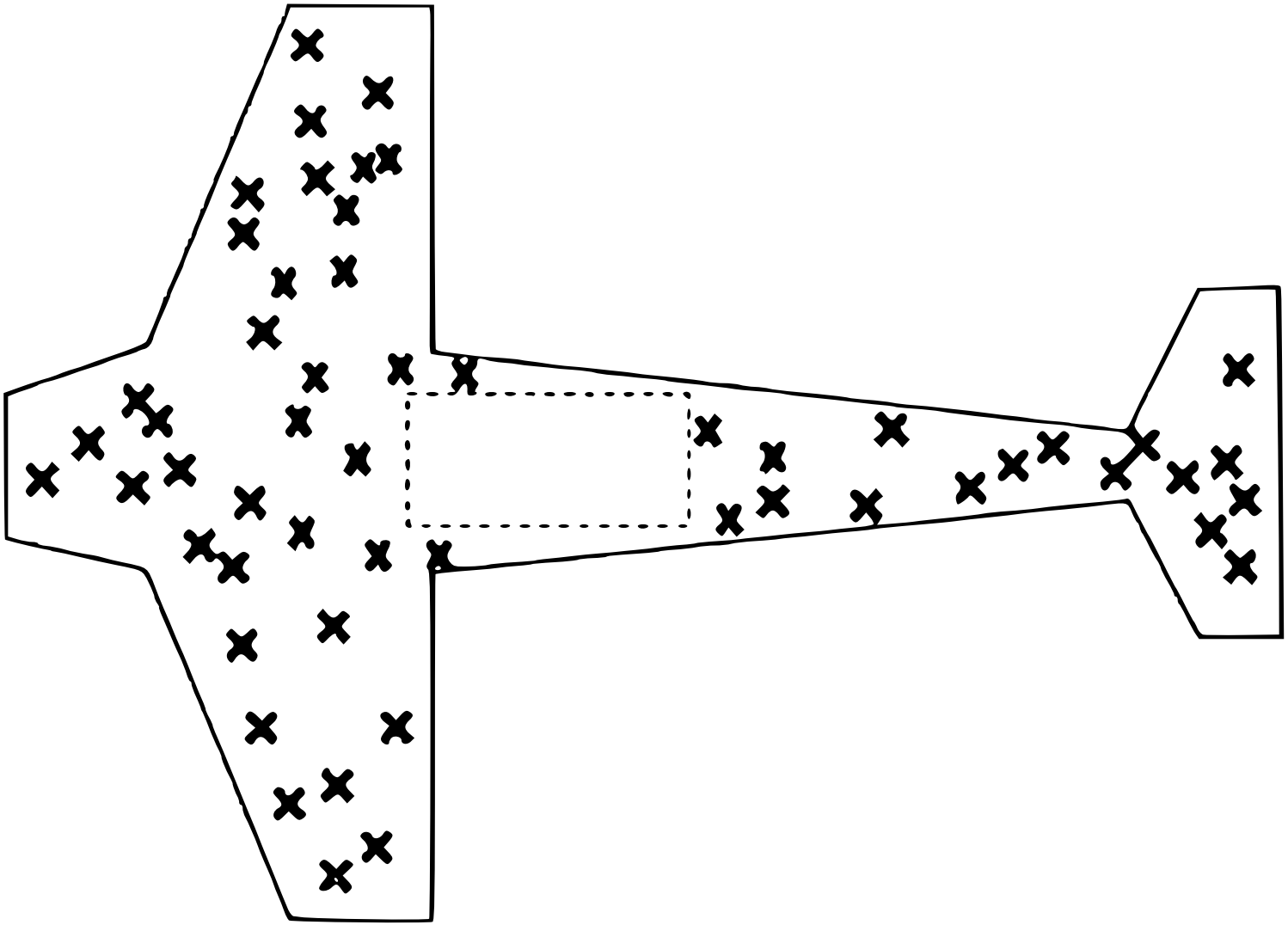
Umfrage Statistik Sommersemester 2014

Bitte beantworten Sie folgende Fragen vollständig und füllen Sie jeweils *nur eine* beliebige Spalte *leserlich* aus.

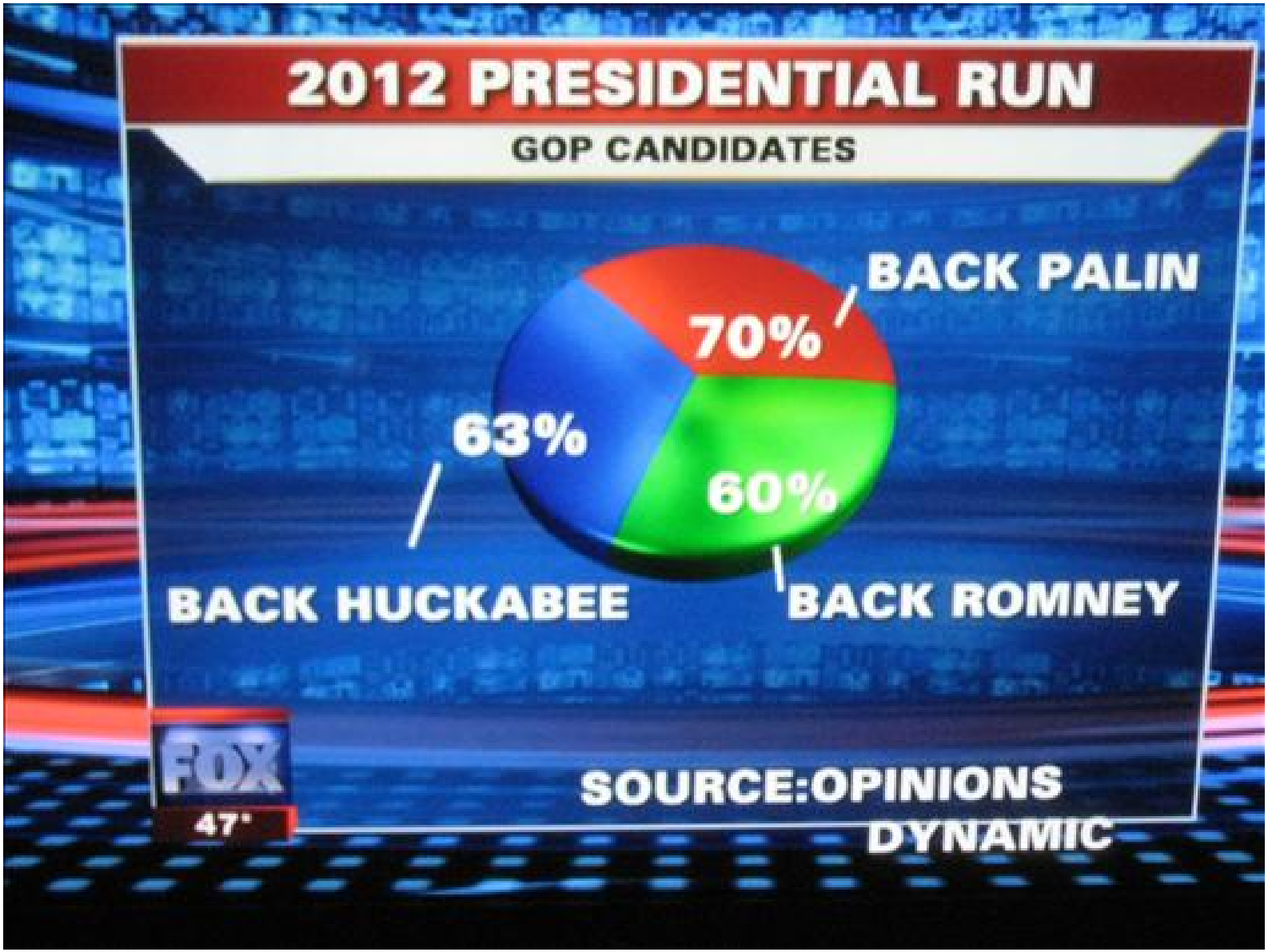
Ihr Alter (in Jahren)										
Ihre Größe (in cm)										
Ihr Geschlecht (m/w)										
Wie alt ist (bzw. wäre) Ihr Vater heute?										
Wie alt ist (bzw. wäre) Ihre Mutter heute?										
Größe Ihres Vaters (cm)										
Größe Ihrer Mutter (cm)										
Wie viele Geschwister haben Sie?										
Wunschfarbe für Ihr nächstes Smartphone; mögliche Auswahl: (si)lber, (sc)hwarz, (w)eiß, (g)elb, (b)lau, (r)ot										
Ausgaben für Ihre mobile Kommunikation (egal wer bezahlt hat) in den vergangenen 12 Monaten inkl. Hardware (Smartphone, Zubehör), Software (Apps), Vertrag										
Wie viele Paar Schuhe besitzen Sie?										
Ausgaben für Ihre Schuhe (egal wer bezahlt hat) in den letzten 12 Monaten										
Ihre Note in der Matheklausur										
Waren Sie mir Ihrer Leistung in Mathe zufrieden? Antworten: (s)ehr zufrieden, (z)ufrieden, (g)eht so, (n)icht zufrieden										

Statistik: Table of Contents

Auswahl der Stichprobe

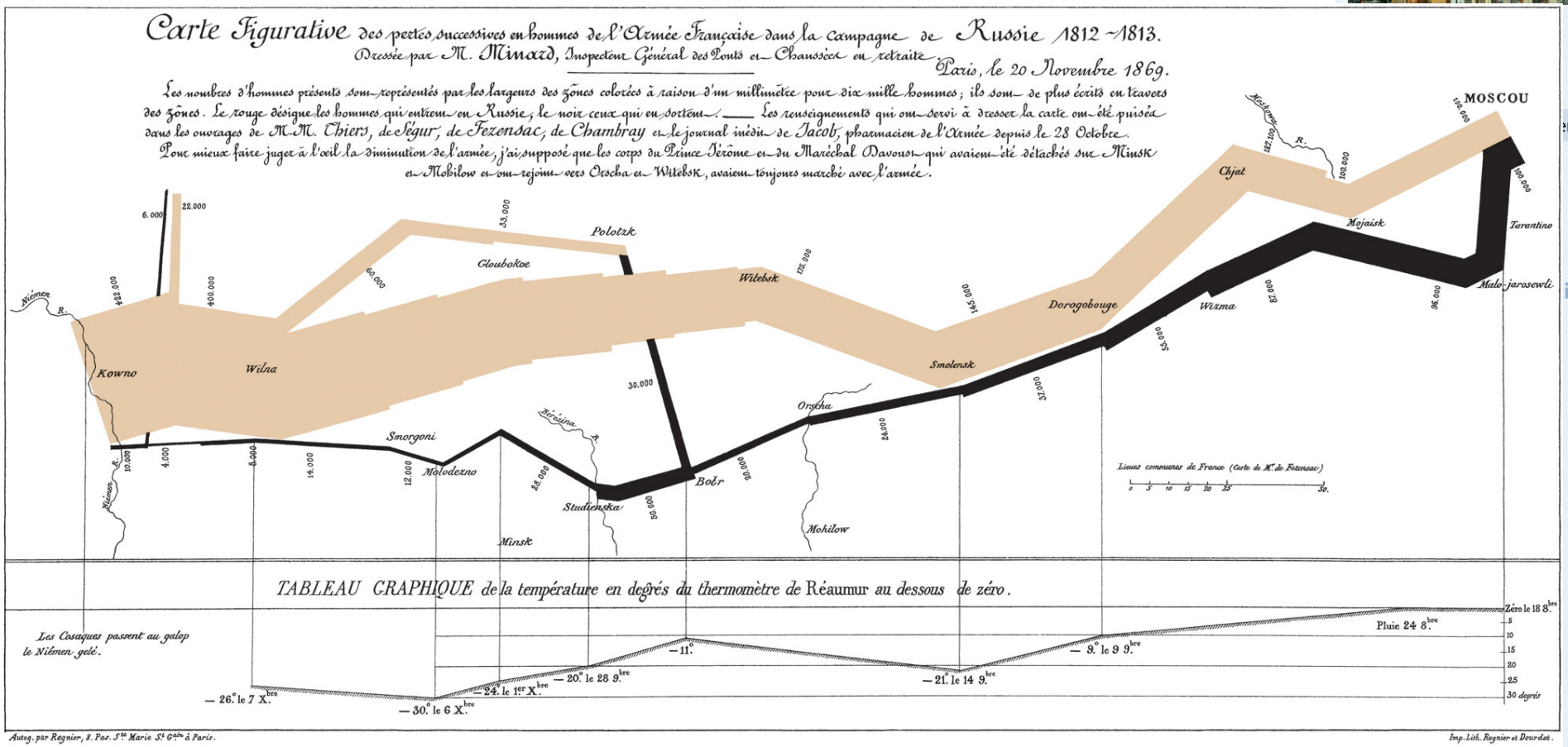


Aussage?



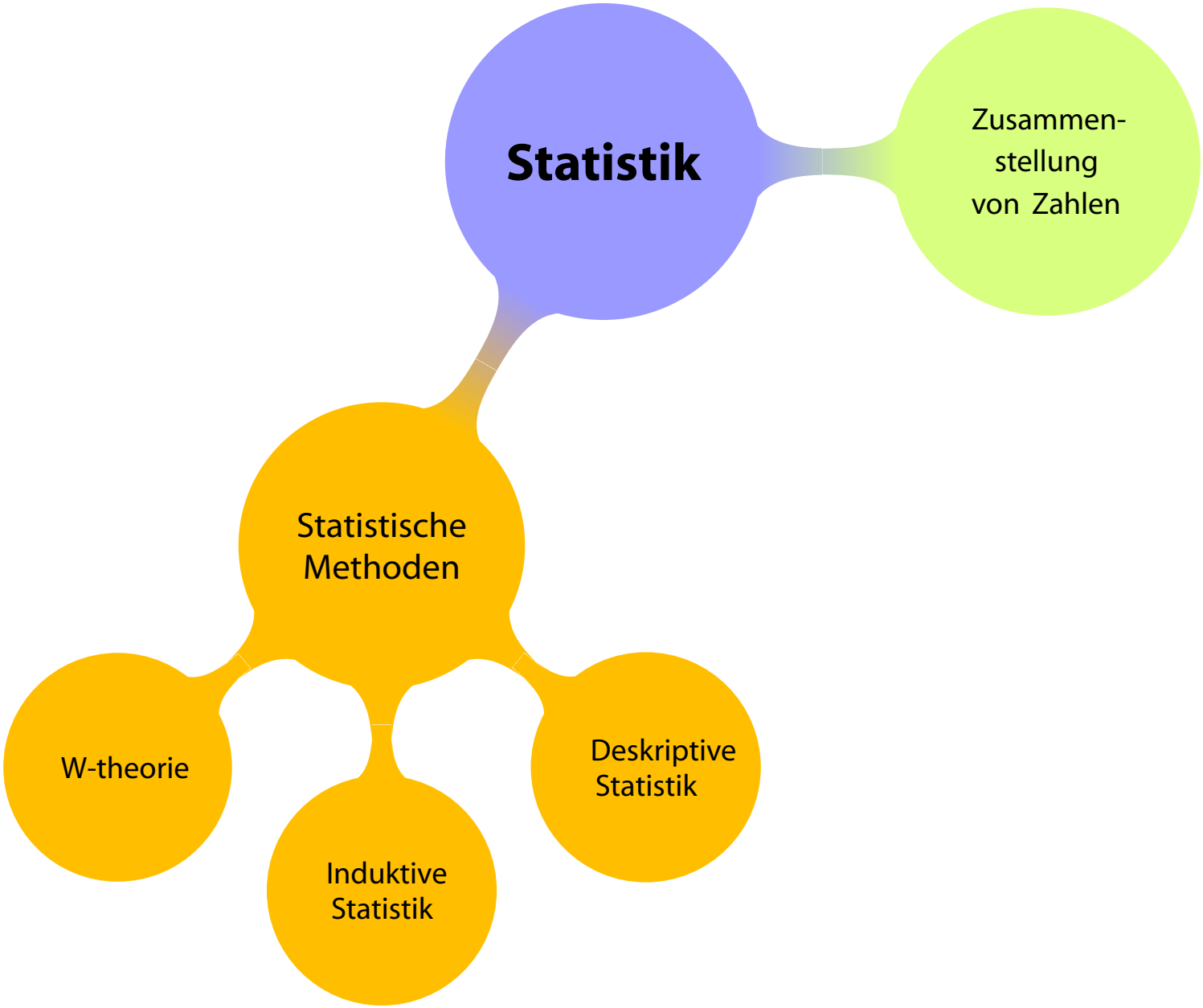
Quelle: Friendly, Michael (2001), Gallery of Data Visualization, `http://www.datavis.ca/gallery/`, Stand November 2014

Minards Grafik von 1869 über Napoleons Rußlandfeldzug



Quelle: Wikimedia Commons, `http://goo.gl/T7ZNme`, Stand November 2014

Bedeutungen des Begriffs Statistik



Beispiel

12 Beschäftigte werden nach der Entfernung zum Arbeitsplatz (in km) befragt.
Antworten: 4, 11, 1, 3, 5, 4, 20, 4, 6, 16, 10, 6

- deskriptiv:
  - Durchschnittliche Entfernung: 7,5
  - Klassenbildung:

Klasse	$[0;5)$	$[5;15)$	$[15;30)$
Häufigkeit	5	5	2

- induktiv:
  - Schätze die mittlere Entfernung **aller** Beschäftigten.
  - Prüfe, ob die mittlere Entfernung geringer als 10 km ist.

Merkmale

Objekte

- Merkmalsträger: Untersuchte statistische Einheit
- Merkmal: Interessierende Eigenschaft des Merkmalträgers
- (Merkmals-)Ausprägung: Konkret beobachteter Wert des Merkmals
- Grundgesamtheit: Menge aller relevanten Merkmalsträger
- Typen von Merkmalen:

  a) qualitativ – quantitativ
     - qualitativ: z.B. Geschlecht
     - quantitativ: z.B. Schuhgröße
     - Qualitative Merkmale sind quantifizierbar (z.B.: weiblich 1, männlich 0)

  b) diskret – stetig
     - diskret: Abzählbar viele unterschiedliche Ausprägungen
     - stetig: Alle Zwischenwerte realisierbar

Skalenniveaus

Nominalskala:

- Zahlen haben nur Bezeichnungsfunktion
- z.B. Artikelnummern, Lieblingsfarbe

Ordinalskala:

- zusätzlich Rangbildung möglich
- z.B. Schulnoten
- Differenzen sind aber **nicht** interpretierbar!
  ➠ Addition usw. ist unzulässig.

Kategorien

Kardinalskala: , metrisch

- zusätzlich Differenzbildung sinnvoll
- z.B. Gewinn
- Noch feinere Unterscheidung in: Absolutskala, Verhältnisskala, Intervallskala

Zahlen

Skalendegression und Skalenprogression

Ziel der Skalierung: Gegebene Information angemessen abbilden, möglichst ohne Über- bzw. Unterschätzungen

Es gilt:

- Grundsätzlich können alle Merkmale nominal skaliert werden.
- Grundsätzlich kann jedes metrische Merkmal ordinal skaliert werden.

Das nennt man Skalendegression. Dabei: Informationsverlust

Aber:

- Nominale Merkmale dürfen **nicht** ordinal- oder metrisch skaliert werden.
- Ordinale Merkmale dürfen **nicht** metrisch skaliert werden.

Das nennt nennt man Skalenprogression. Dabei: Interpretation von **mehr Informationen** in die Merkmale, als inhaltlich vertretbar.
(Gefahr der **Fehlinterpretation**)

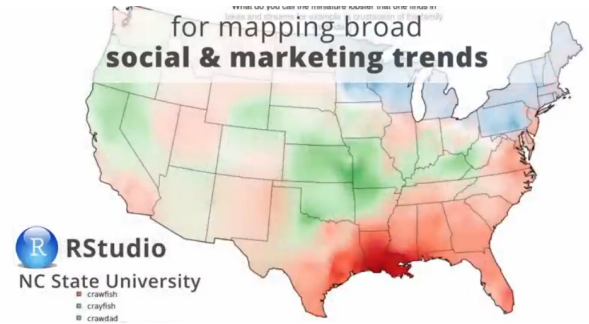
Was ist R und warum soll man es benutzen?

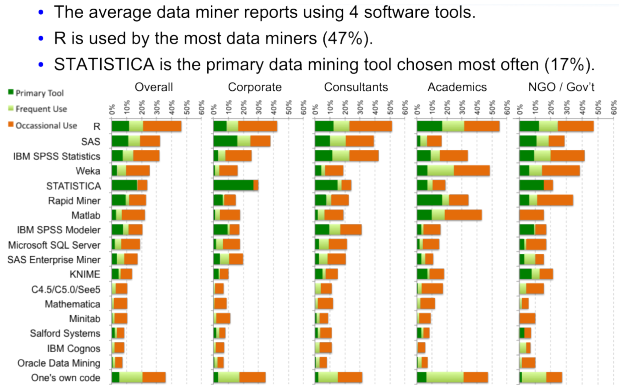
- R ist ein freies Softwarepaket zu Statistik und Datenanalyse

graphics source: http://goo.gl/W70kms

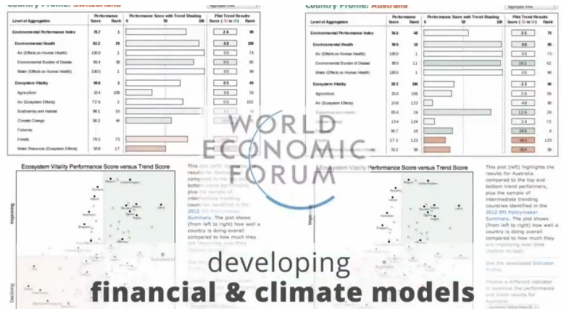
Was ist R und warum soll man es benutzen?

- R ist ein freies Softwarepaket zu Statistik und Datenanalyse
- R ist sehr mächtig und weit verbreitet in Wissenschaft und Industrie (sogar von mehr Leuten benutzt als z.B. SPSS)





source: http://goo.gl/axhGhh



graphics source: http://goo.gl/W70kms

Was ist R und warum soll man es benutzen?

- R ist ein freies Softwarepaket zu Statistik und Datenanalyse
- R ist sehr mächtig und weit verbreitet in Wissenschaft und Industrie (sogar von mehr Leuten benutzt als z.B. SPSS)
- Ursprung von R: 1993 an der Universität Auckland von Ross Ihaka and Robert Gentleman entwickelt

graphics source: http://goo.gl/W70kms

Was ist R und warum soll man es benutzen?

- R ist ein freies Softwarepaket zu Statistik und Datenanalyse
- R ist sehr mächtig und weit verbreitet in Wissenschaft und Industrie (sogar von mehr Leuten benutzt als z.B. SPSS)
- Ursprung von R: 1993 an der Universität Auckland von Ross Ihaka and Robert Gentleman entwickelt
- Seitdem: Viele Leute haben R verbessert mit tausenden von Paketen für viele Anwendungen

graphics source: http://goo.gl/W70kms

Was ist R und warum soll man es benutzen?

- R ist ein freies Softwarepaket zu Statistik und Datenanalyse
- R ist sehr mächtig und weit verbreitet in Wissenschaft und Industrie (sogar von mehr Leuten benutzt als z.B. SPSS)
- Ursprung von R: 1993 an der Universität Auckland von Ross Ihaka and Robert Gentleman entwickelt
- Seitdem: Viele Leute haben R verbessert mit tausenden von Paketen für viele Anwendungen
- Nachteil (auf den ersten Blick): Kein point und click tool

```
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```

graphics source: http://goo.gl/W70kms

Was ist R und warum soll man es benutzen?

- R ist ein freies Softwarepaket zu Statistik und Datenanalyse
- R ist sehr mächtig und weit verbreitet in Wissenschaft und Industrie (sogar von mehr Leuten benutzt als z.B. SPSS)
- Ursprung von R: 1993 an der Universität Auckland von Ross Ihaka and Robert Gentleman entwickelt
- Seitdem: Viele Leute haben R verbessert mit tausenden von Paketen für viele Anwendungen
- Nachteil (auf den ersten Blick): Kein point und click tool
- Großer Vorteil (auf den zweiten Blick): Kein point und click tool

```
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```

Download: `R-project.org`

graphics source: `http://goo.gl/W70kms`

Was ist RStudio?

- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.

Was ist RStudio?

- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.
- Gibt's für OSX, Linux und Windows

Was ist RStudio?

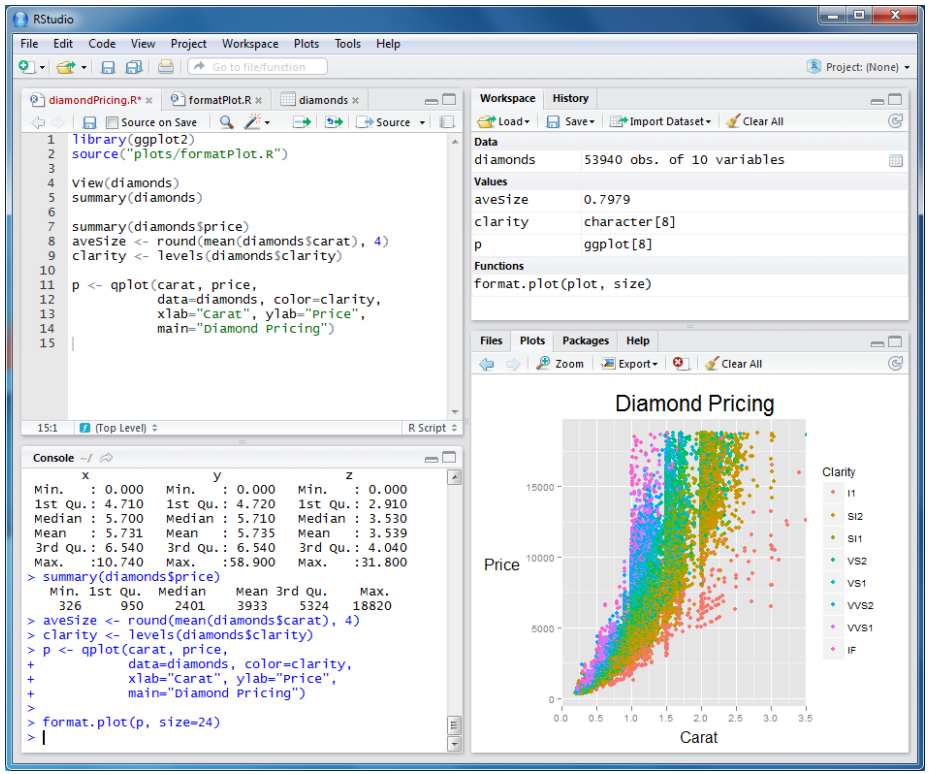
- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.
- Gibt's für OSX, Linux und Windows
- Ist auch frei

Was ist RStudio?

- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.
- Gibt's für OSX, Linux und Windows
- Ist auch frei
- Trotzdem: Sie müssen Kommandos schreiben

RStudio™

Free & Open-Source IDE for R

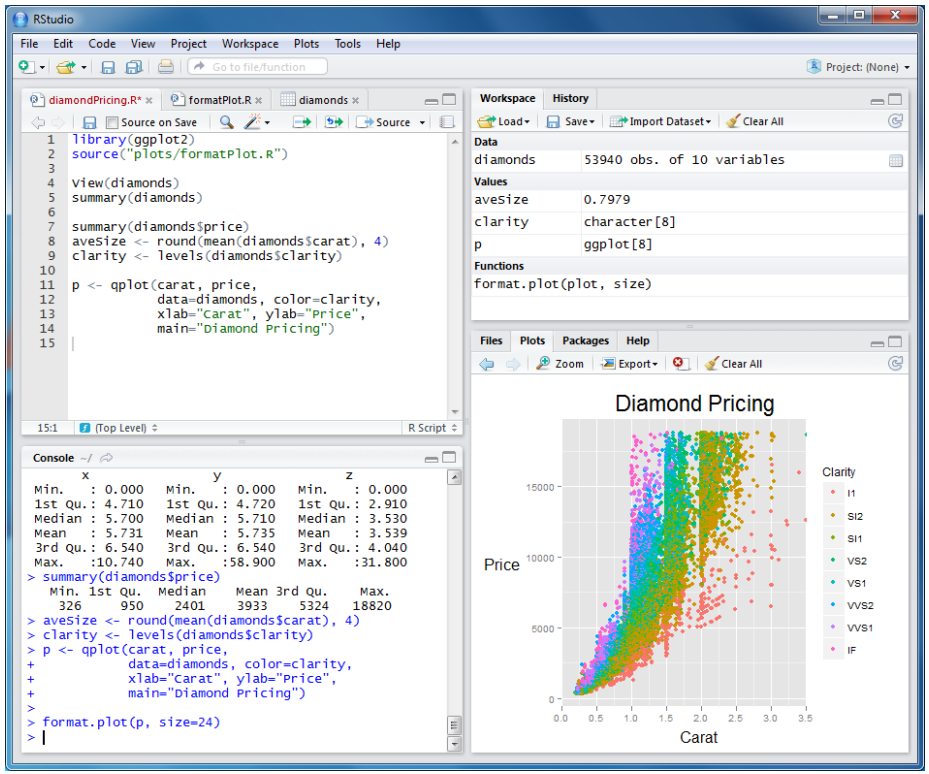


Was ist RStudio?

- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.
- Gibt's für OSX, Linux und Windows
- Ist auch frei
- Trotzdem: Sie müssen Kommandos schreiben
- Aber: RStudio unterstützt Sie dabei

RStudio™

Free & Open-Source IDE for R

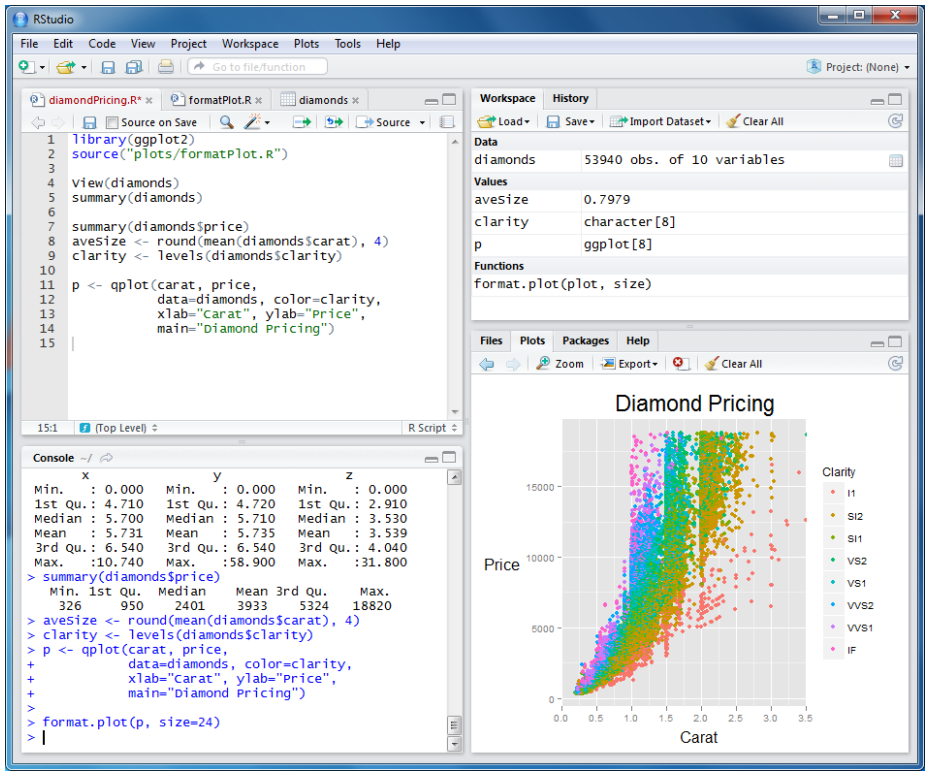


Was ist RStudio?

- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.
- Gibt's für OSX, Linux und Windows
- Ist auch frei
- Trotzdem: Sie müssen Kommandos schreiben
- Aber: RStudio unterstützt Sie dabei
- **Download:** `RStudio.com`

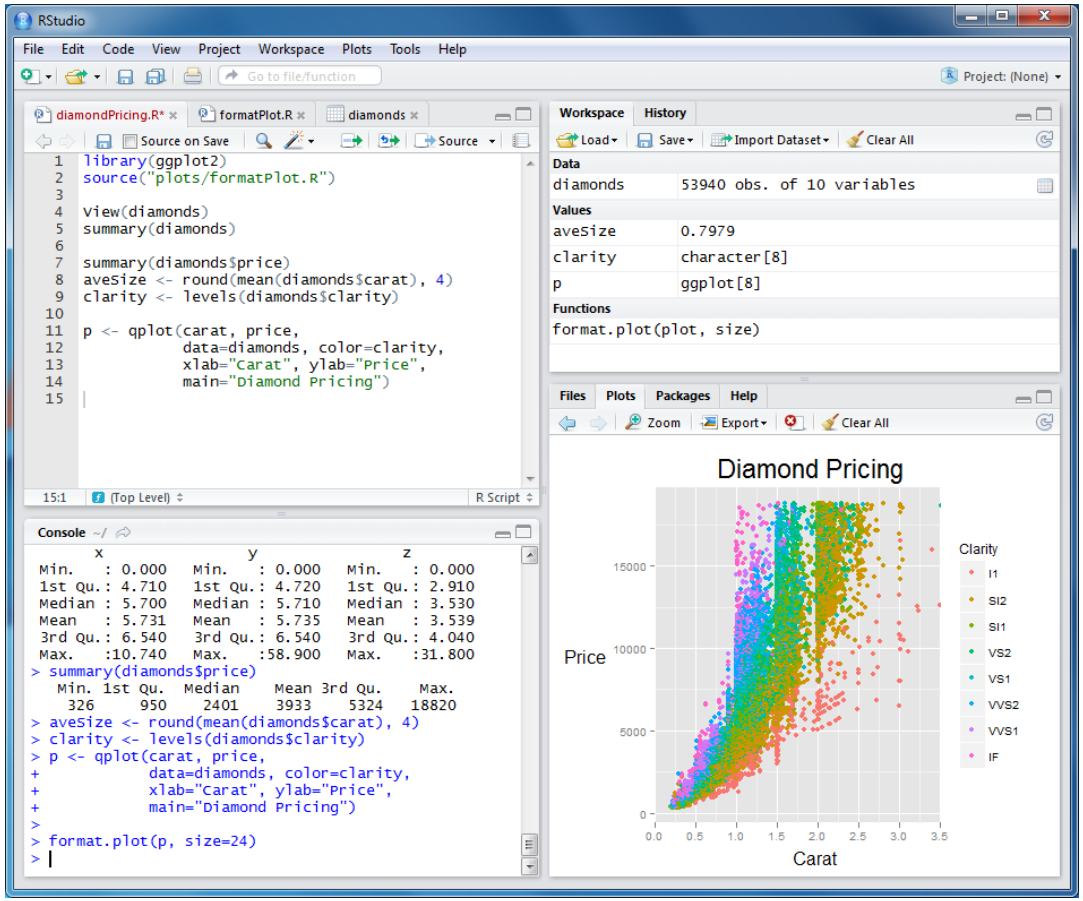
RStudio™

Free & Open-Source IDE for R



RStudio Kennenlernen

- Code
- Console
- Workspace
- History
- Files
- Plots
- Packages
- Help
- Auto-Completion
- Data Import



Daten einlesen und Struktur anschauen


```
# Arbeitsverzeichnis setzen (alternativ über Menü)
setwd("C:/ste/work/vorlesungen/2014WS_Doktorandenworkshop/2015_01_Statistik_Workshop")

# Daten einlesen aus einer csv-Datei (Excel)
MyData = read.csv2(file="../Daten/Umfrage_HSA_2014_03.csv", header=TRUE)
```

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

## 'data.frame': 205 obs. of  10 variables:
##  $ Alter      : int  21 20 19 20 20 24 20 27 23 21 ...
##  $ Geschlecht : Factor w/ 2 levels "Frau","Mann": 1 1 1 1 1 2 1 1 2 2 ...
##  $ AlterV     : int  54 57 49 45 43 54 49 53 52 55 ...
##  $ AlterM     : int  51 57 58 49 42 52 53 53 48 55 ...
##  $ Geschwister: int  1 0 3 3 5 2 2 1 2 1 ...
##  $ Farbe      : Factor w/ 6 levels "blau","gelb",..: 6 6 4 4 6 4 3 6 4 6 ...
##  $ AusgSchuhe : int  50 500 400 100 450 90 250 200 300 200 ...
##  $ AnzSchuhe  : int  17 22 15 15 22 8 20 10 3 7 ...
##  $ AusgKomm   : num  156 450 240 35.8 450 250 100 300 450 1300 ...
##  $ MatheZufr  : Ord.factor w/ 4 levels "nicht"<"geht so"<..: 1 4 4 4 4 2 1 1 3 3 ...
```

Erste Zeilen der Datentabelle

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

```
##   Alter Geschlecht AlterV AlterM Geschwister   Farbe AusgSchuhe AnzSchuhe AusgKomm MatheZufr
## 1    21       Frau     54     51           1   weiss         50        17    156.0     nicht
## 2    20       Frau     57     57           0   weiss        500        22    450.0      sehr
## 3    19       Frau     49     58           3 schwarz        400        15    240.0      sehr
## 4    20       Frau     45     49           3 schwarz        100        15     35.8      sehr
## 5    20       Frau     43     42           5   weiss        450        22    450.0      sehr
## 6    24       Mann     54     52           2 schwarz         90         8    250.0   geht so
```

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

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

```
## [1] 205
```

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

```
## [1] 10
```


Daten kennenlernen


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

##   Geschlecht AlterV AlterM Geschwister
## 1       Frau     54     51           1
## 2       Frau     57     57           0
## 3       Frau     49     58           3
```

```
# Auswahl von Objekten über logische Ausdrücke
head(Geschlecht=="Frau" & Alter<19, 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

# Einsetzen in Klammern und Ausgabe von Alter des Studenten, seines Vaters und seiner Mutter
MyData[Geschlecht=="Frau" & Alter<19, # Objektauswahl
      c("Alter", "AlterM", "AlterV")  # Welche Merkmale anzeigen?
      ]

##     Alter AlterM AlterV
## 23     18     50     52
## 44     18     37     43
## 51     18     51     54
## 57     18     53     57
## 74     18     53     49
## 126    18     44     45
## 139    18     51     58
## 185    18     46     48
## 193    18     49     47
```

Daten kennenlernen


```
# Zeige die Männer, die mehr als 1000 Euro für Schuhe
# und Mobilfunk zusammen ausgegeben haben
MyData[Geschlecht=="Mann" & AusgSchuhe + AusgKomm > 1000,
       c("Alter", "Geschwister", "Farbe", "AusgSchuhe", "AusgKomm")]

##     Alter Geschwister   Farbe AusgSchuhe AusgKomm
## 10     21           1   weiss        200     1300
## 15     20           1     rot        400      815
## 26     20           1 schwarz        200     1250
## 40     21           0  silber        300      825
## 87     20           1    blau       1000      350
## 113    25           0 schwarz        280     1200
## 146    24           1 schwarz        300      900
## 177    19           2 schwarz        500      720
## 178    23           1 schwarz        450      630
## 192    20           0 schwarz        400      950
```

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 7 6 6 7 4 5 4 7 7 5 5 5 5 6 6 4 5 2 5 4
## [49] 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	—

Alter von Studierenden in Statistik Vorlesung

- Absolute Häufigkeiten

	18	19	20	21	22	23	24	25	26	27	28	29	31	32	33	34	36
$h(a_j)$	10	27	39	29	26	23	14	6	7	6	7	4	1	3	1	1	1
$H(a_j) = \sum_{i=1}^{j} h(a_j)$	10	37	76	105	131	154	168	174	181	187	194	198	199	202	203	204	205

- Relative frequencies

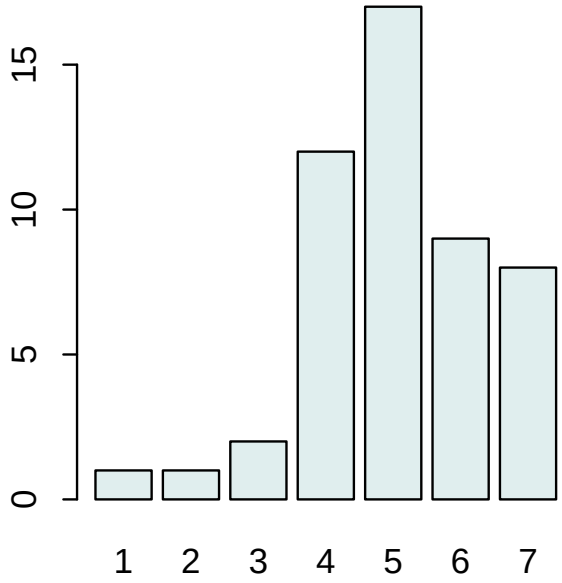
	18	19	20	21	22	23	24	25	26	27	28	29	31	32	33	34	36
$f(a_j) = \frac{h(a_j)}{n}$	0.05	0.13	0.19	0.14	0.13	0.11	0.07	0.03	0.03	0.03	0.03	0.02	0.00	0.01	0.00	0.00	0.00
$F(a_j) = \sum_{i=1}^{j} f(a_j)$	0.05	0.18	0.37	0.51	0.64	0.75	0.82	0.85	0.88	0.91	0.95	0.97	0.97	0.99	0.99	1.00	1.00

❶ Balkendiagramm

```
table(x)

## x
##  1  2  3  4  5  6  7
##  1  1  2 12 17  9  8

barplot(table(x), col="azure2")
```



(Höhe proportional zu Häufigkeit)

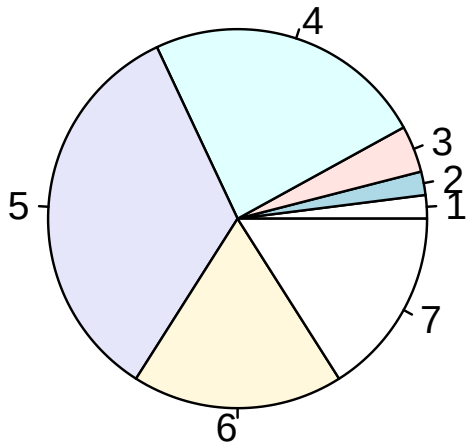
❷ Kreissektorendiagramm

Winkel: $w_j = 360° \cdot f(a_j)$

z.B. $w_1 = 360° \cdot \frac{1}{50} = 7{,}2°$

$w_7 = 360° \cdot \frac{8}{50} = 57{,}6°$

```
pie(table(x))
```

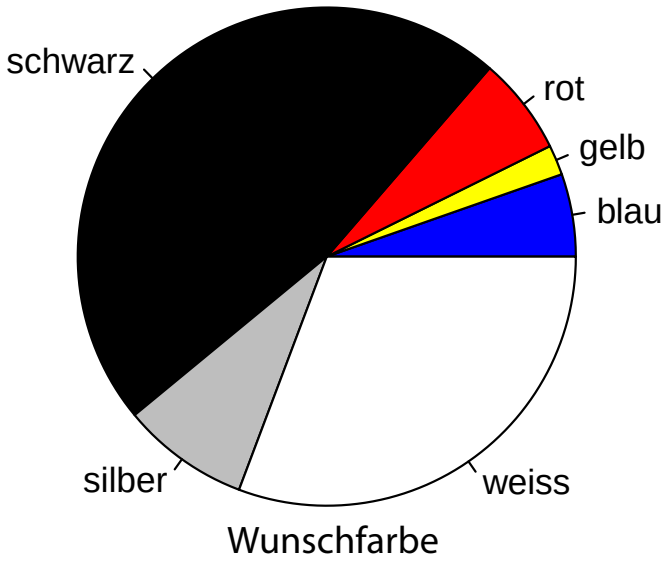


(Fläche proportional zu Häufigkeit)

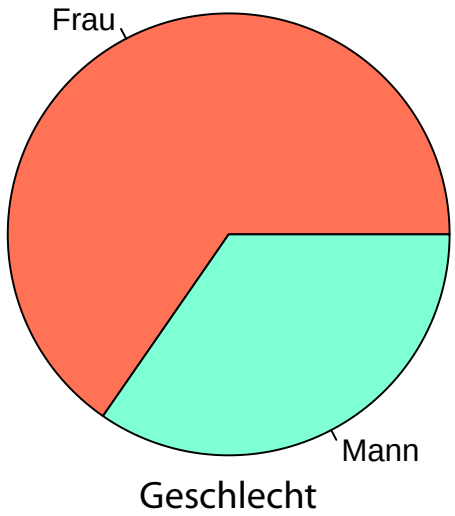
Umfrage: Kuchen

Kreissektorendiagramm

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

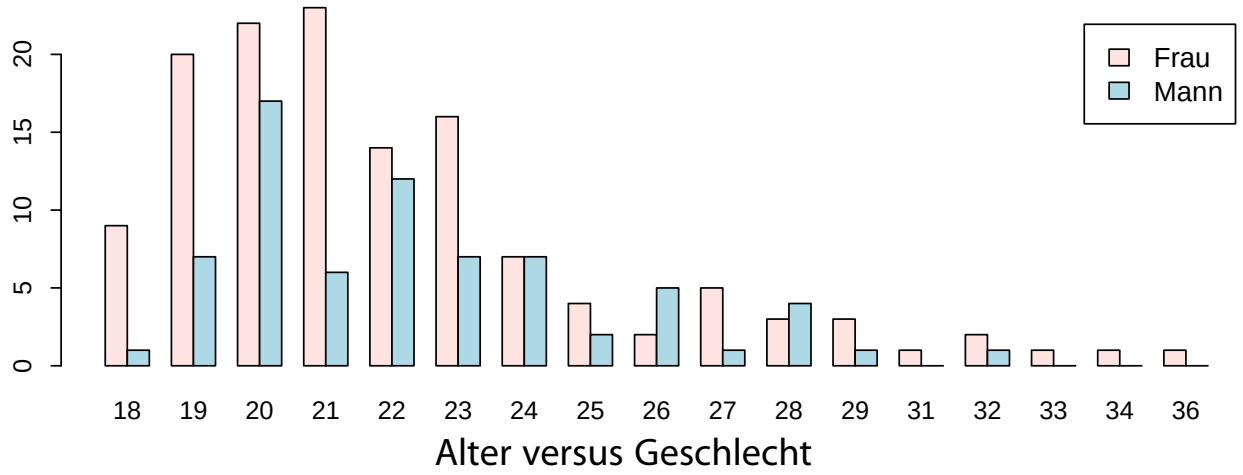



```
pie(table(MyData$Geschlecht),
    col=c("coral1", "aquamarine"))
```

