# Statistik

## für Betriebswirtschaft und internationales Management

Sommersemester 2015

Häufigkeiten mit TR

Setup → STAT → Häufigkeit On Ein

Prof. Dr. Stefan Etschberger
Hochschule Augsburg

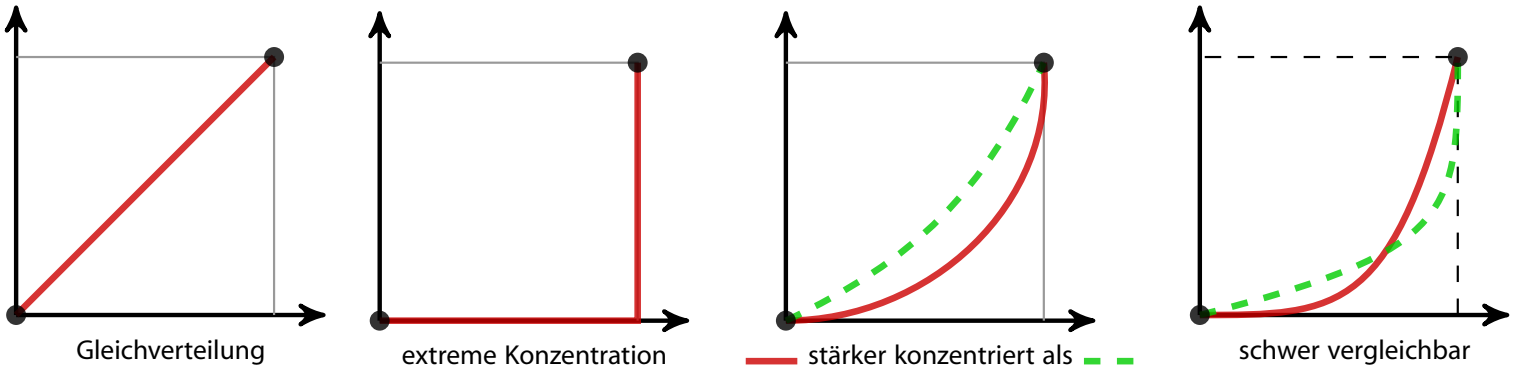
1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

## 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 |

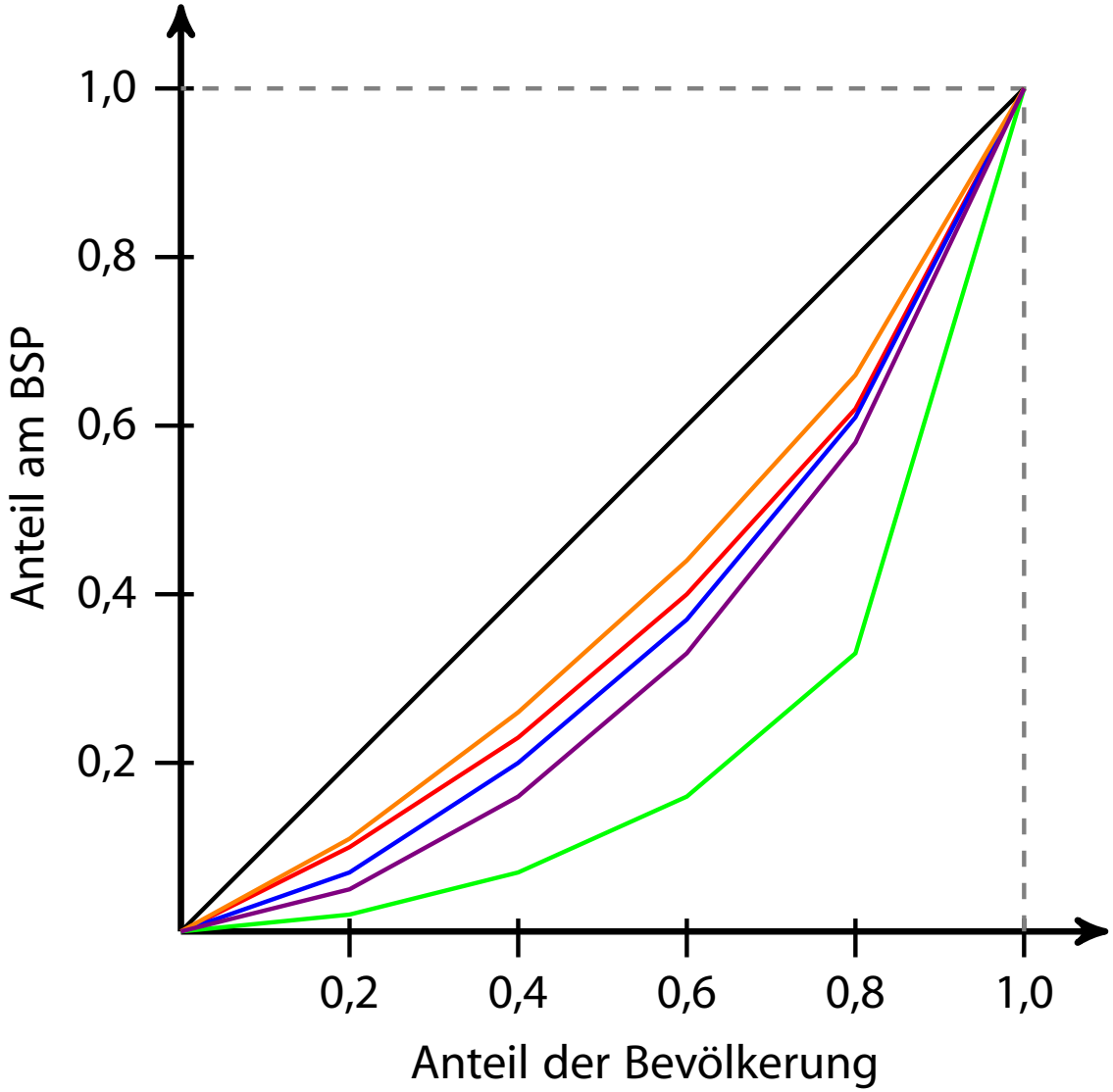
## Vergleich von Lorenzkurven:



# Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

(Stand 2000)






1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

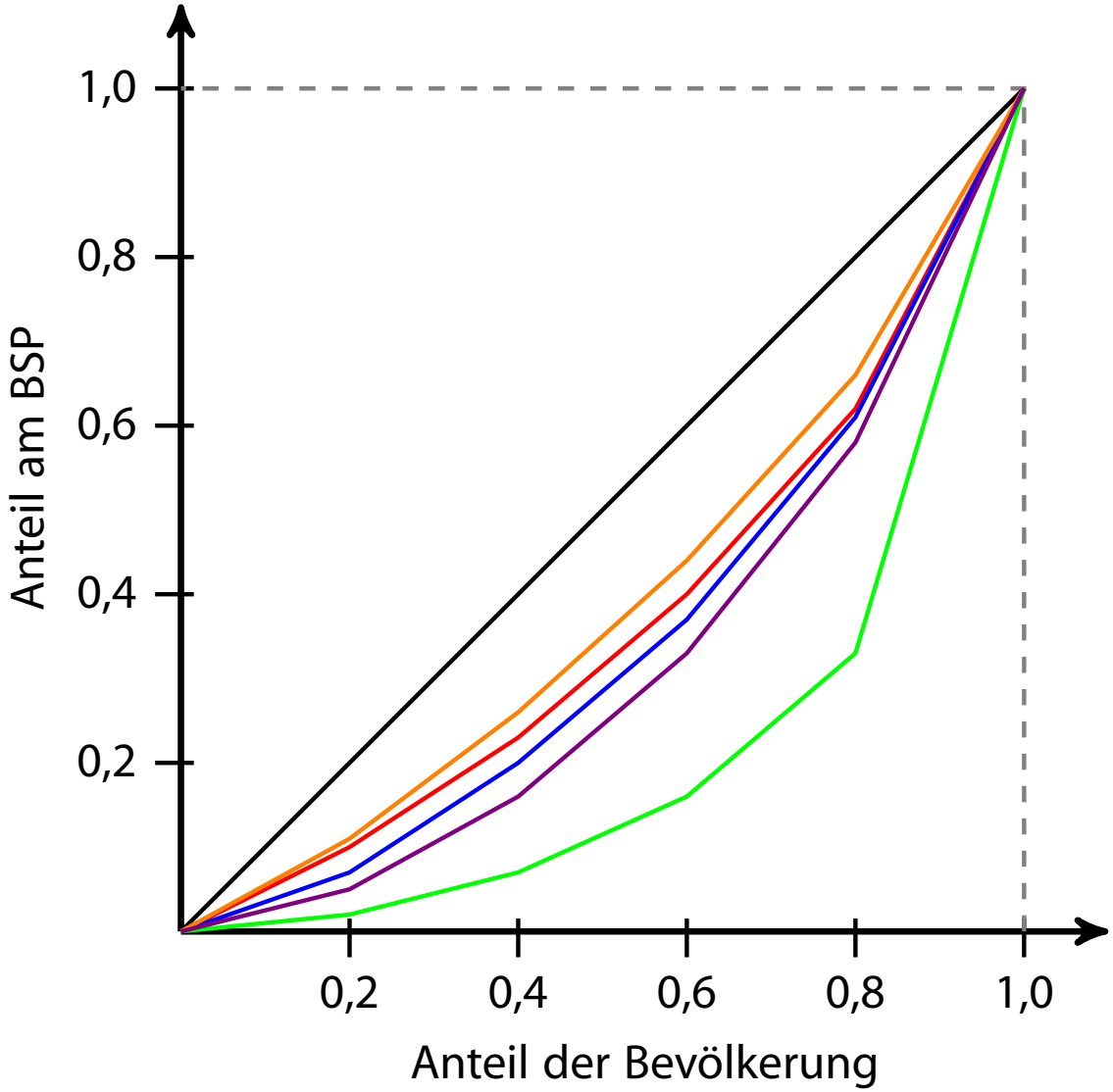
# Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP