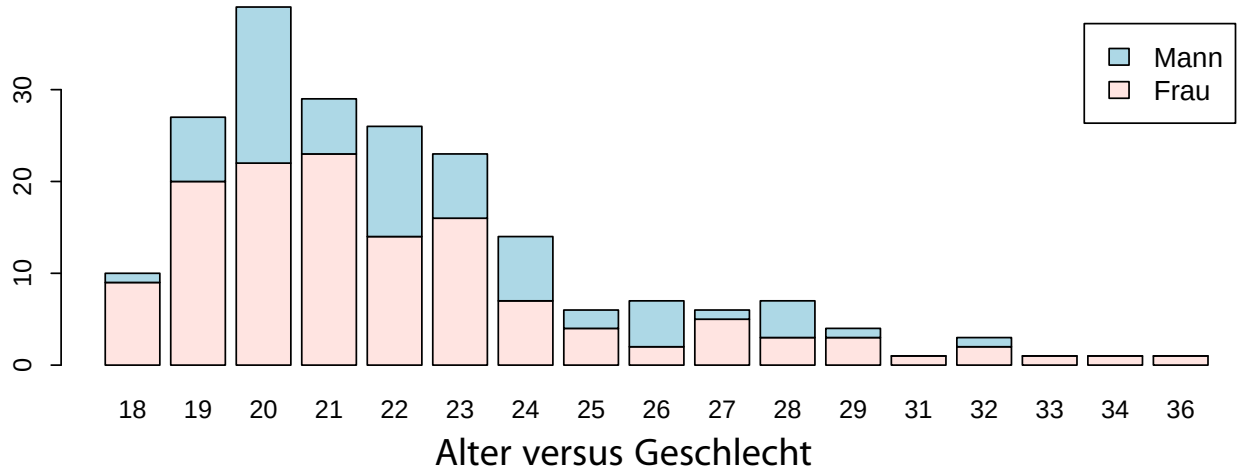


Balkendiagramm, Klassen getrennt oder gestapelt

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




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



❸ Histogramm

- für klassierte Daten
- Fläche proportional zu Häufigkeit:

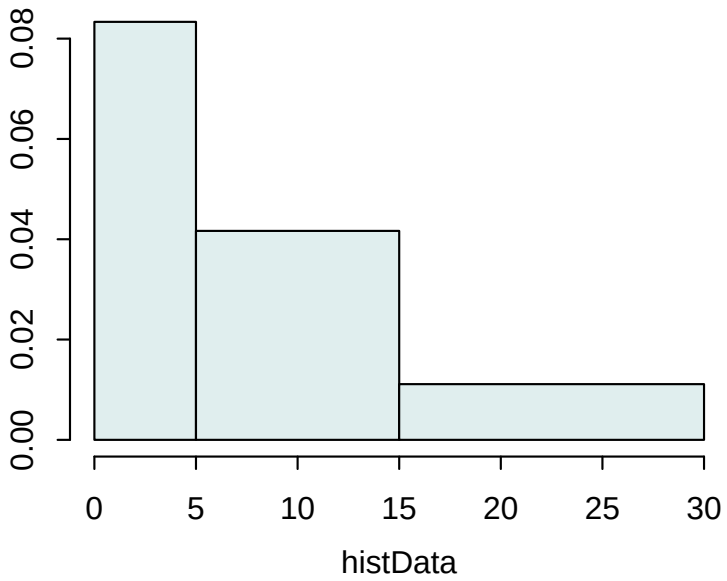
$$\text{Höhe}_j \cdot \text{Breite}_j = c \cdot h(a_j)$$

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

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

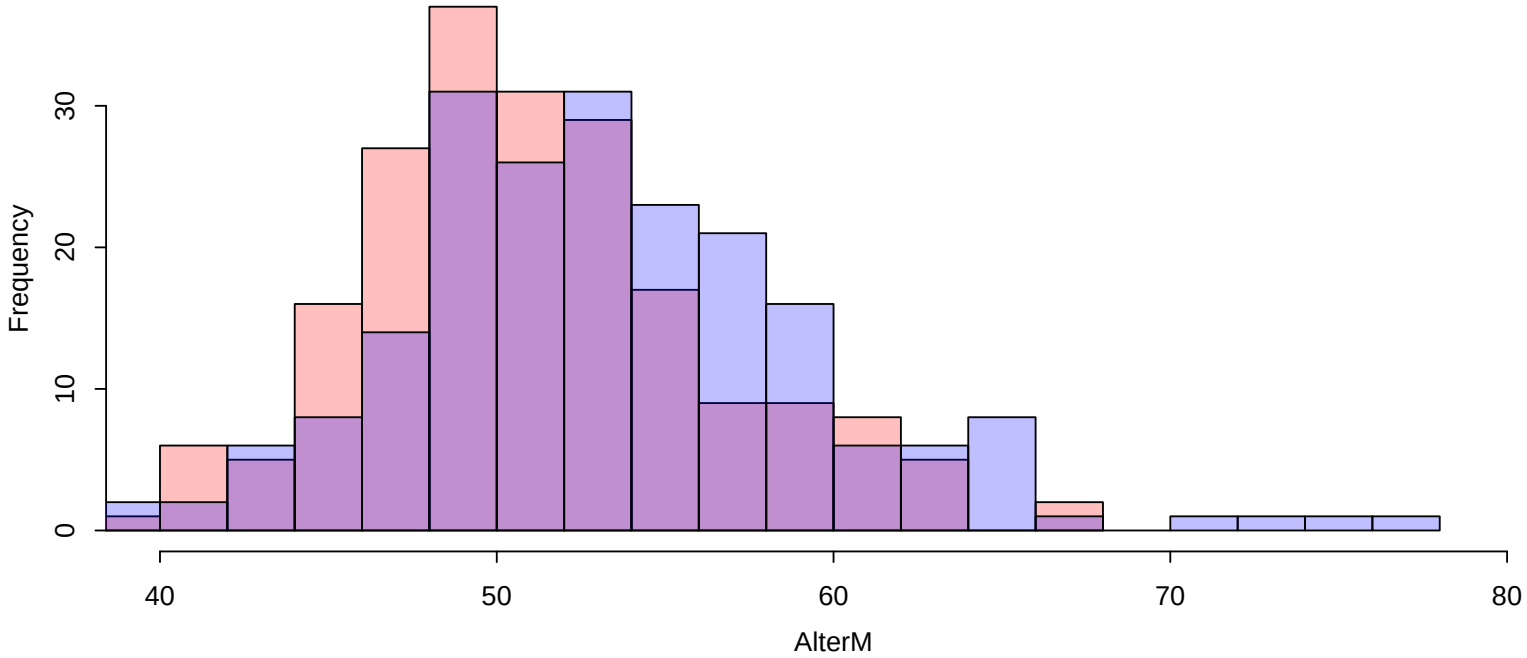
Klasse	$[0;5)$	$[5;15)$	$[15;30]$
$h(a_j)$	5	5	2
Breite_j	5	10	15
Höhe_j	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{90}$

```
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

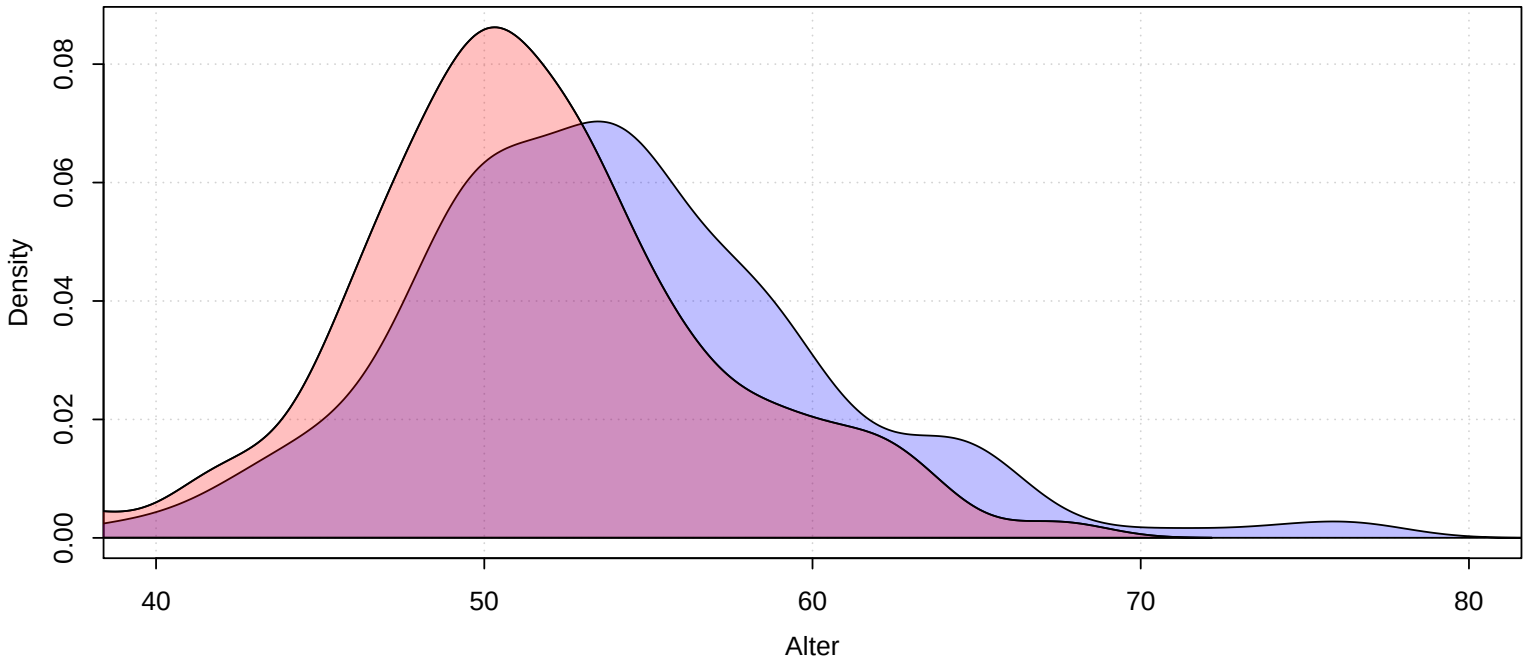
```
plot(hist(AlterM, plot=F, breaks=20),
          col=rgb(1,0,0,1/4), # make red transparent
          main="",
          xlim=c(40,80)) # draw from 40 to 80
plot(hist(AlterV, plot=F, breaks=20),
          col=rgb(0,0,1,1/4),
          add=TRUE)
```



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

Dichteplot

```
densMutter = density(AlterM)
densVater = density(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)

Lageparameter

"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:

Eink.
1000
1000
⋮
1000
40000
} 10

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

$\Rightarrow x_{Mod} = 1$

Sinnvoll bei allen Skalenniveaus.

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.

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

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:

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

Lageparameter

Ausgaben für Schuhe

```
median(AusgSchuhe)

## [1] 250

mean(AusgSchuhe)

## [1] 296.6244
```

Alter

```
median(Alter)

## [1] 21

mean(Alter)

## [1] 22.21951
```

Lieblingsfarbe

```
summary(Geschlecht)

## Frau Mann
##  134   71
```

Alter der Mutter

```
median(AlterM)

## [1] 51

mean(AlterM)

## [1] 51.49756
```

- 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:

$$\begin{aligned} &\text{a)} \quad SP = 2050 - 1950 = 100 \\ &\text{b)} \quad SP = 6000 - 0 \quad\;\; = 6000 \end{aligned}$$

- Mittlere quadratische Abweichung:

  Empirische Varianz

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

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

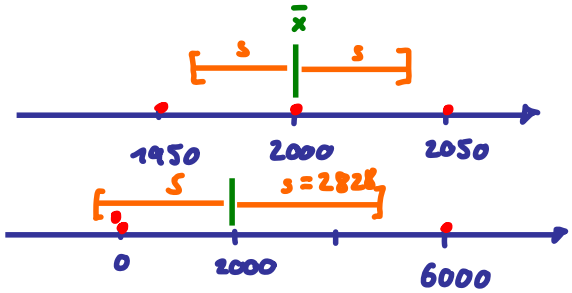
$= \frac{1}{3} \cdot (0^2 + 0^2 + 6000^2) - 2000^2 \qquad = 8000000$

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

Im Beispiel:

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

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

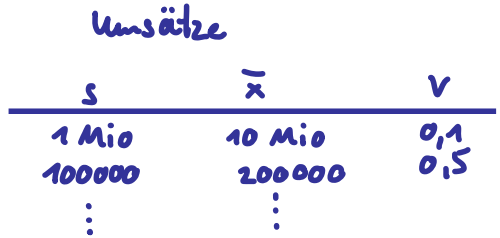


- **Variationskoeffizient:** $V = \frac{s}{\bar{x}}$ (maßstabsunabhängig)

Im Beispiel:

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

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



Lage und Streuung: Überblick

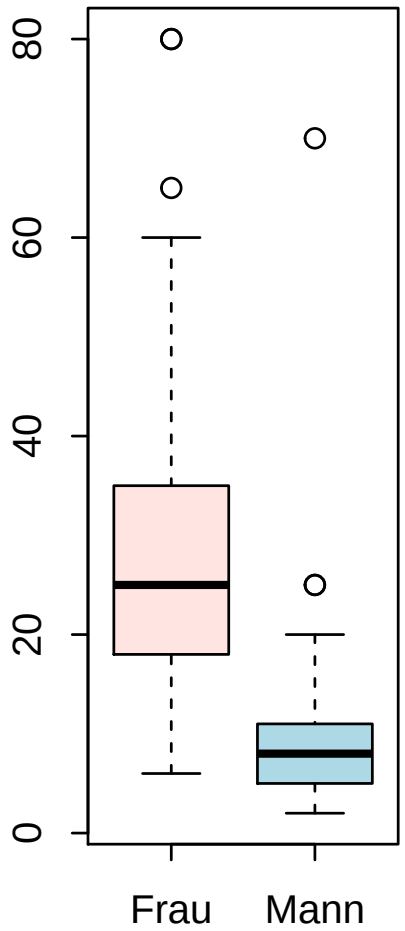

```
LageStreuung = function(x) {
  x=na.omit(x)  # ignoriere fehlende Werte
  n = length(x) # Anzahl nicht fehlender Werte
  popV = var(x)*(n-1)/n # var() ist nicht mittl. qu. Abweichung
  return(list(mean=mean(x),
              median=median(x),
              Variance=popV,
              StdDev=sqrt(popV),
              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	22.22	53.95	51.50	1.47	21.58	296.62
median	21.00	54.00	51.00	1.00	20.00	250.00
Variance	11.04	39.16	29.83	1.62	232.43	48418.70
StdDev	3.32	6.26	5.46	1.27	15.25	220.04
VarCoeff	0.15	0.12	0.11	0.86	0.71	0.74

Lage und Streuung als Grafik: Boxplot

- Graphische Darstellung von Lage und Streuung
- Box: Oberer/Unterer Rand: 3. bzw. 1. Quartil,
- 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(AnzSchuhe ~ Geschlecht,
        col=c("mistyrose", "lightblue"),
        data=MyData, main="")
```



„Wieviel Paar Schuhe besitzen Sie?“

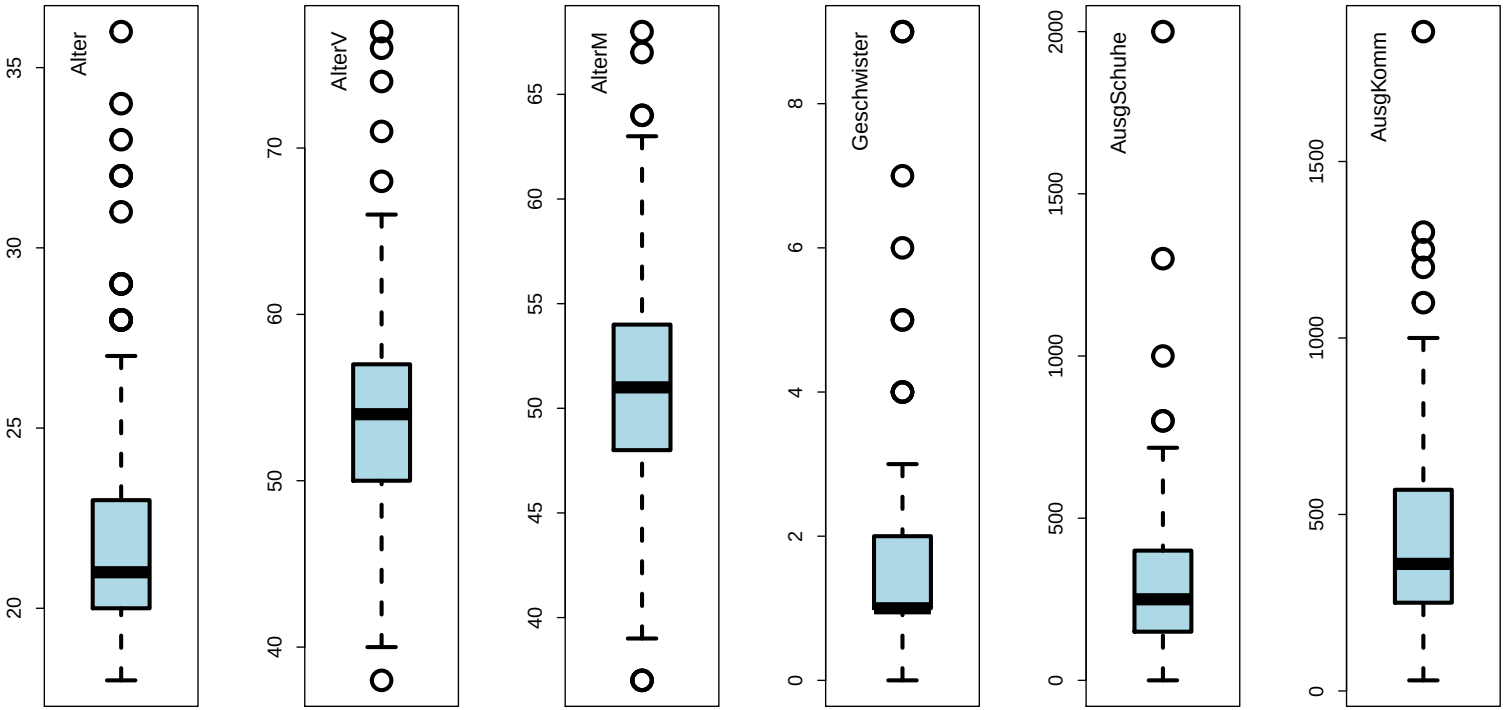
Dateninspektion: Überblick über alle Variablen


```
summary(MyData)

##     Alter       Geschlecht    AlterV          AlterM     Geschwister       Farbe
##  Min.   :18.00   Frau:134   Min.   :38.00   Min.   :37.0   Min.   :0.000   blau   :11
##  1st Qu.:20.00   Mann: 71   1st Qu.:50.00   1st Qu.:48.0   1st Qu.:1.000   gelb   : 4
##  Median :21.00              Median :54.00   Median :51.0   Median :1.000   rot    :13
##  Mean   :22.22              Mean   :53.95   Mean   :51.5   Mean   :1.473   schwarz:97
##  3rd Qu.:23.00              3rd Qu.:57.00   3rd Qu.:54.0   3rd Qu.:2.000   silber :17
##  Max.   :36.00              Max.   :77.00   Max.   :68.0   Max.   :9.000   weiss  :63
##    AusgSchuhe       AnzSchuhe        AusgKomm          MatheZufr
##  Min.   :   0.0   Min.   : 2.00   Min.   :  30.0   nicht    :68
##  1st Qu.: 150.0   1st Qu.:10.00   1st Qu.: 250.0   geht so  :47
##  Median : 250.0   Median :20.00   Median : 360.0   zufrieden:43
##  Mean   : 296.6   Mean   :21.58   Mean   : 429.4   sehr     :26
##  3rd Qu.: 400.0   3rd Qu.:30.00   3rd Qu.: 570.0   NA's     :21
##  Max.   :2000.0   Max.   :80.00   Max.   :1868.0
```

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)
}
```



- Gegeben: kardinale Werte $0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n$
- **Achtung!** Die Werte müssen aufsteigend sortiert werden!
- Lorenzkurve:

> Wieviel Prozent der Merkmalssumme entfällt auf die x Prozent kleinsten Merkmalsträger?

- **Beispiel:** Die 90 % ärmsten besitzen 20 % des Gesamtvermögens.
- Streckenzug: $(0,0),\ (u_1,v_1),\ \ldots,\ (u_n,v_n) = (1,1)$ mit

$$v_k = \text{Anteil der } k \text{ kleinsten MM-Träger an der MM-Summe} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i}$$

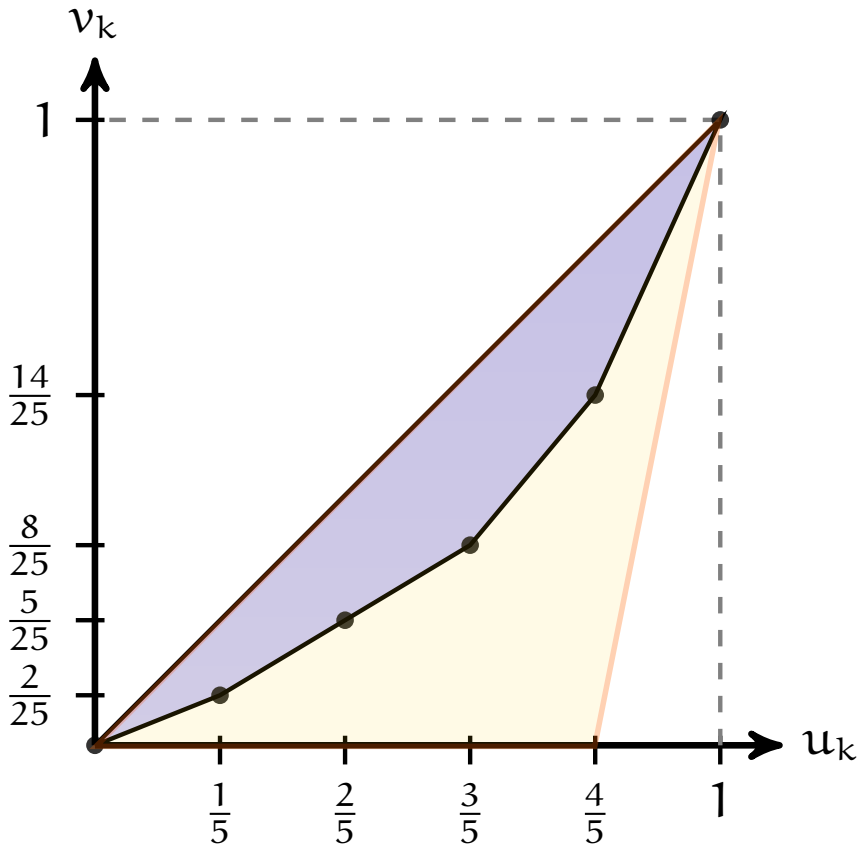
$$u_k = \text{Anteil der } k \text{ kleinsten an der Gesamtzahl der MM-Träger} = \frac{k}{n}$$

Lorenzkurve: Beispiel

Markt mit fünf Unternehmen; Umsätze: 6, 3, 11, 2, 3 (Mio. €)

$\Rightarrow n = 5,\ \sum_{k=1}^{5} x_k = 25$

k	1	2	3	4	5
x_k	2	3	3	6	11
p_k	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{3}{25}$	$\frac{6}{25}$	$\frac{11}{25}$
v_k	$\frac{2}{25}$	$\frac{5}{25}$	$\frac{8}{25}$	$\frac{14}{25}$	1
u_k	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

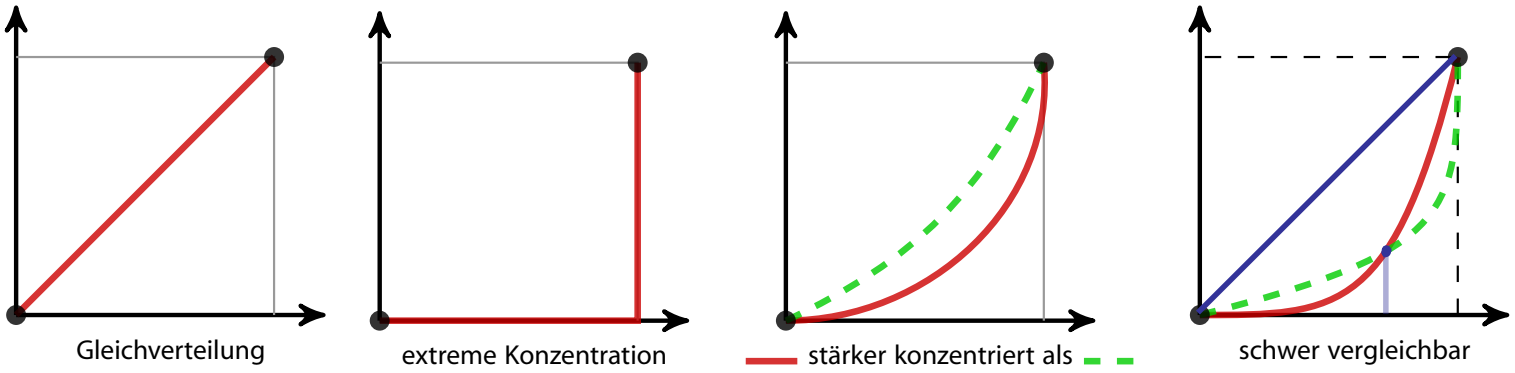


Knickstellen:

- Bei i-tem Merkmalsträger $\Longleftrightarrow$ $x_{i+1} > x_i$
- Empirische Verteilungsfunktion liefert Knickstellen:

a_j	2	3	6	11
$h(a_j)$	1	2	1	1
$f(a_j)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
$F(a_j)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

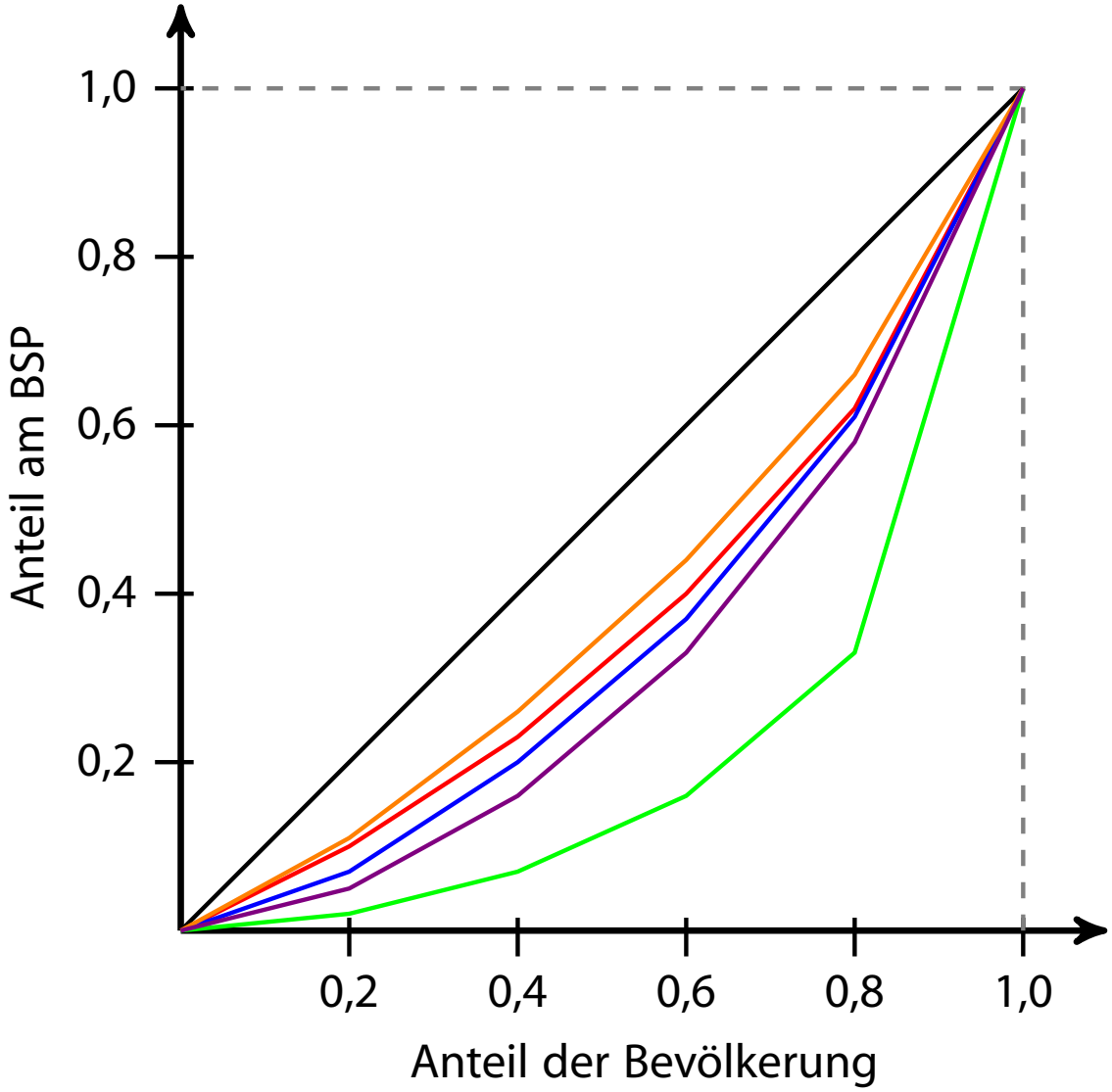
Vergleich von Lorenzkurven:



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

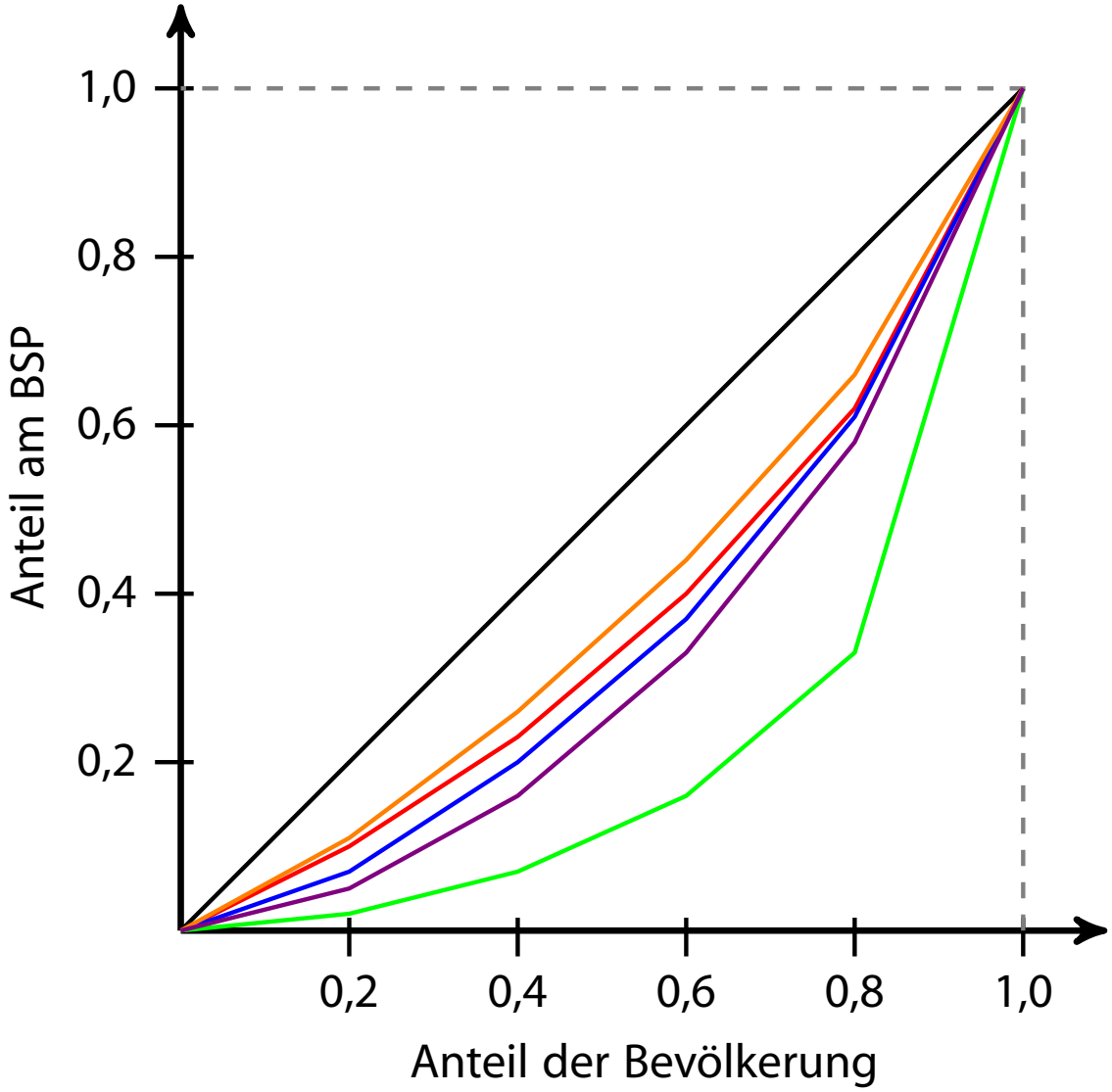
(Stand 2000)



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

(Stand 2000)



Gini-Koeffizient

- Numerisches Maß der Konzentration: Gini-Koeffizient G

$$G = \frac{\text{Fläche zwischen } 45°\text{-Linie und L}}{\text{Fläche unter } 45°\text{-Linie}} = \frac{\text{◢}}{\text{◢}}$$

- Aus den Daten:

$$G = \frac{2\sum_{i=1}^{n} i\,x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i} = \frac{2\sum_{i=1}^{n} i\,p_i - (n+1)}{n} \quad \text{wobei} \quad p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

- Problem: $G_{max} = \frac{n-1}{n}$

➠ Normierter Gini-Koeffizient:

$$G_* = \frac{n}{n-1} \cdot G \in [0;1]$$

Beispiel:

i	1	2	3	4	$\sum$
x_i	1	2	2	15	20
p_i	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{15}{20}$	1

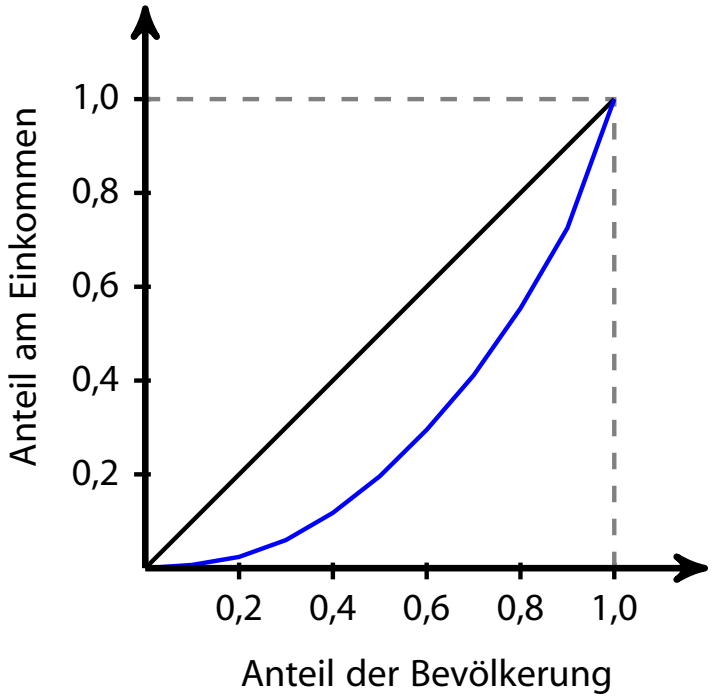
$$G = \frac{2 \cdot \left(1 \cdot \frac{1}{20} + 2 \cdot \frac{2}{20} + 3 \cdot \frac{2}{20} + 4 \cdot \frac{15}{20}\right) - (4+1)}{4} = 0{,}525$$

Mit $G_{max} = \frac{4-1}{4} = 0{,}75$ folgt

$$G_* = \frac{4}{4-1} \cdot 0{,}525 = 0{,}7$$

Armutsbericht der Bundesregierung 2008

- Verteilung der Bruttoeinkommen in Preisen von 2000
- aus unselbständiger Arbeit der Arbeitnehmer/-innen insgesamt

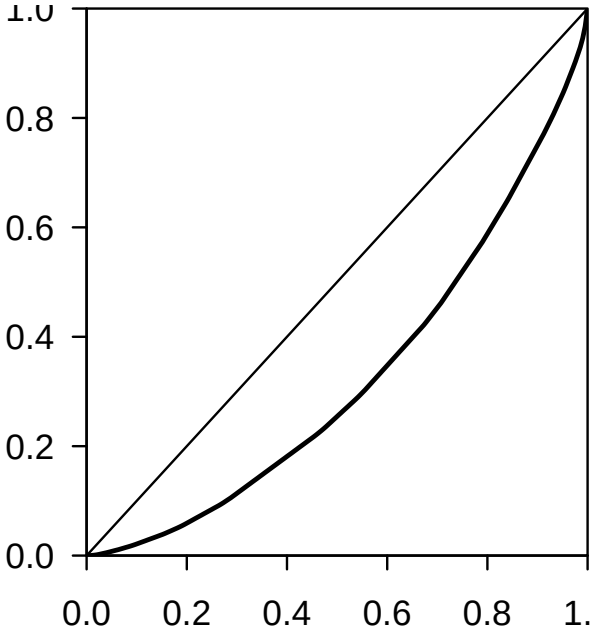


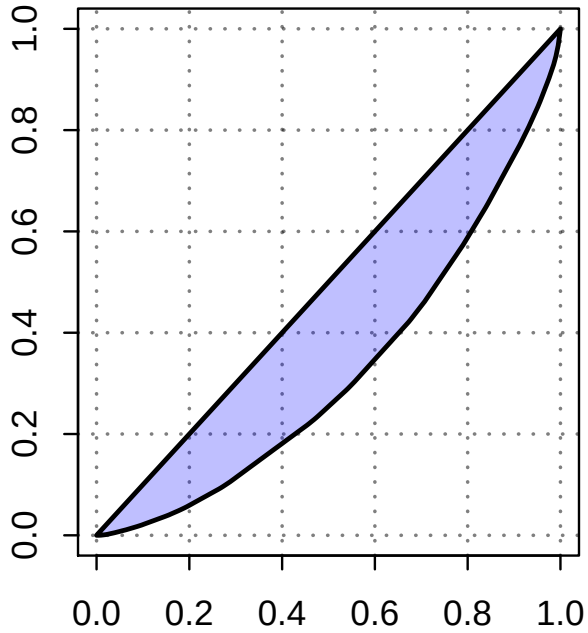
	2002	2003	2004	2005
Arithmetisches Mittel	24.873	24.563	23.987	23.648
Median	21.857	21.531	20.438	20.089
Gini-Koeffizient	0,433	0,441	0,448	0,453

Lorenzkurve mit R

```
require(ineq) # inequality Paket
Lorenz = Lc(AusgSchuhe)
plot(Lorenz, xlab="", ylab="", main="") # Standard plot

plot(c(0,1), c(0,1), type="n", # bisschen netter
    panel.first=grid(lwd=1.5, col=rgb(0,0,0,1/2)),
    xlab="", main="", ylab="")
polygon(Lorenz$p, Lorenz$L, density=-1, col=rgb(0,0,1,1/4), lwd=2)
```






```
Gini(AusgSchuhe) # Gini-Koeffizient

## [1] 0.3556353
```


Weitere Konzentrationsmaße

- Konzentrationskoeffizient:

$$CR_g = \text{Anteil, der auf die } g \text{ größten entfällt} = \sum_{i=n-g+1}^{n} p_i = 1 - \nu_{n-g}$$

- Herfindahl-Index:

$$H = \sum_{i=1}^{n} p_i^2 \qquad (\in [\tfrac{1}{n}; 1])$$

Es gilt: $H = \frac{1}{n}(V^2 + 1)$ bzw. $V = \sqrt{n \cdot H - 1}$

- Exponentialindex:

$$E = \prod_{i=1}^{n} p_i^{p_i} \qquad \left(\in [\tfrac{1}{n}; 1]\right) \qquad \text{wobei} \qquad 0^0 = 1$$

- Im Beispiel mit $x = (1, 2, 2, 15)$:

$$CR_2 = \tfrac{17}{20} = 0{,}85$$

$$H = \left(\frac{1}{20}\right)^2 + \cdots + \left(\frac{15}{20}\right)^2 = 0{,}59$$

$$E = \left(\frac{1}{20}\right)^{\frac{1}{20}} \cdots \left(\frac{15}{20}\right)^{\frac{15}{20}} = 0{,}44$$

Zweidimensionale Urliste

Urliste vom Umfang n zu **zwei** Merkmalen X und Y:

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

Kontingenztabelle:

Sinnvoll bei wenigen Ausprägungen bzw. bei klassierten Daten.

	Ausprägungen von Y			
Ausprägungen von X	b_1	b_2	$\ldots$	b_l
a_1	h_{11}	h_{12}	$\ldots$	h_{1l}
a_2	h_{21}	h_{22}	$\ldots$	h_{2l}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_k	h_{k1}	h_{k2}	$\ldots$	h_{kl}

Unterscheide:

- **Gemeinsame Häufigkeiten:**

$$h_{ij} = h(a_i, b_j)$$

- **Randhäufigkeiten:**

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

- **Bedingte (relative) Häufigkeiten:**

$$f_1(a_i \mid b_j) = \frac{h_{ij}}{h_{\cdot j}} \qquad \text{und} \qquad f_2(b_j \mid a_i) = \frac{h_{ij}}{h_{i\cdot}}$$

Beispiel: 400 unfallbeteiligte Autoinsassen:

	leicht verletzt $(= b_1)$	schwer verletzt $(= b_2)$	tot $(= b_3)$	
angegurtet $(= a_1)$	264 $(= h_{11})$	90 $(= h_{12})$	6 $(= h_{13})$	360 $(= h_{1.})$
nicht angegurtet $(= a_2)$	2 $(= h_{21})$	34 $(= h_{22})$	4 $(= h_{23})$	40 $(= h_{2.})$
	266 $(= h_{.1})$	124 $(= h_{.2})$	10 $(= h_{.3})$	400 $(= n)$

$f_2(b_3 \mid a_2) = \frac{4}{40} = 0{,}1$ (10 % der nicht angegurteten starben.)

$f_1(a_2 \mid b_3) = \frac{4}{10} = 0{,}4$ (40 % der Todesopfer waren nicht angegurtet.)

Beispiel:

i	1	2	3	4	5	$\sum$
x_i	2	4	3	9	7	25
y_i	4	3	6	7	8	28

$$\Rightarrow \bar{x} = \frac{25}{5} = 5$$
$$\bar{y} = \frac{28}{5} = 5{,}6$$

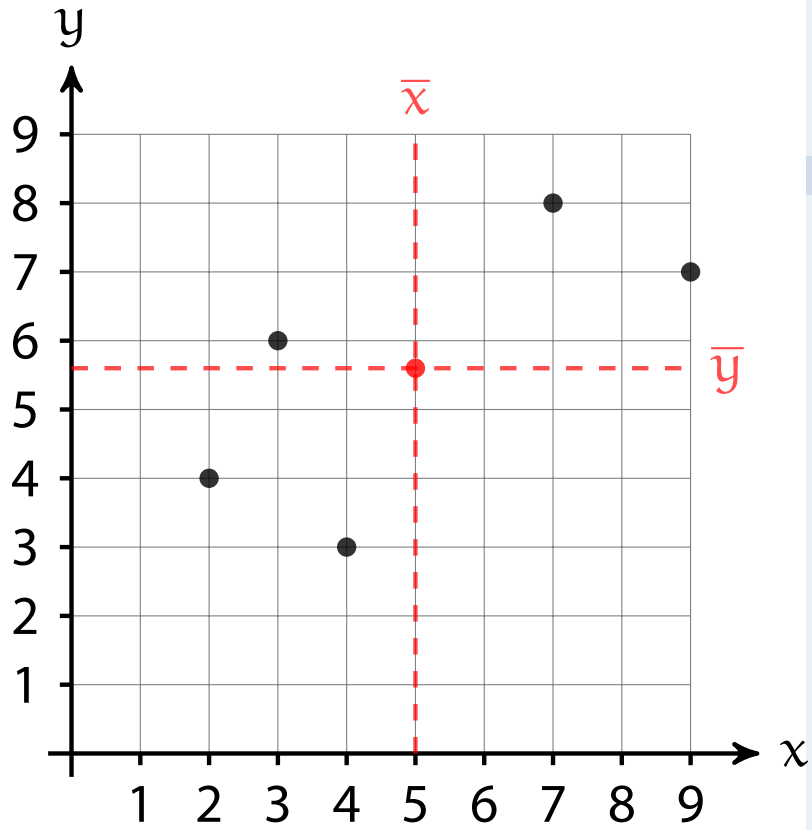
Streuungsdiagramm sinnvoll bei vielen verschiedenen Ausprägungen (z.B. stetige Merkmale)

➠ Alle (x_i, y_i) sowie $(\bar{x}, \bar{y})$ in Koordinatensystem eintragen.

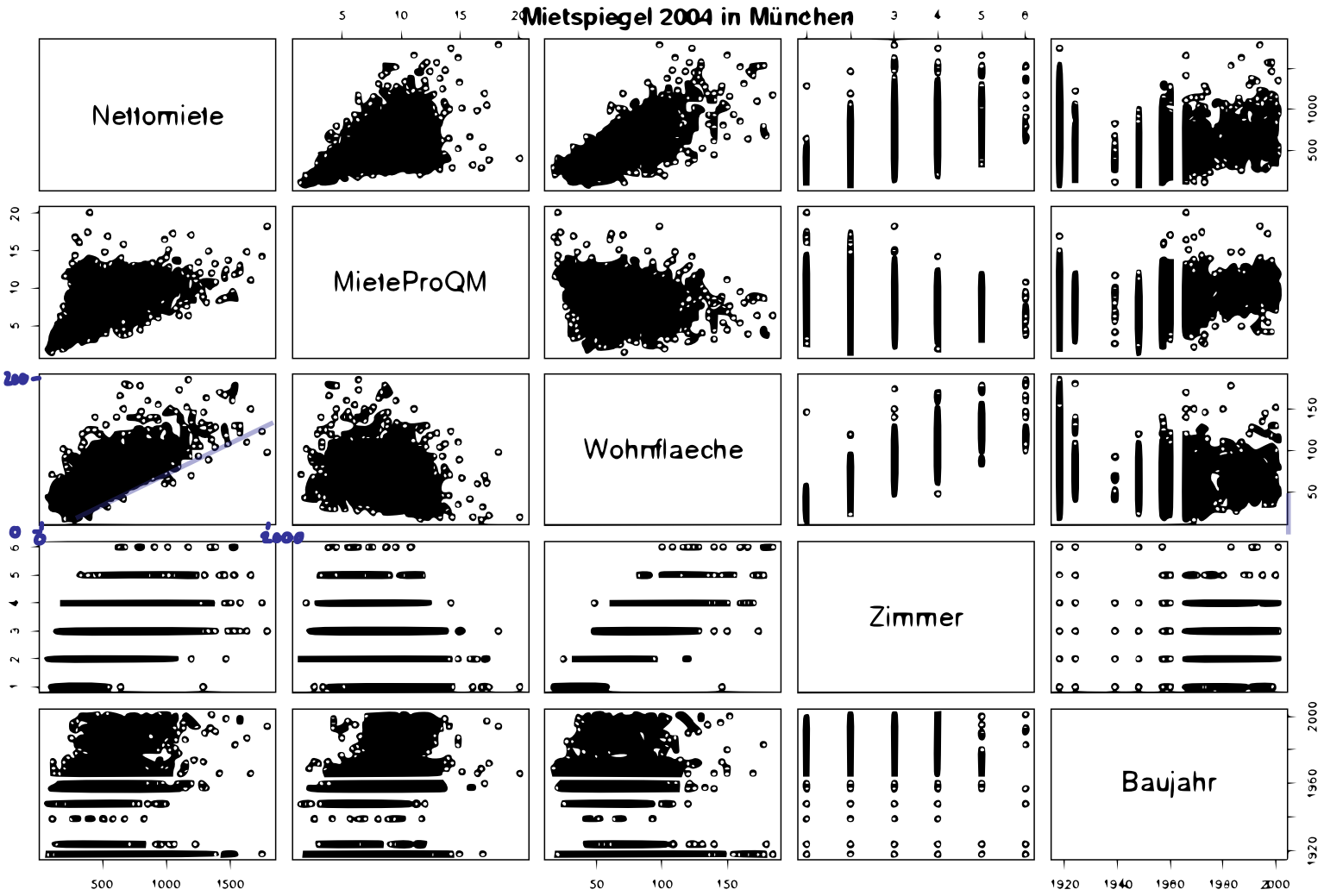
Beispiel:

i	1	2	3	4	5	$\sum$
x_i	2	4	3	9	7	25
y_i	4	3	6	7	8	28

$$\Rightarrow \bar{x} = \frac{25}{5} = 5$$
$$\bar{y} = \frac{28}{5} = 5{,}6$$

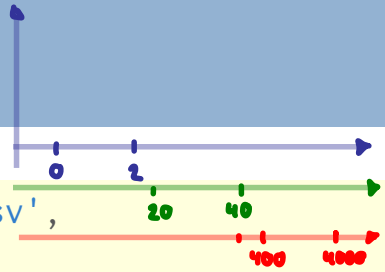


Beispiel Streuungsdiagramm



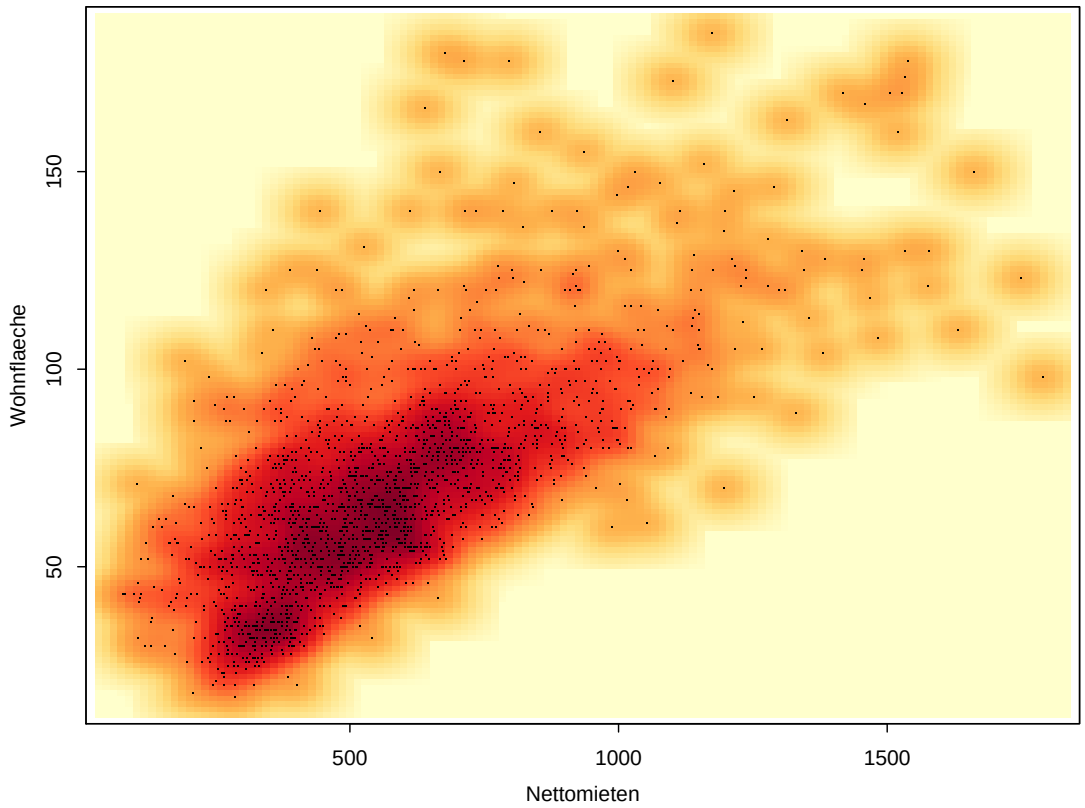
(Datenquelle: Fahrmeir u. a. (2009))

Beispiel Streuungsdiagramm



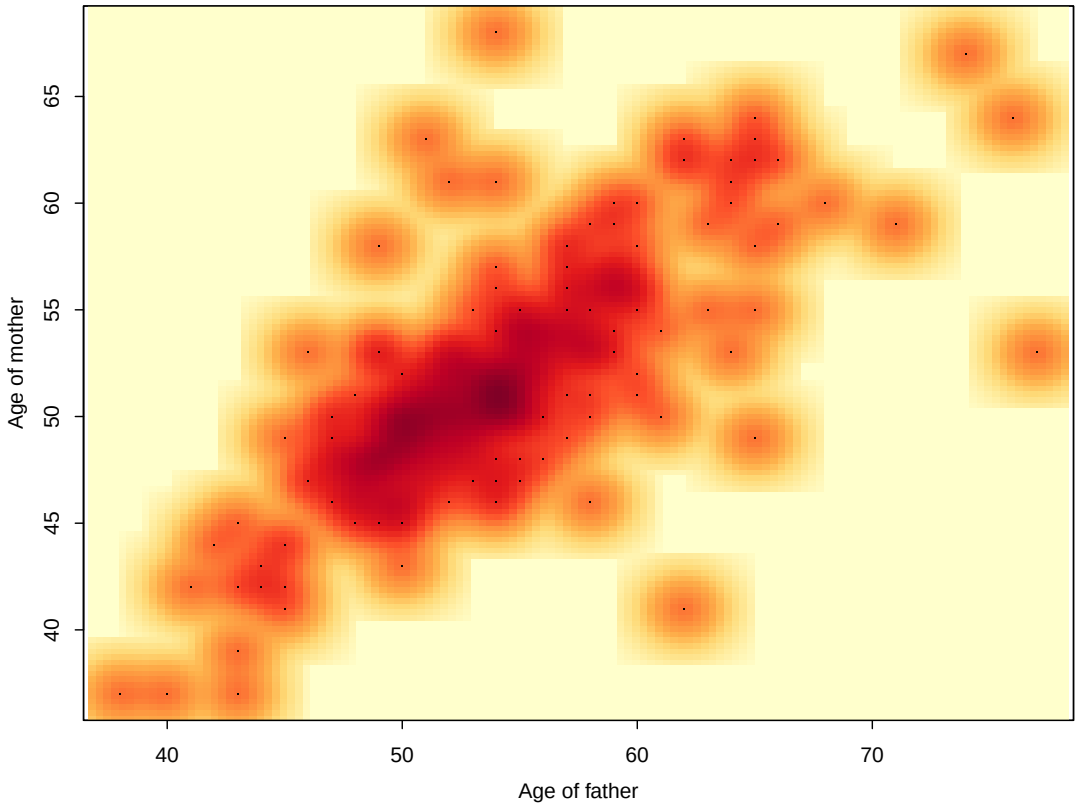