1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

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:

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

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

## 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 |

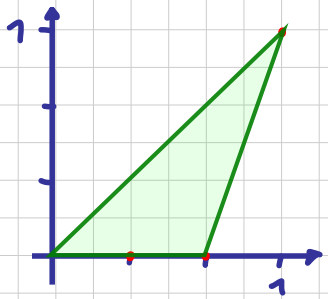
$$G = \frac{2 \cdot \left(1 \cdot \frac{1}{20} + 2 \cdot \frac{2}{20} + 3 \cdot \frac{2}{20} + 4 \cdot \frac{15}{20}\right) - (4+1)}{4} = 0{,}525$$

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

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

## Gini-Koeffizient: Maximum

Beispiel: 0, 0, 1 Mio



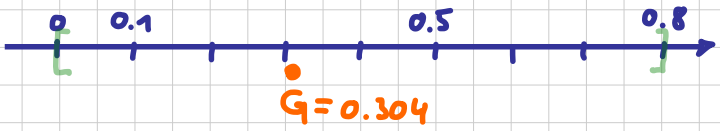

$G = 2 \cdot \triangle = 2 \cdot \left(\frac{1}{2} \cdot \frac{2}{3} \cdot 1\right) = \frac{2}{3}$

$\Rightarrow G_{max} = \frac{n-1}{n}$

(Beispiel 24.3.2015

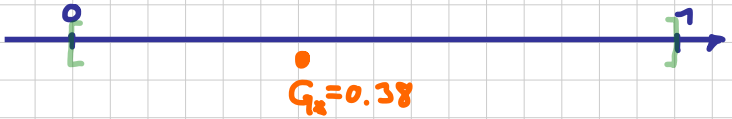
$G = 0{,}304$

hier: $n = 5 \Rightarrow G_{max} = \frac{4}{5} = 0.8$




Normierter Gini-Koeffizient: $G^* = \frac{G}{G_{max}}$

hier: $G^* = \frac{0.304}{0.8} = 0.38$



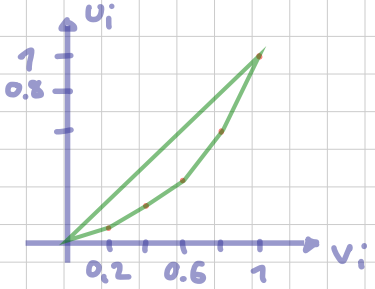

## Beispiel (Gini-Koeffizient)

Urliste 100 500 200 150 300

gesucht: Lorenzkurve, Gini-Koeffizient

$\sum x_i = 1250$

| sortiert: | 100 | 150 | 200 | 300 | 500 |
|---|---|---|---|---|---|
| $p_i$ | $\frac{1}{5}$ | $\frac{2}{5}$ | $\frac{3}{5}$ | $\frac{4}{5}$ | $\frac{5}{5}$ |
| $u_i$ | $\frac{100}{1250}$ | $\frac{250}{1250}$ | $\frac{450}{1250}$ | $\frac{750}{1250}$ | 1 |
| $v_i$ | $\frac{1}{5}$ | $\frac{3}{5}$ | $\frac{3}{5}$ | $\frac{4}{5}$ | $\frac{5}{5}$ |




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

$= \frac{2}{5}\left[1 \cdot \frac{100}{1250} + 2 \cdot \frac{150}{1250} + 3 \cdot \frac{200}{1250} + 4 \cdot \frac{300}{1250} + 5 \cdot \frac{500}{1250}\right] - \frac{6}{5}$