```
mieten <- read.table('../../_data/MietenMuenchen.csv',
                     header=TRUE, sep='\t',
                     check.names=TRUE, fill=TRUE,
                     na.strings=c('',''))
x   <- cbind(Nettomieten=mieten$nm, Wohnflaeche=mieten$wfl)

library("geneplotter")  ## from BioConductor
smoothScatter(x, nrpoints=Inf,
              colramp=colorRampPalette(brewer.pal(9,"YlOrRd")),
              bandwidth=c(30,3))
```



Beispiel Streuungsdiagramm

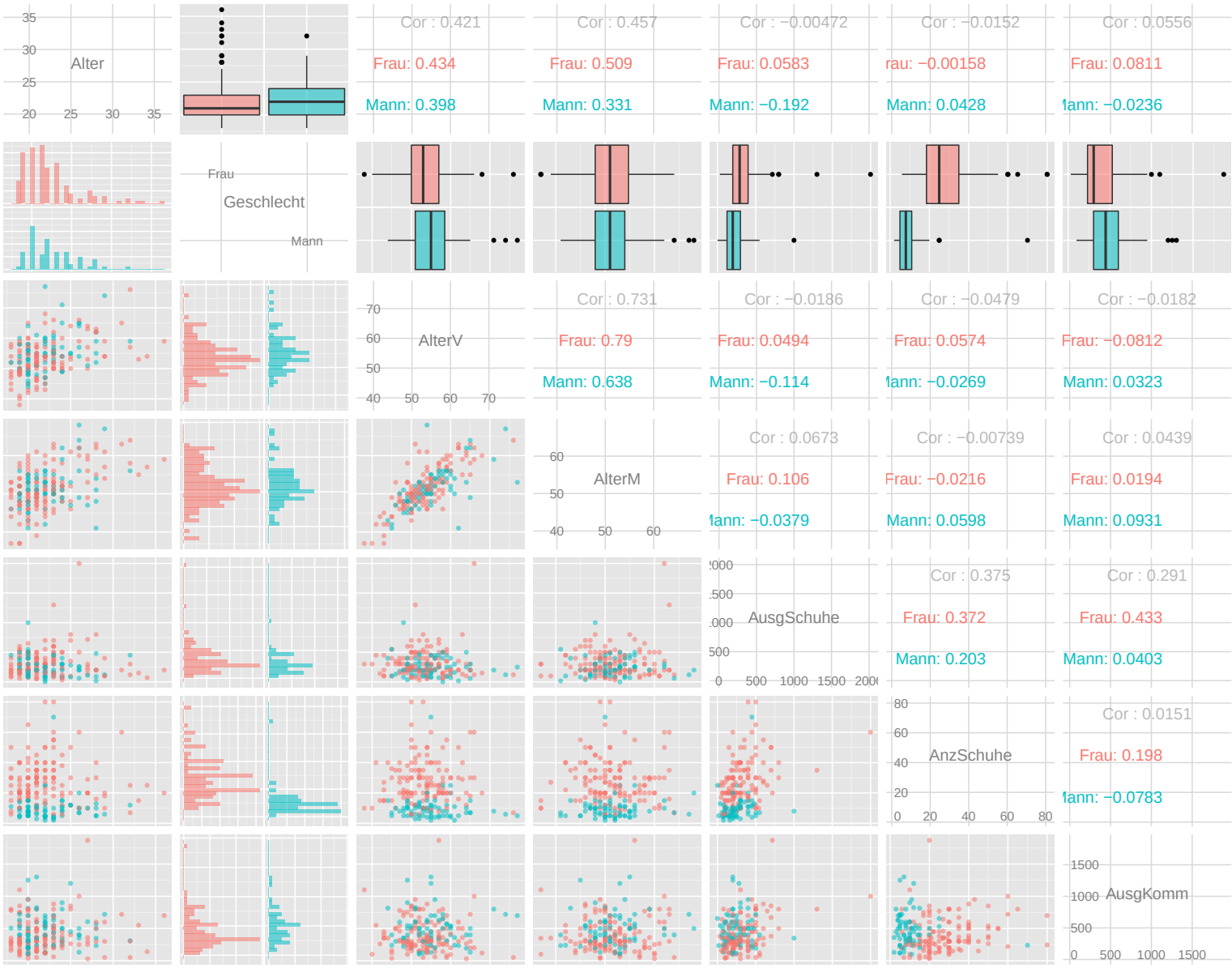
```
x = cbind("Age of father"=AlterV, "Age of mother"=AlterM)
require("geneplotter") ## from BioConductor
smoothScatter(x, colramp=colorRampPalette(brewer.pal(9,"YlOrRd")) )
```



Visualisiere Paare

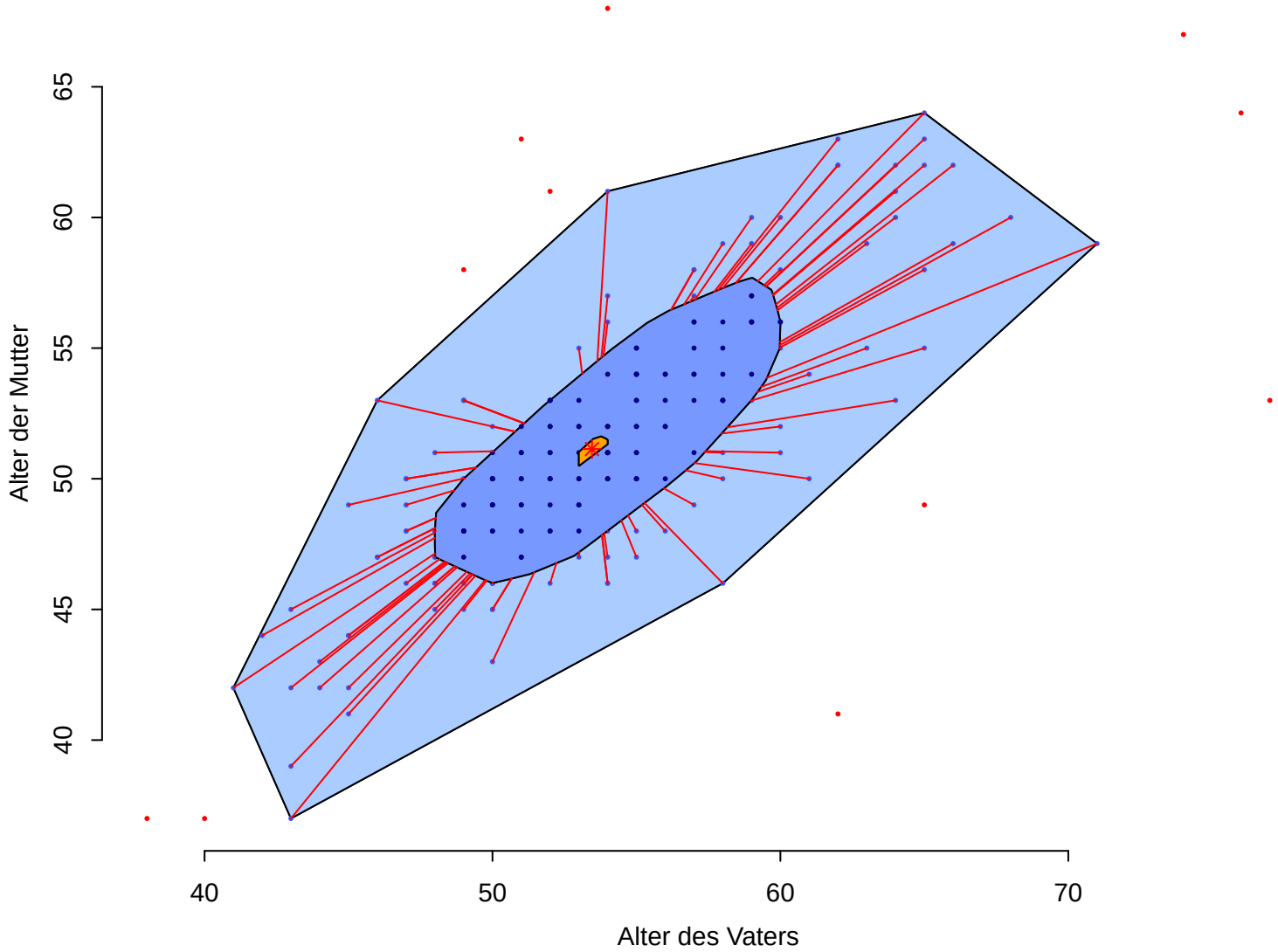
$\frac{1}{2}n(n-1)$

```
require(GGally)
ggpairs(MyData[, -c(5, 6, 10)], colour='Geschlecht', alpha=0.4)
```



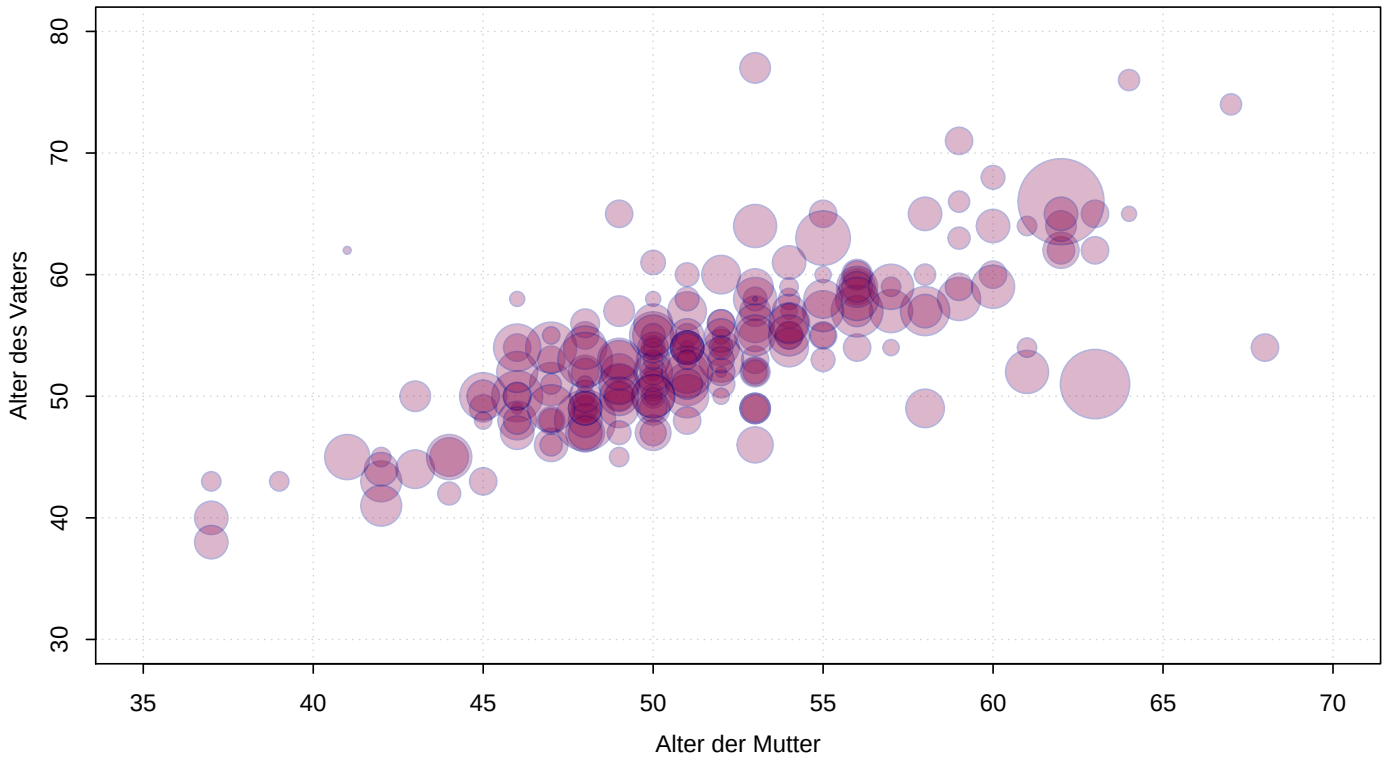
Bagplot: Boxplot in 2 Dimensionen

```
require(aplpack)
bagplot(AlterV, AlterM, xlab="Alter des Vaters", ylab="Alter der Mutter")
```



Bubbleplot: 3 metrische Variablen


```
require(DescTools)
PlotBubble(AlterM, AlterV, AusgSchuhe/400,
col=SetAlpha("deeppink4",0.3),
border=SetAlpha("darkblue",0.3),
xlab="Alter der Mutter", ylab="Alter des Vaters",
panel.first=grid(),
main="")
```



Größe der Blasen: Ausgaben für Schuhe

- Frage: Wie stark ist der Zusammenhang zwischen X und Y?
- Dazu: Korrelationskoeffizienten
- Verschiedene Varianten: Wahl abhängig vom Skalenniveau von X und Y:

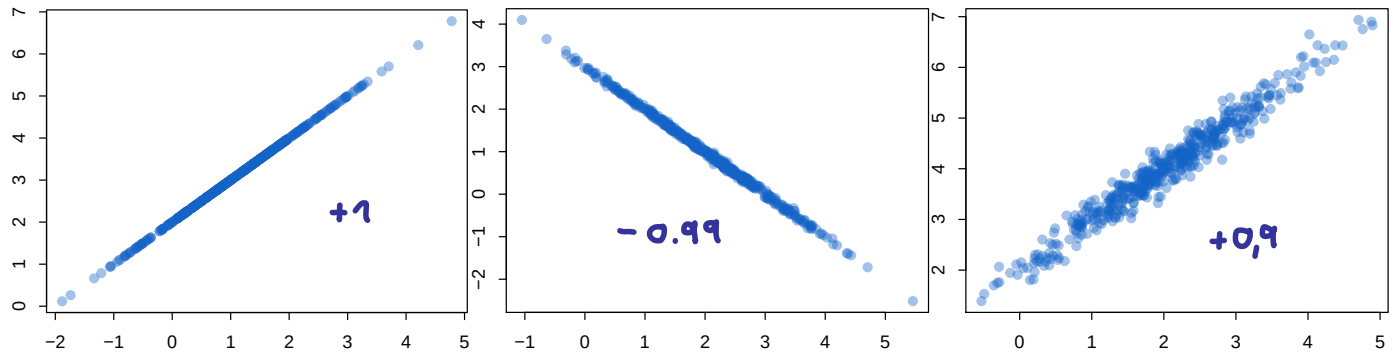
	Skalierung von Y		
Skalierung von X	kardinal	ordinal	nominal
kardinal	Bravais-Pearson-Korrelations-koeffizient		
ordinal		Rangkorrelations-koeffizient von Spearman	
nominal			Kontingenz-koeffizient

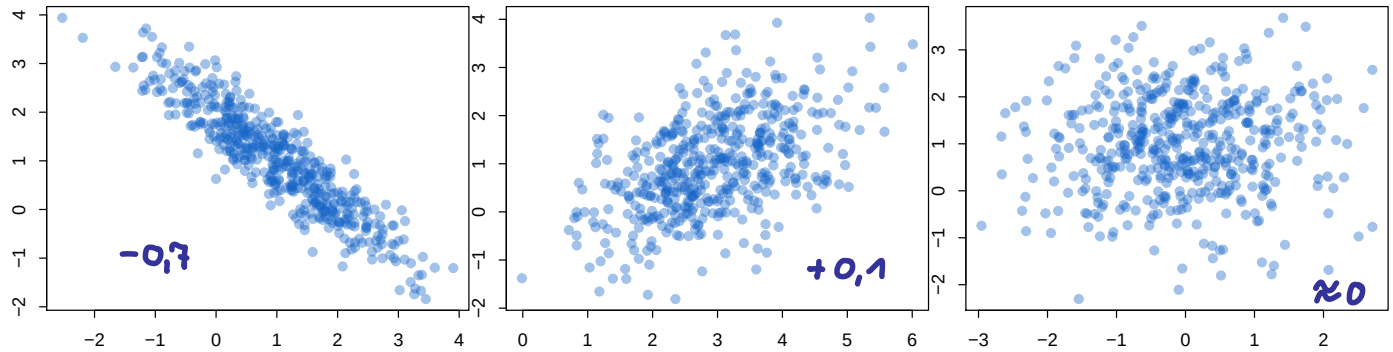
Korrelationskoeffizient von Bravais und Pearson

Bravais-Pearson-Korrelationskoeffizient

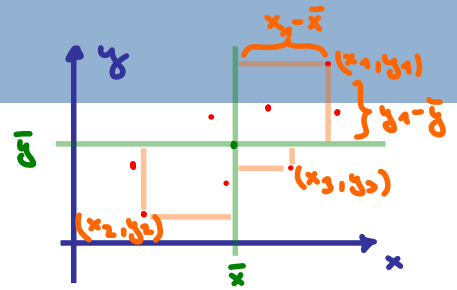
Voraussetzung: X, Y kardinalskaliert

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \in [-1;+1]$$





Korrelationskoeffizient von Bravais und Pearson



$(x_1-\bar{x})(y_1-\bar{y})>0$

$\underbrace{(x_2-\bar{x})}_{<0}\underbrace{(y_2-\bar{y})}_{<0}>0$

$\underbrace{(x_3-\bar{x})}_{>0}\underbrace{(y_3-\bar{y})}_{<0}<0$

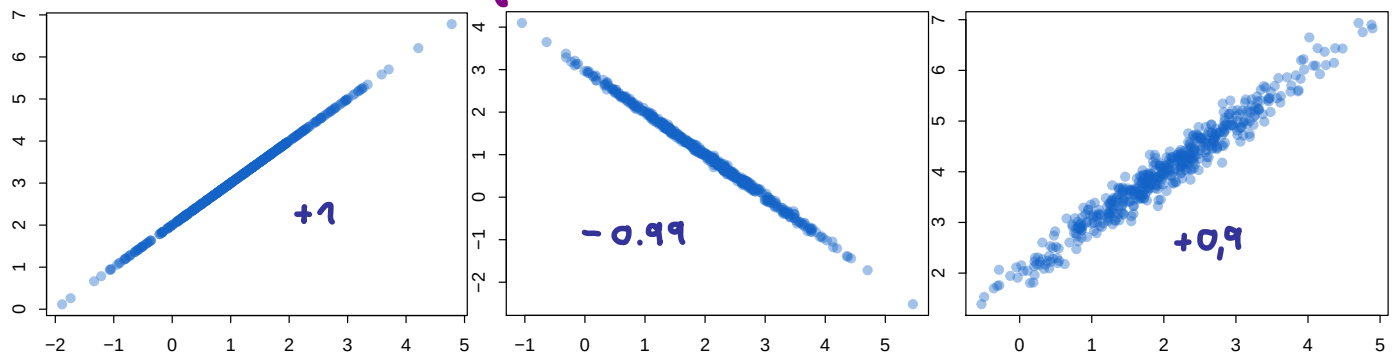
Bravais-Pearson-Korrelationskoeffizient

Voraussetzung: X, Y kardinalskaliert

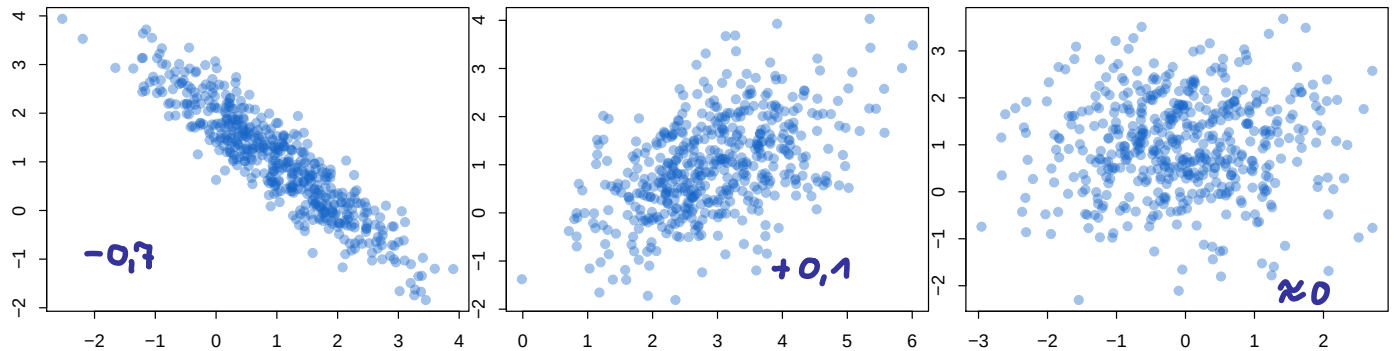
$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}} \in [-1;+1]$$

Kovarianz

s_x s_y



$r = 1$ $\quad r = -0{,}999$ $\quad r = 0{,}979$



$r = -0{,}915$ $\quad r = 0{,}493$ $\quad r = 0{,}043$

Bravais-Pearson-Korrelationskoeffizient

Im Beispiel:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	4	4	16	8
2	4	3	16	9	12
3	3	6	9	36	18
4	9	7	81	49	63
5	7	8	49	64	56
$\sum$	25	28	159	174	157

$$\Rightarrow \quad \bar{x} = 25/5 = 5 \qquad \bar{y} = 28/5 = 5{,}6$$

$$r = \frac{157 - 5 \cdot 5 \cdot 5{,}6}{\sqrt{159 - 5 \cdot 5^2}\sqrt{174 - 5 \cdot 5{,}6^2}} = 0{,}703$$

(deutliche positive Korrelation)

Rangkorrelation

Beispiel:

$R_{Zuf.}$	Zufrieden?		Preis		R_{Preis}
1,5	1	sehr	11,99	2	2
4	3	zufrieden	9,99	3	3
1,5	2	sehr	19,99	1	1
6	6	geht so	7,99	5	5,5
7	7	gar nicht	7,00	7	7
4	4	zufrieden	8,00	4	4
4	5	zufrieden	7,99	6	5,5

Rang Korrelation $\hat{=}$ Bravais-Pearson der Ränge beider Merkmale

Rangkorrelationskoeffizient von Spearman

- Voraussetzungen: X, Y (mindestens) ordinalskaliert, Ränge eindeutig (keine Doppelbelegung von Rängen)
- Vorgehensweise:
  1. Rangnummern R_i (X) bzw. R'_i (Y) mit $R_i^{(\prime)} = 1$ bei größtem Wert usw.
  2. Berechne

$$r_{SP} = 1 - \frac{6 \sum_{i=1}^{n} (R_i - R'_i)^2}{(n-1)\, n\, (n+1)} \in [-1; +1]$$

- Hinweise:
  - $r_{SP} = +1$ wird erreicht bei $R_i = R'_i \qquad \forall\, i = 1, \ldots, n$
  - $r_{SP} = -1$ wird erreicht bei $R_i = n + 1 - R'_i \quad \forall\, i = 1, \ldots, n$

Im Beispiel:

x_i	R_i	y_i	R'_i
2	5	4	4
4	3	3	5
3	4	6	3
9	1	7	2
7	2	8	1

$$r_{SP} = 1 - \frac{6 \cdot [(5-4)^2 + (3-5)^2 + (4-3)^2 + (1-2)^2 + (2-1)^2]}{(5-1) \cdot 5 \cdot (5+1)} = 0{,}6$$

Kontingenzkoeffizient

- Gegeben: Kontingenztabelle mit k Zeilen und l Spalten (vgl. hier)
- Vorgehensweise:
  1. Ergänze Randhäufigkeiten

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

  2. Berechne theoretische Häufigkeiten

$$\tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}$$

  3. Berechne

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

χ^2 hängt von n ab! $(h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2)$

④ Kontingenzkoeffizient:

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} \qquad \in [0; K_{max}]$$

wobei

z.B. $\min\{2,3\} = 2$
$\Rightarrow K_{max} = \sqrt{\frac{2-1}{2}} \approx 0{,}71$

$$K_{max} = \sqrt{\frac{M-1}{M}} \qquad \text{mit} \qquad M = \min\{k, l\}$$

⑤ Normierter Kontingenzkoeffizient:

$$K_* = \frac{K}{K_{max}} \qquad \in [0; 1]$$

$K_* = +1 \iff$
bei Kenntnis von x_i kann y_i erschlossen werden u.u.

Beispiel

X : Staatsangehörigkeit (d,a)
Y : Geschlecht (m,w)

h_{ij}	m	w	$h_{i\cdot}$
d	30	30	60
a	10	30	40
$h_{\cdot j}$	40	60	100

$\Rightarrow$

$\tilde{h}_{ij}$	m	w
d	24	36
a	16	24

wobei $\tilde{h}_{11} = \frac{60 \cdot 40}{100} = 24$ usw.

$$\chi^2 = \frac{(30-24)^2}{24} + \frac{(30-36)^2}{36} + \frac{(10-16)^2}{16} + \frac{(30-24)^2}{24} = 6{,}25$$

$$K = \sqrt{\frac{6{,}25}{100+6{,}25}} = 0{,}2425; \quad M = \min\{2,2\} = 2; \quad K_{max} = \sqrt{\frac{2-1}{2}} = 0{,}7071$$

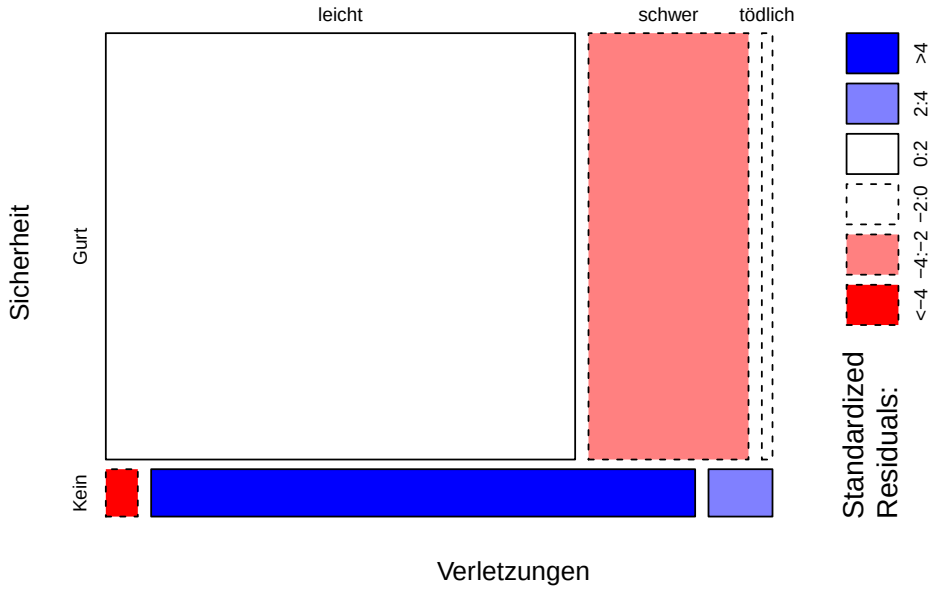
$$K_* = \frac{0{,}2425}{0{,}7071} = 0{,}3430$$

Beispiel Autounfälle

$$\chi^2 = \frac{(264-239{,}4)^2}{239{,}4} + \frac{(90-111{,}6)^2}{111{,}6} + \ldots + \frac{(4-1)^2}{1}$$

	Verletzung			
	leicht	schwer	tödlich	
angegurtet	264 (239,4)	90 (111,6)	6 (9)	360
nicht angegurtet	2 (26,6)	34 (12,4)	4 (1)	40
	266	124	10	400

$\frac{40}{400} \cdot 10$



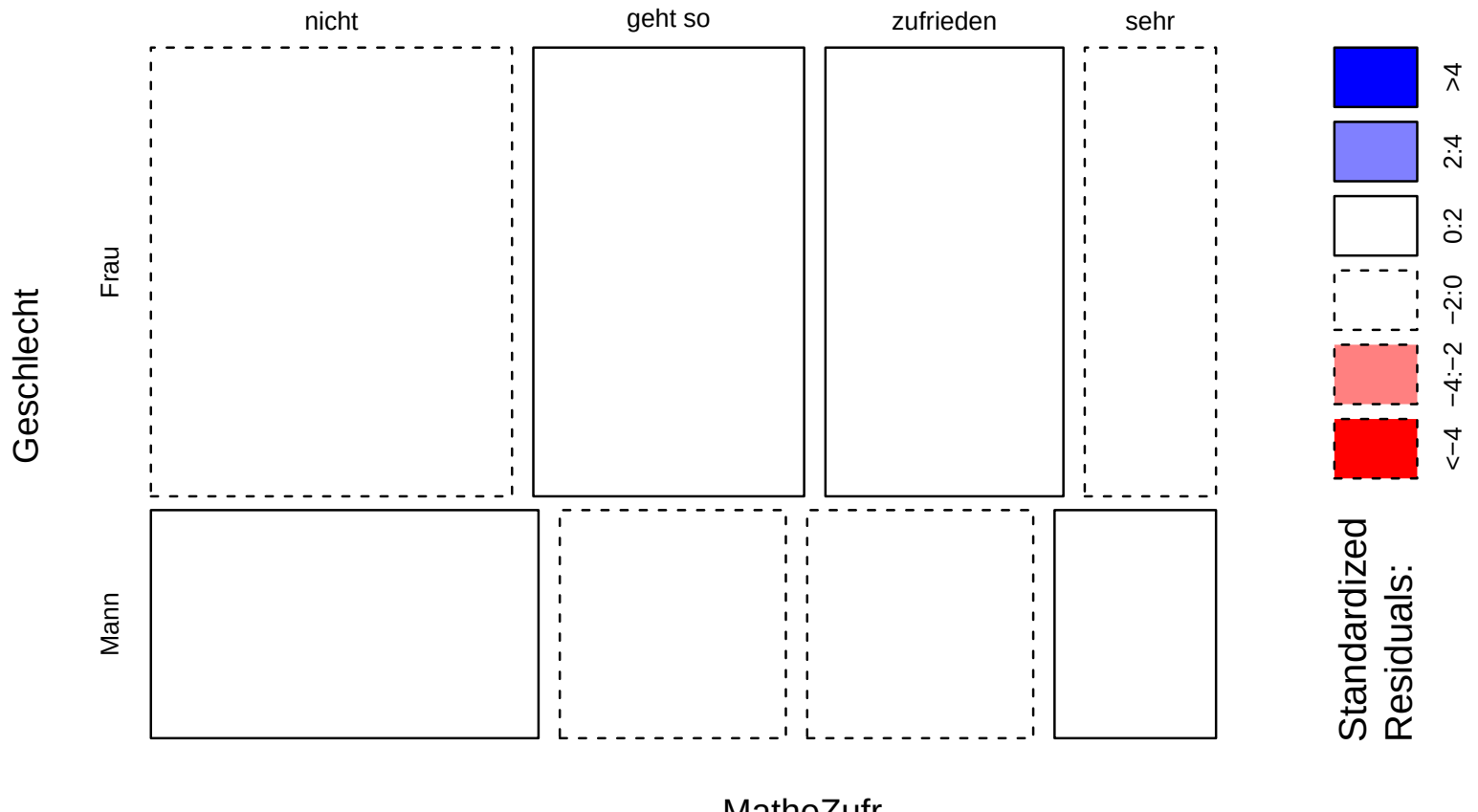
Mosaikplot Autounfälle

Mosaicplot Geschlecht, Zufriedenheit über Note in Matheklausur

```
tab = table(MatheZufr,Geschlecht)
tab
```

```
##            Geschlecht
## MatheZufr   Frau Mann
##   nicht       44   24
##   geht so     33   14
##   zufrieden   29   14
##   sehr        16   10
```

```
mosaicplot(tab, shade = TRUE, sort=2:1, main="")
```

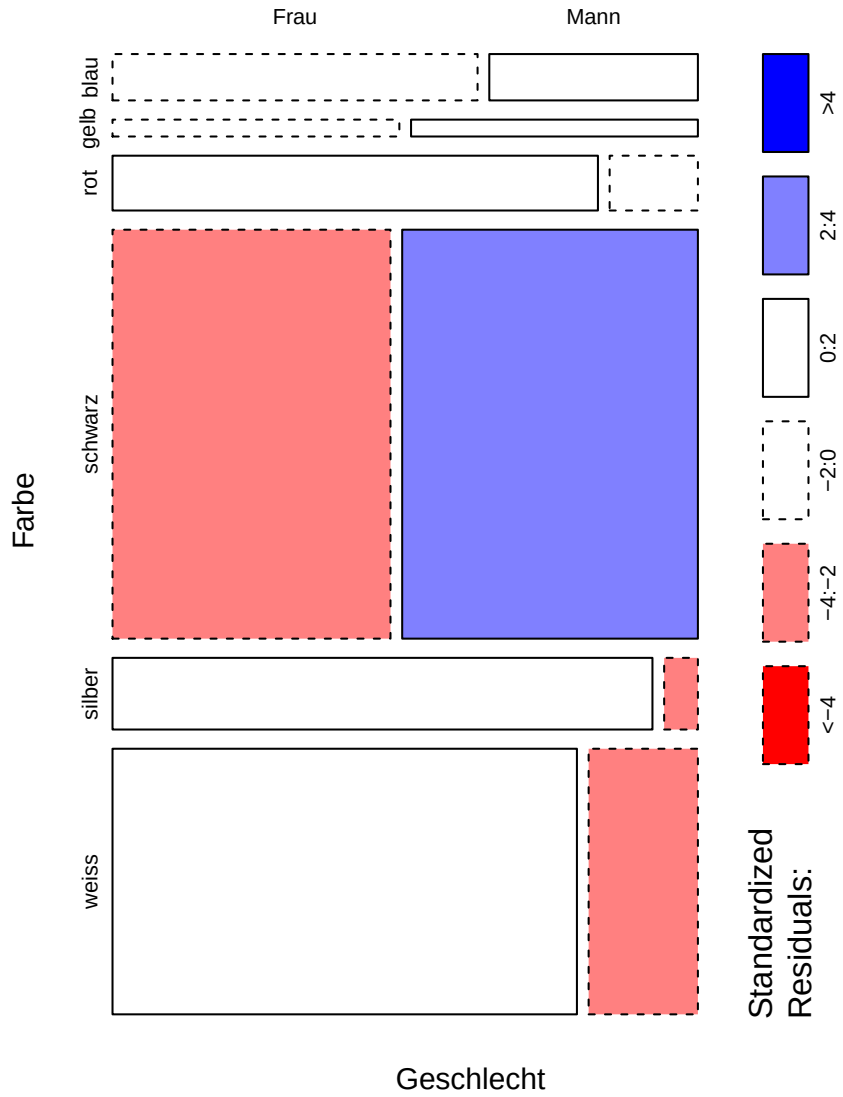


Mosaicplot Geschlecht, Wunschfarbe für Smartphone

```
tab = table(Farbe, Geschlecht)
tab
```

```
##          Geschlecht
## Farbe     Frau Mann
##   blau       7    4
##   gelb       2    2
##   rot       11    2
##   schwarz   47   50
##   silber    16    1
##   weiss     51   12
```

```
mosaicplot(t(tab), shade = TRUE,
           sort=2:1, main="")
```

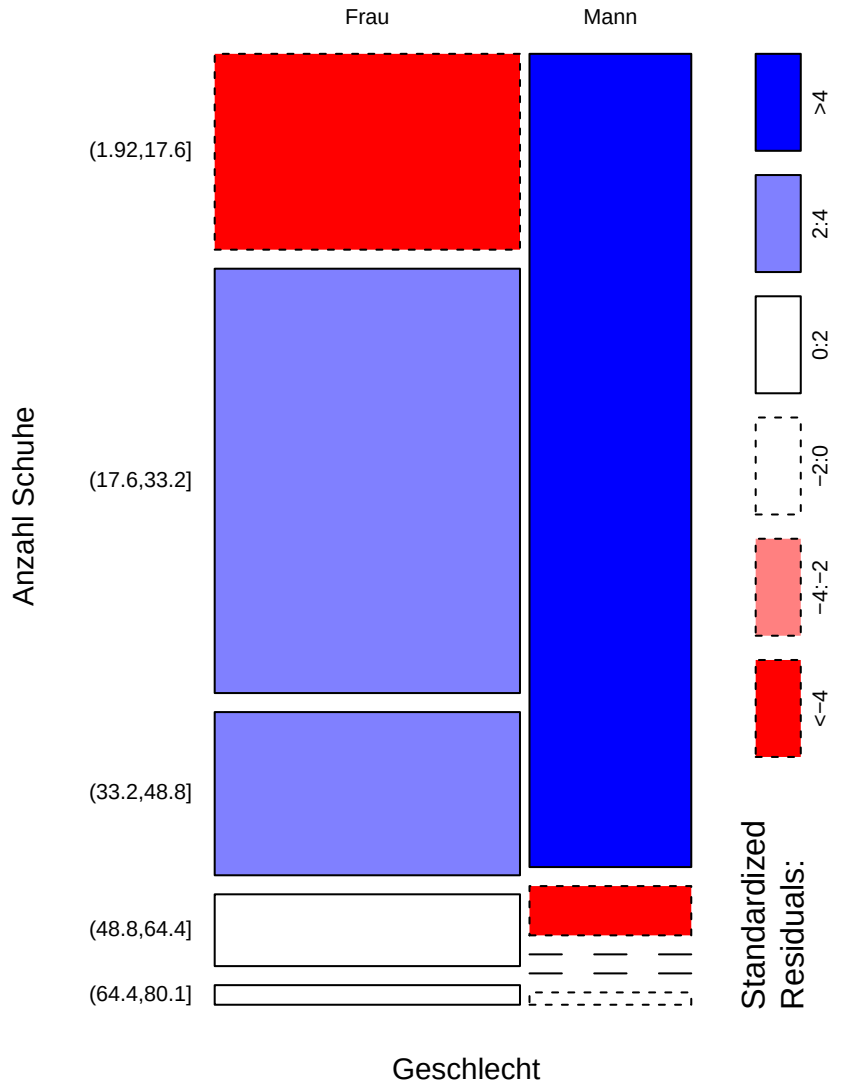


Mosaicplot Geschlecht, Anzahl Schuhe

```
tab = table(
  "Anzahl Schuhe"=cut(AnzSchuhe, 5),
  Geschlecht)
tab
```

```
##                Geschlecht
## Anzahl Schuhe Frau Mann
##   (1.92,17.6]   30   66
##   (17.6,33.2]   65    4
##   (33.2,48.8]   25    0
##   (48.8,64.4]   11    0
##   (64.4,80.1]    3    1
```

```
mosaicplot(t(tab), shade = TRUE,
           main="", las=1)
```



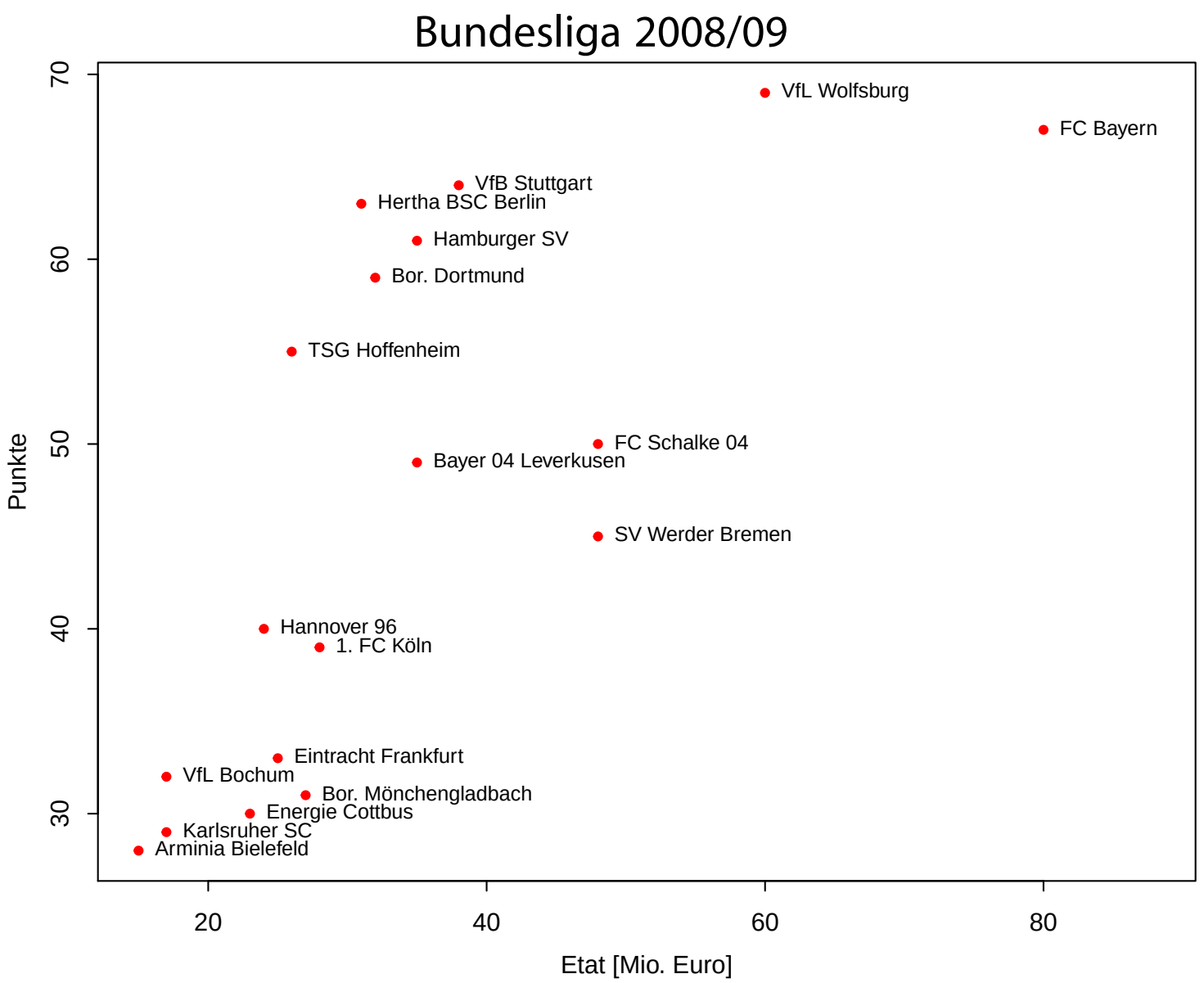
Bundesliga 2008/2009

- Gegeben: Daten zu den 18 Vereinen der ersten Bundesliga in der Saison 2008/09
- Merkmale: **Vereinssetat** für Saison (nur direkte Gehälter und Spielergehälter)
- und **Ergebnispunkte** in Tabelle am Ende der Saison

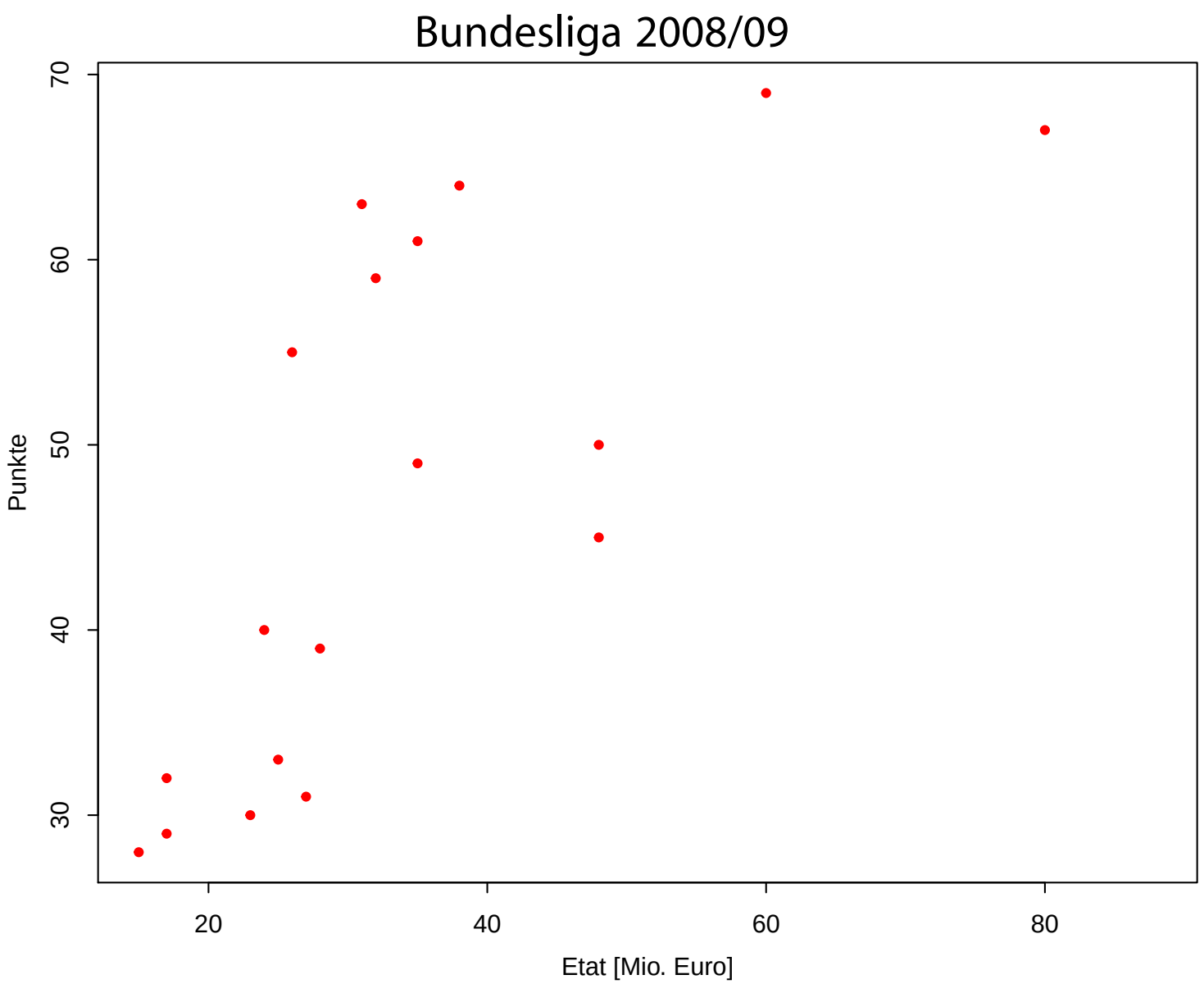
	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96	24	40
Energie Cottbus	23	30
VfL Bochum	17	32
Karlsruher SC	17	29
Arminia Bielefeld	15	28

(Quelle: Welt)

Darstellung der Daten in Streuplot



Darstellung der Daten in Streuplot



Trend als lineares Modell

- Kann man die **Tabellenpunkte** näherungsweise über einfache Funktion **in Abhängigkeit des Vereinsetats** darstellen?
- Allgemein: Darstellung einer Variablen Y als Funktion von X:

$$y = f(x)$$

- Dabei:
  - X heißt Regressor bzw. unabhängige Variable
  - Y heißt Regressand bzw. abhängige Variable
- Wichtiger (und einfachster) Spezialfall: f beschreibt einen linearen Trend:

$$y = a + b\,x$$

- Dabei anhand der Daten zu schätzen: a (Achsenabschnitt) und b (Steigung)
- Schätzung von a und b: Lineare Regression

Fehlerquadratsumme

- Pro Datenpunkt gilt mit Regressionsmodell:

$$y_i = a + bx_i + \epsilon_i$$

- Dabei: ϵ_i ist jeweils Fehler (der Grundgesamtheit),
- mit $e_i = y_i - (\hat{a} + \hat{b}x_i)$: Abweichung (Residuen) zwischen gegebenen Daten der Stichprobe und durch Modell geschätzten Werten
- Modell gut wenn alle Residuen e_i zusammen möglichst klein
- Einfache Summe aber nicht möglich, denn e_i positiv oder negativ
- Deswegen: Summe der Quadrate von e_i
- Prinzip der kleinsten Quadrate: Wähle a und b so, dass

$$Q(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \rightarrow \min$$

Beste Lösung

- Beste und eindeutige Lösung:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

- Regressionsgerade:

$$\hat{y} = \hat{a} + \hat{b}\,x$$

Bundesligabeispiel

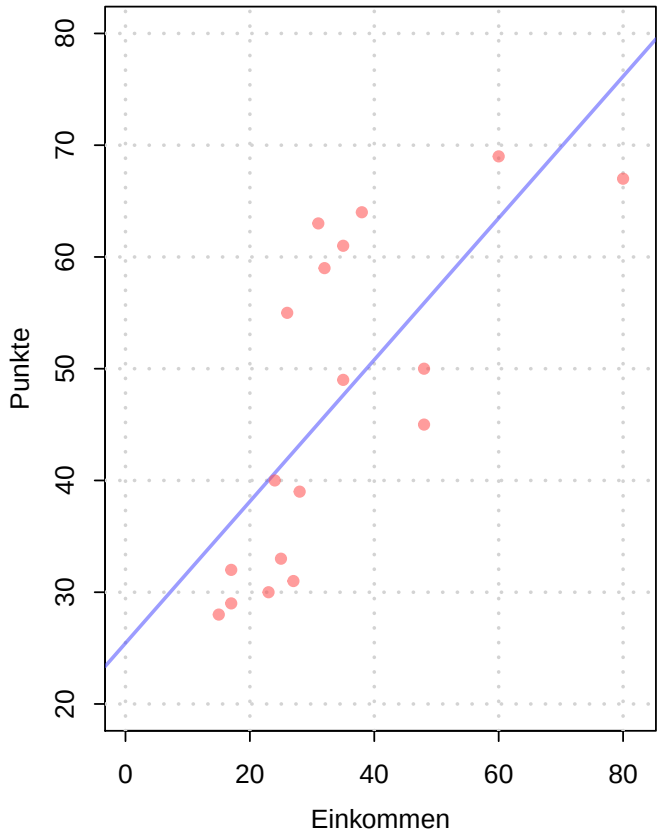
- Berechnung eines linearen Modells der Bundesligadaten
- dabei: Punkte $\hat{=} y$ und Etat $\hat{=} x$:

$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
n	18

$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33{,}83 \cdot 46{,}89}{25209 - 18 \cdot 33{,}83^2} \approx 0{,}634$$

$$\Rightarrow \hat{a} = 46{,}89 - \hat{b} \cdot 33{,}83 \approx 25{,}443$$

- Modell: $\hat{y} = 25{,}443 + 0{,}634 \cdot x$

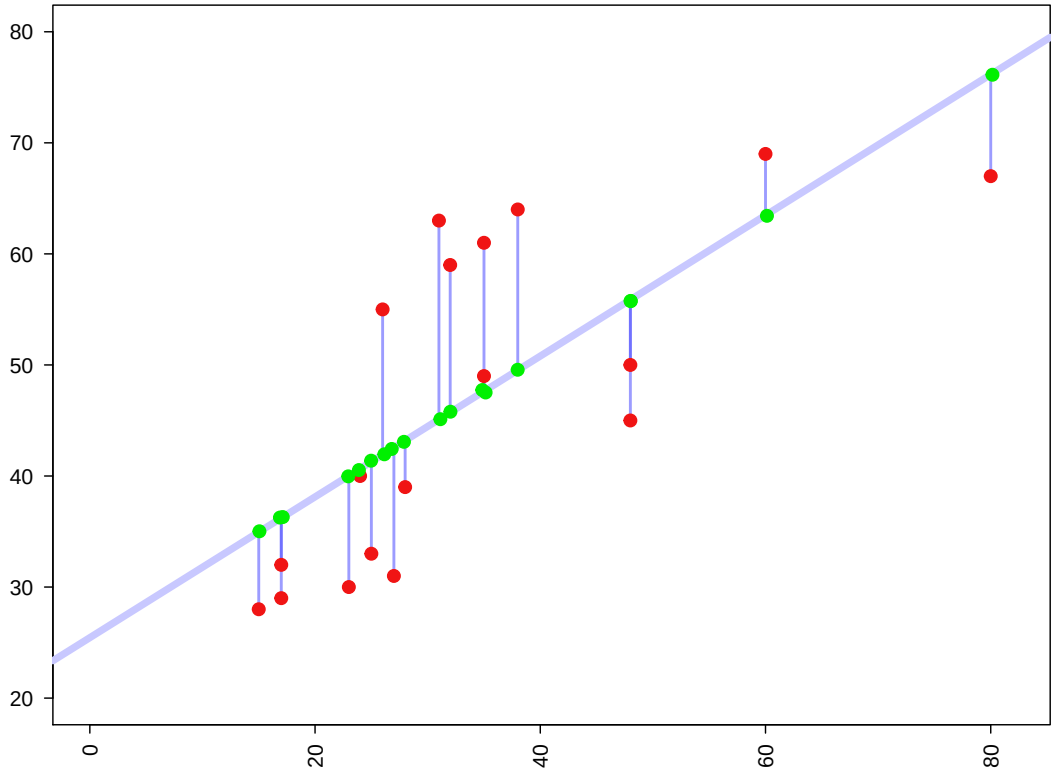


- Prognosewert für Etat $= 30$:

$$\hat{y}(30) = 25{,}443 + 0{,}634 \cdot 30 \approx 44{,}463$$

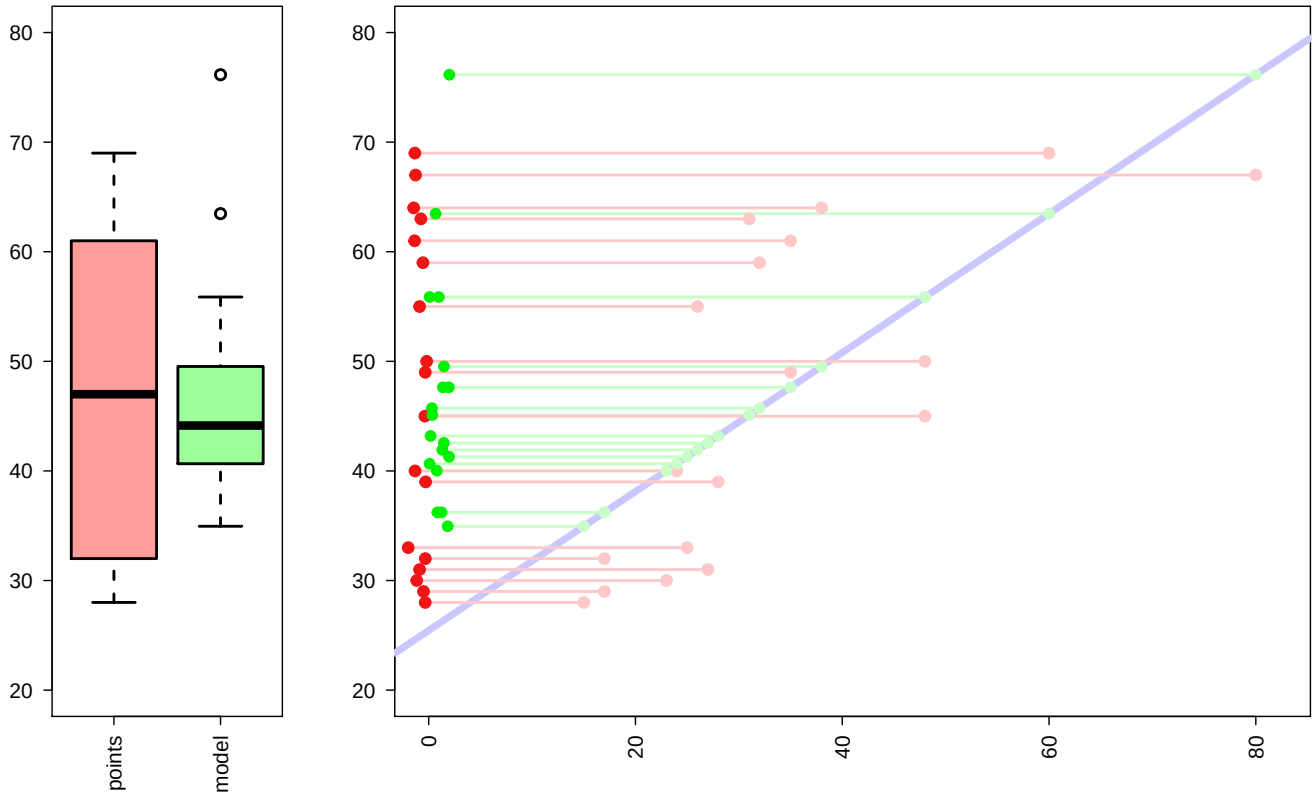
Varianz und Information

- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Information**sgehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden



Varianz und Information

- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Informations**gehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden



- Empirische Varianz (mittlere quadratische Abweichung) für „rot“ bzw. „grün“ ergibt jeweils

$$\frac{1}{18}\sum_{i=1}^{18}(y_i-\bar{y})^2 \approx 200{,}77 \qquad \text{bzw.} \qquad \frac{1}{18}\sum_{i=1}^{18}(\hat{y}_i-\bar{y})^2 \approx 102{,}78$$

Determinationskoeffizient

- Gütemaß für die Regression: Determinationskoeffizient (Bestimmtheitskoeffizient):

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2} = r^2 \in [0;1]$$

- Mögliche Interpretation von R^2:
  Durch die Regression erklärter Anteil der Varianz
- $R^2 = 0$ wird erreicht wenn X, Y unkorreliert
  $R^2 = 1$ wird erreicht wenn $\hat{y}_i = y_i \; \forall \, i$ (alle Punkte auf Regressionsgerade)
- Im (Bundesliga-)Beispiel:

$$R^2 = \frac{\sum_{i=1}^{18} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{18} (y_i - \bar{y})^2} \approx \frac{102{,}78}{200{,}77} \approx 51{,}19\,\%$$

Regression: 4 eindimensionale Beispiele

- Berühmte Daten aus den 1970er Jahren:

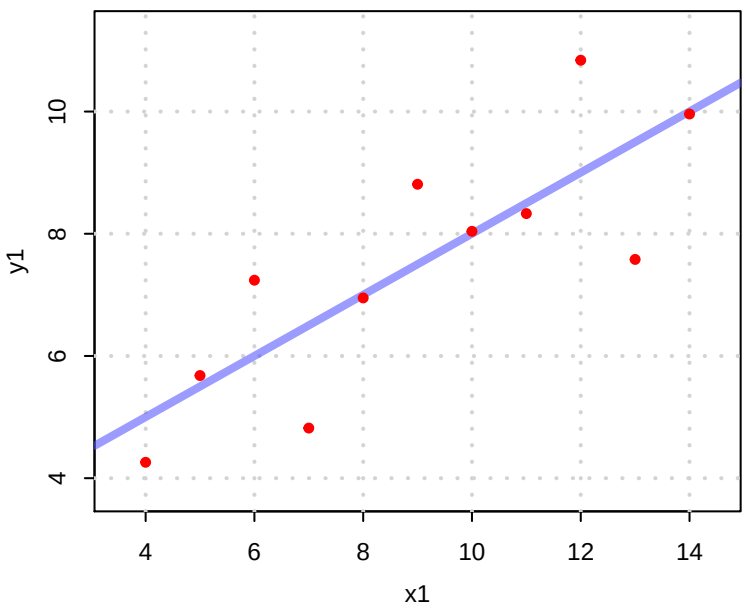
i	x_{1i}	x_{2i}	x_{3i}	x_{4i}	y_{1i}	y_{2i}	y_{3i}	y_{4i}
1	10	10	10	8	8,04	9,14	7,46	6,58
2	8	8	8	8	6,95	8,14	6,77	5,76
3	13	13	13	8	7,58	8,74	12,74	7,71
4	9	9	9	8	8,81	8,77	7,11	8,84
5	11	11	11	8	8,33	9,26	7,81	8,47
6	14	14	14	8	9,96	8,10	8,84	7,04
7	6	6	6	8	7,24	6,13	6,08	5,25
8	4	4	4	19	4,26	3,10	5,39	12,50
9	12	12	12	8	10,84	9,13	8,15	5,56
10	7	7	7	8	4,82	7,26	6,42	7,91
11	5	5	5	8	5,68	4,74	5,73	6,89

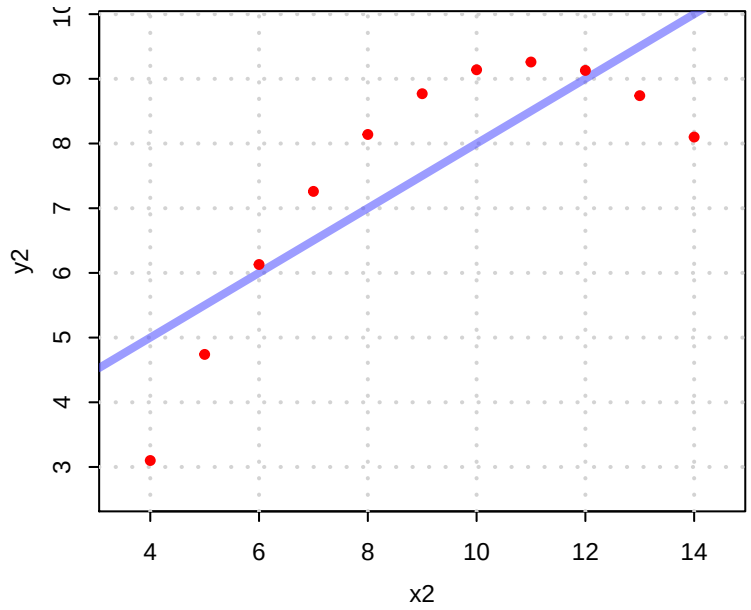
(Quelle: Anscombe (1973))

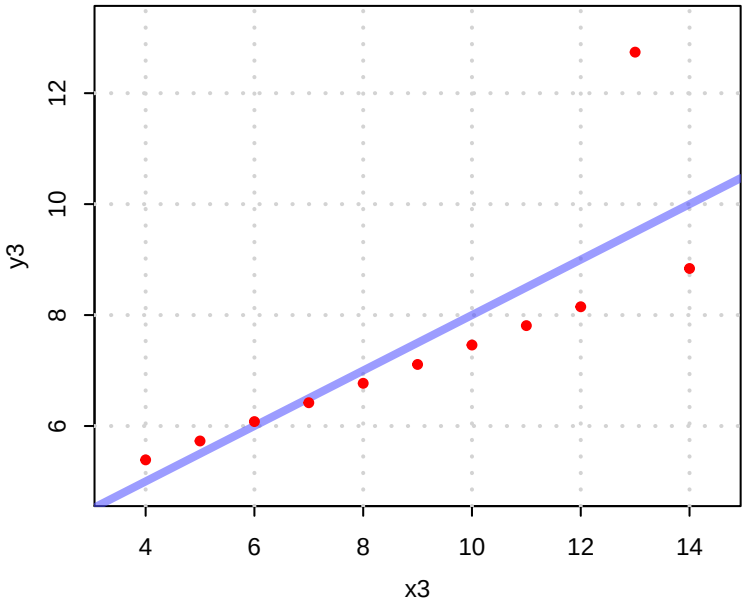
- In folgender Tabelle: Jeweils Ergebnisse der linearen Regressionsanalyse
- dabei: x_k unabhängige Variable und y_k abhängige Variable
- Modell jeweils: $y_k = a_k + b_k x_k$

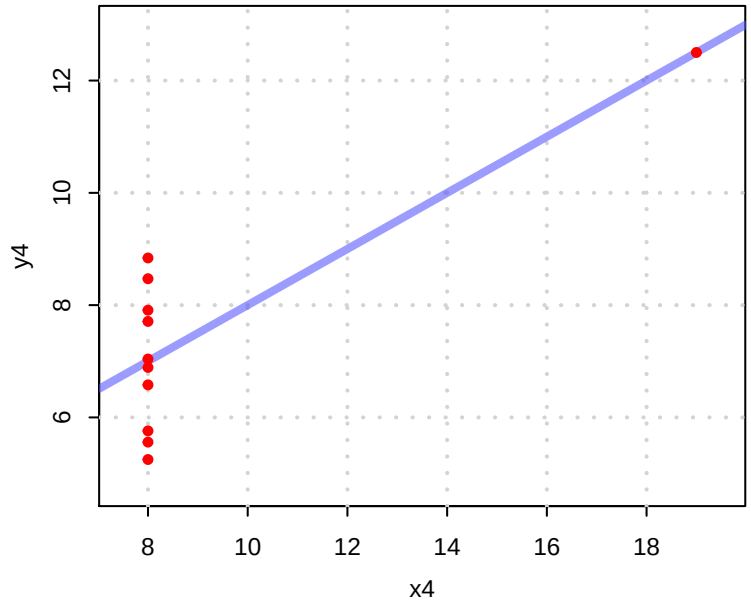
k	$\hat{a}_k$	$\hat{b}_k$	R_k^2
1	3,0001	0,5001	0,6665
2	3,0010	0,5000	0,6662
3	3,0025	0,4997	0,6663
4	3,0017	0,4999	0,6667

Plot der Anscombe-Daten







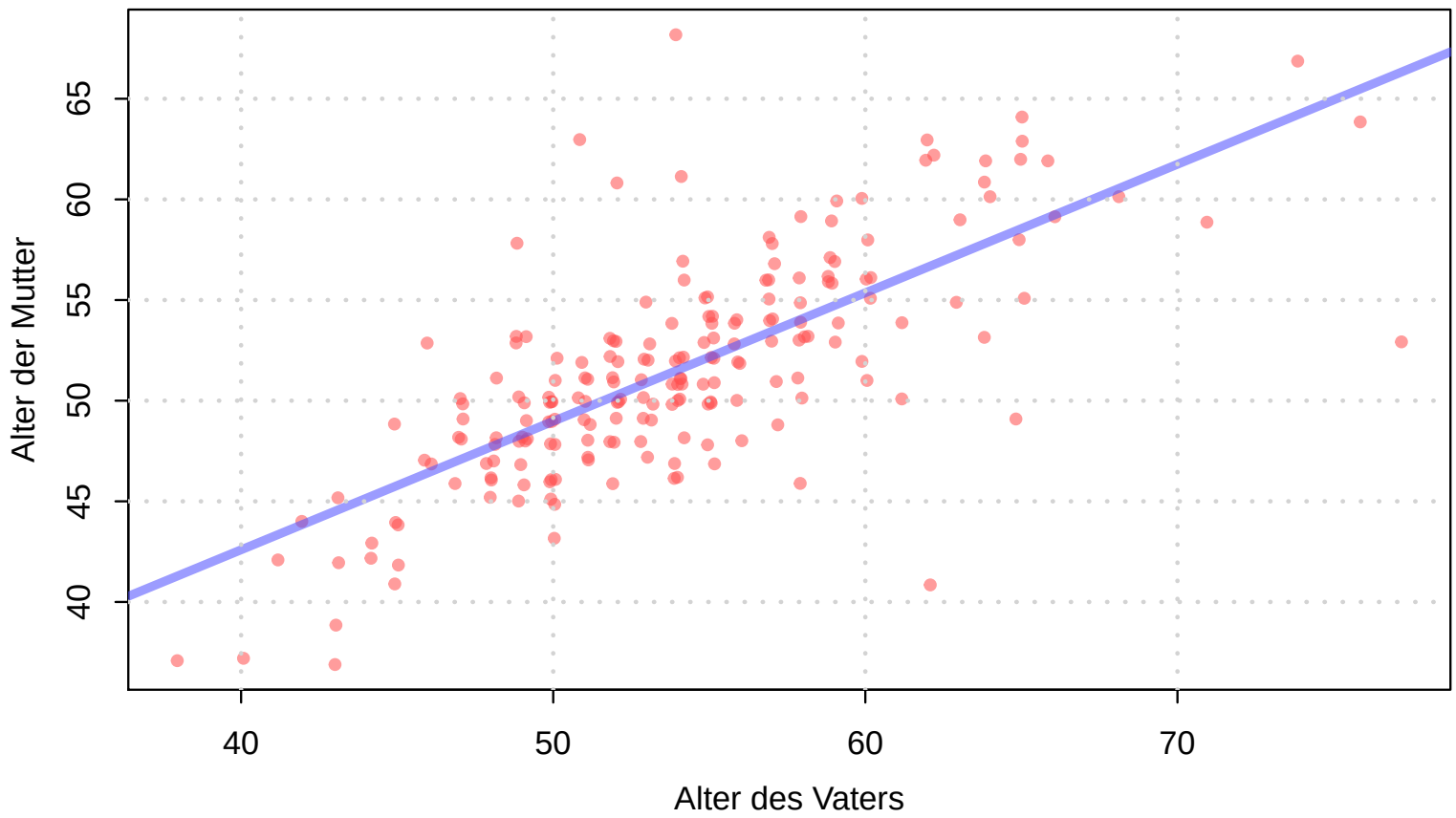


Beispieldaten

```r
meineRegression = lm(AlterM ~ AlterV)
meineRegression

plot(AlterV, AlterM,
     xlab="Alter des Vaters",
     ylab="Alter der Mutter")
abline(meineRegression)
```

```
##
## Call:
## lm(formula = AlterM ~ AlterV)
##
## Coefficients:
## (Intercept)       AlterV
##     17.0537       0.6384
```



Cook's Distanz

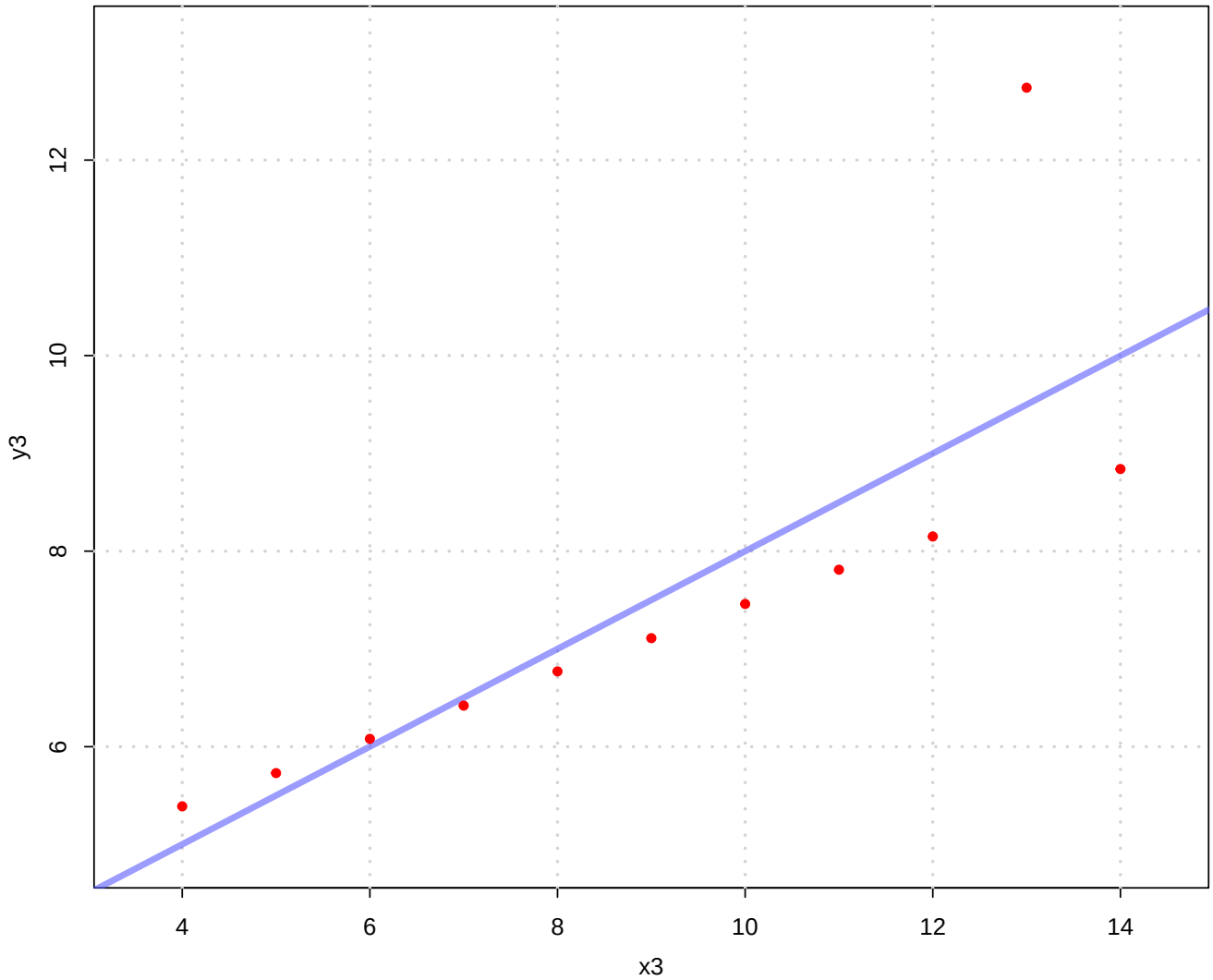
- Oft Kritisch: Einzelne Punkte, die Modell stark beeinflussen
- Idee: Was würde sich ändern, wenn solche Punkte weggelassen würden?
- Cook-Distanz: Misst den Effekt eines gelöschten Objekts
- Formel für ein lineares Modell mit einem unabh. Merkmal:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(\text{ohne } i)})^2}{\text{MSE}}$$

- Dabei bedeutet:
  - $\hat{y}_j$: Prognosewert des kompletten Modells für das j-te Objekt
  - $\hat{y}_{j(\text{ohne } i)}$: Prognosewert des Modells ohne Objekt i für das j-te Objekt
  - $\text{MSE} = \frac{1}{n} \cdot \sum (\hat{y}_i - y_i)^2$: Normierender Term (Schätzwert für Fehlerstreuung)

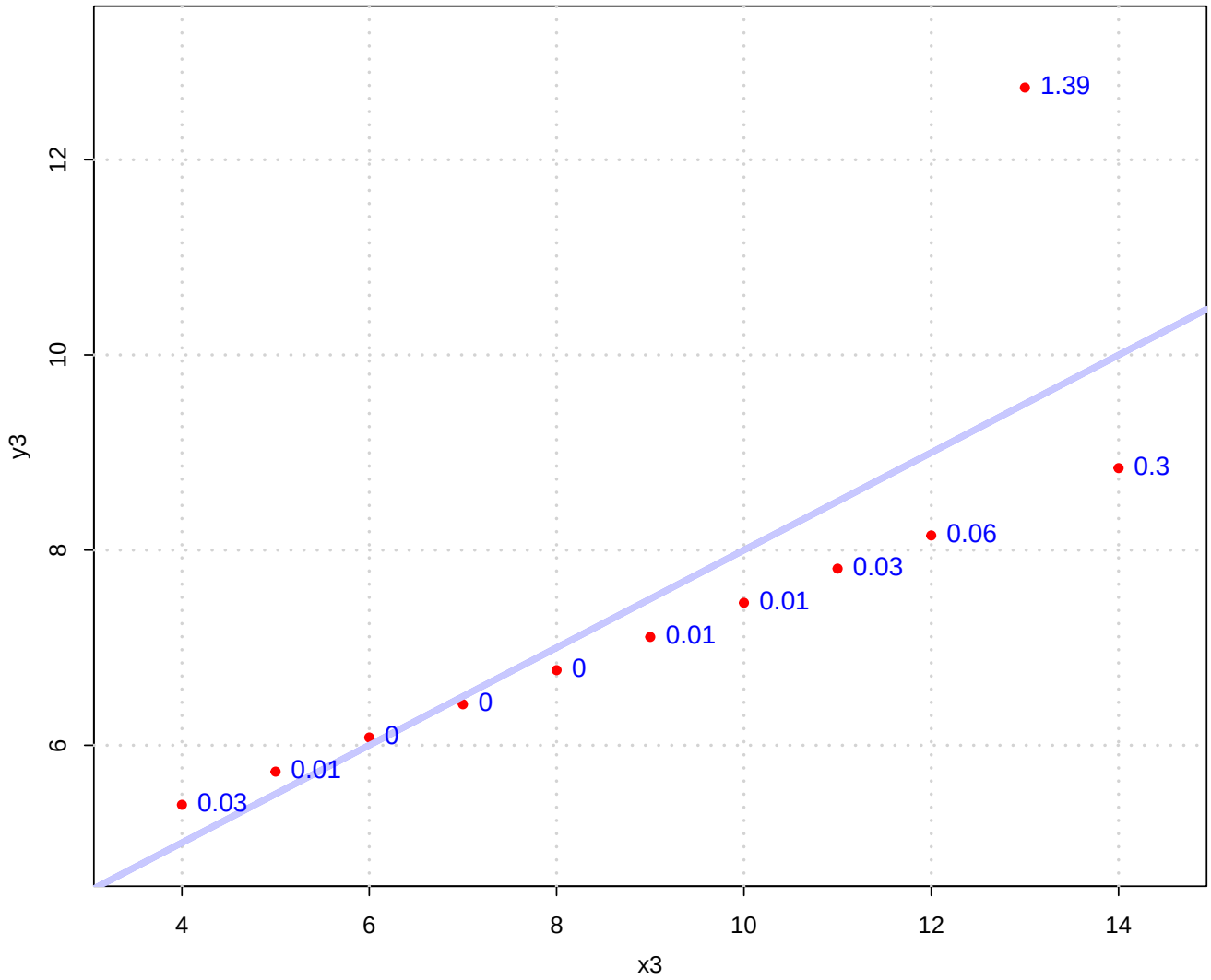
Ausreißer?

- Anscombe-Daten: Regressionsmodell Nr. 3

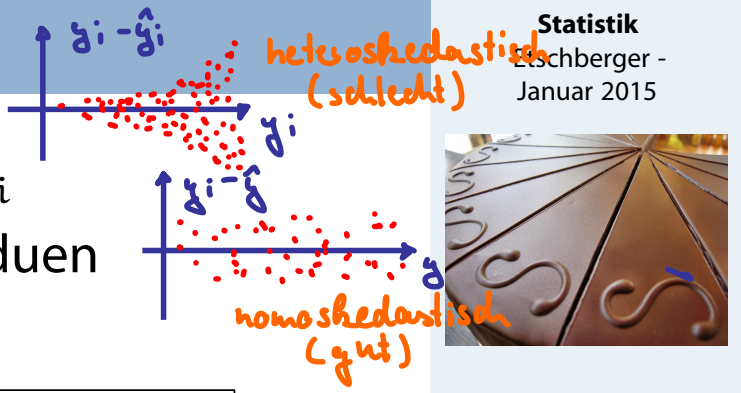


Ausreißer?

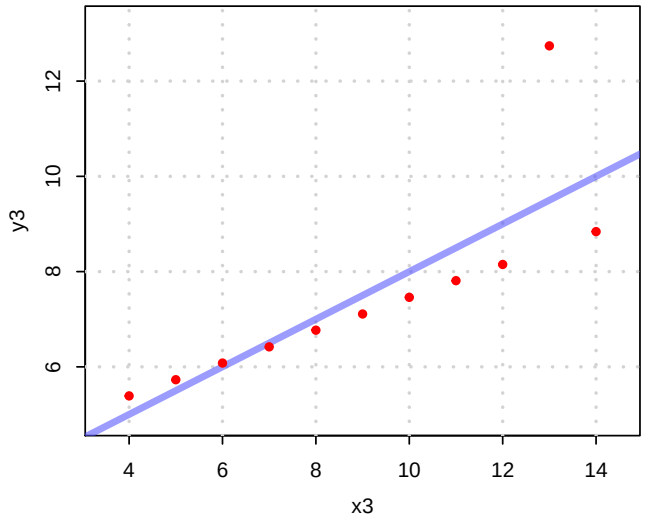
- Anscombe-Daten: Regressionsmodell Nr. 3
- Darstellung der Cook-Distanz neben Punkten
- Faustformel: Werte über 1 sollten genau untersucht werden

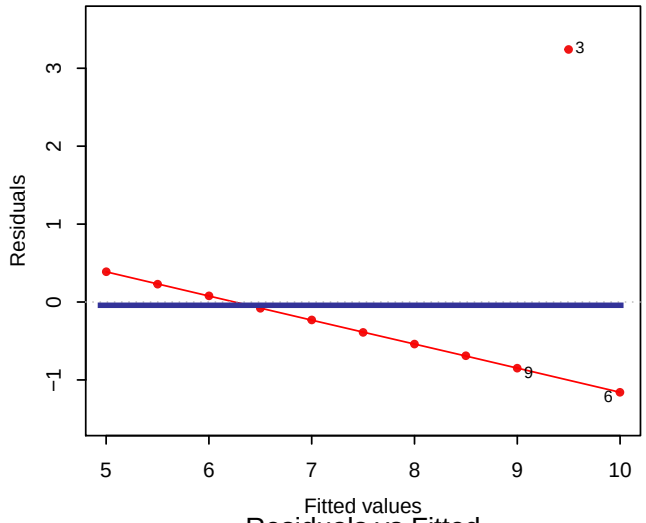


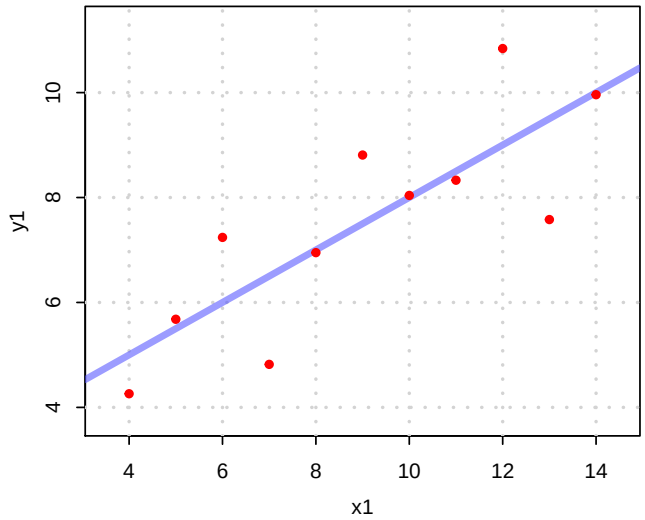
Residualanalyse

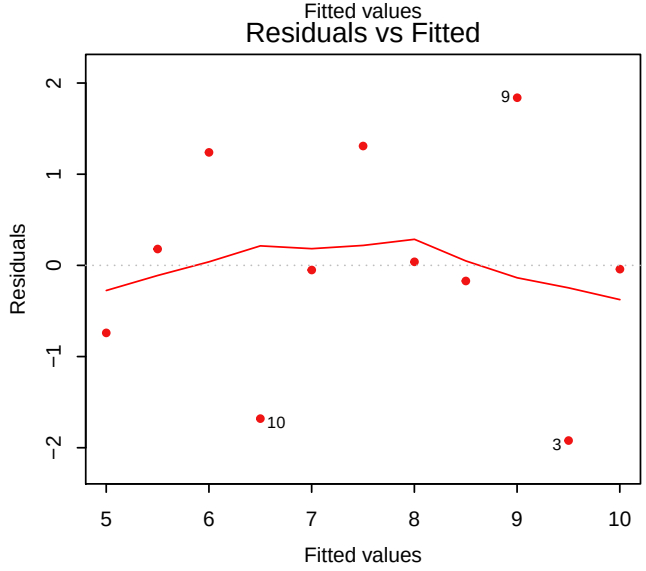


- Oft aufschlussreich: Verteilung der Residuen e_i
- Verbreitet: Graphische Darstellungen der Residuen
- Z.B.: e_i über $\hat{y}_i$



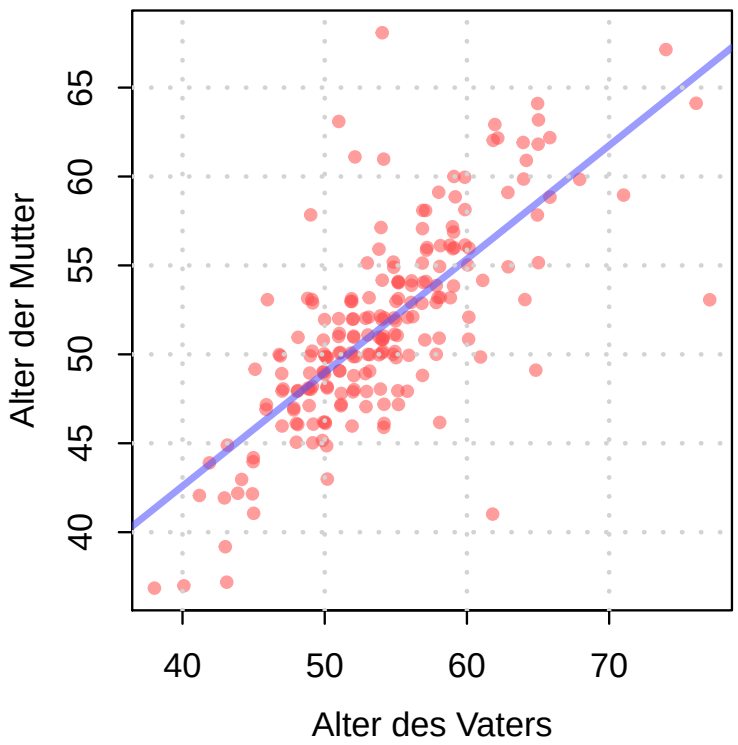


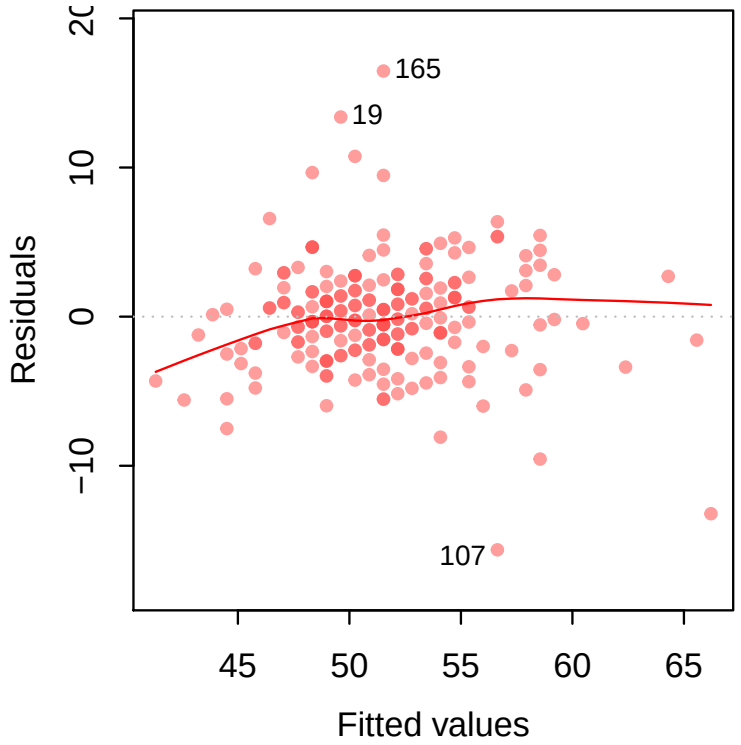




Wichtige Eigenschaften der Residuenverteilung

- Möglichst keine systematischen Muster
- Keine Änderung der Varianz in Abhängigkeit von $\hat{y}_i$ (Homoskedastizität)
- Nötig für inferentielle Analysen: Näherungsweise Normalverteilung der Residuen (q-q-plots)





Exkurs: Kausalität vs. Korrelation

- Meist wichtig für sinnvolle Regressionsanalysen:
- Kausale Verbindung zwischen unabhängigem und abhängigem Merkmal
- Sonst bei Änderung der unabhängigen Variablen keine sinnvollen Prognosen möglich
- Oft: Latente Variablen im Hintergrund