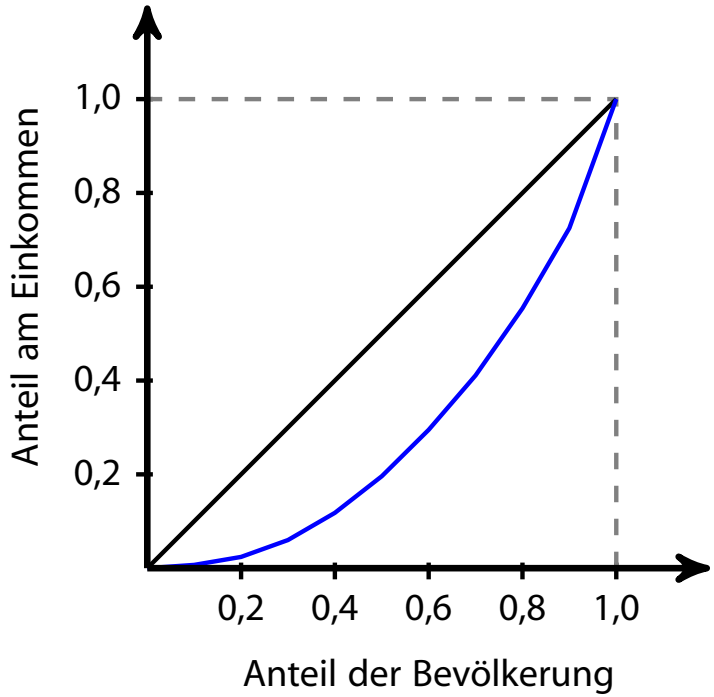
$= \frac{2}{1250}\left[1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + 4 \cdot 6 + 5 \cdot 10\right] - \frac{6}{5}$

$= \frac{188}{125} - \frac{6}{5} = \frac{38}{125} \approx 0{,}304$

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

## Armutsbericht der Bundesregierung 2008

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

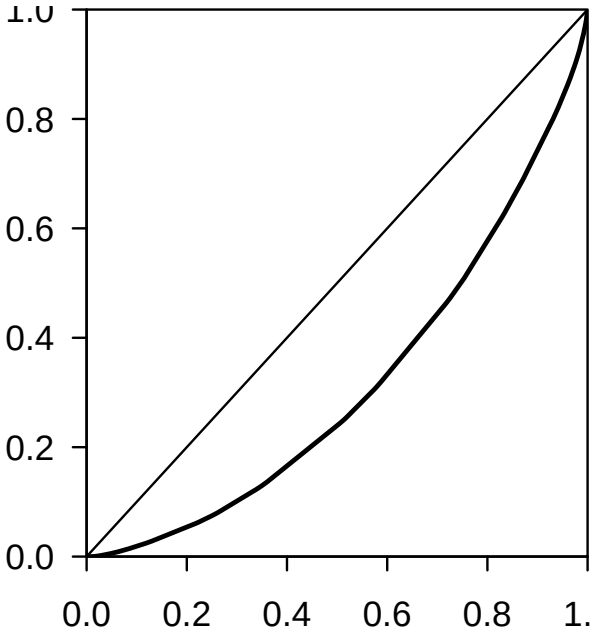


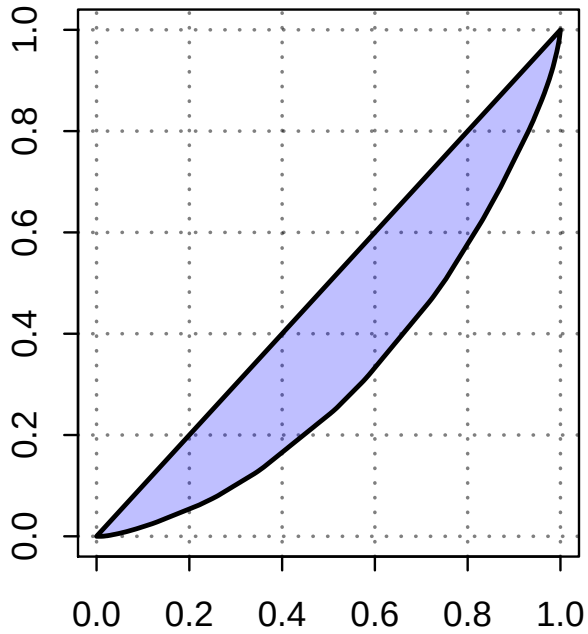

| | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|
| Arithmetisches Mittel | 24.873 | 24.563 | 23.987 | 23.648 |
| Median | 21.857 | 21.531 | 20.438 | 20.089 |
| Gini-Koeffizient | 0,433 | 0,441 | 0,448 | 0,453 |

# 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] 0.3730148
```




1. Einführung
2. Deskriptive Statistik
   - Häufigkeiten
   - Lage und Streuung
   - Konzentration
   - Zwei Merkmale
   - Korrelation
   - Preisindizes
   - Lineare Regression
3. W-Theorie
4. Induktive Statistik

Quellen

Tabellen

# 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:**

z.B.: 10,10,10,10,10 $\Rightarrow \Sigma = 50$
$\Rightarrow H = \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \cdots + \left(\frac{1}{5}\right)^2 = 5\left(\frac{1}{5}\right)^2 = \frac{1}{5}$

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

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

- **Exponentialindex:**

Bsp: 10, 10, 10 $\Sigma = 30$
$\Rightarrow E = \left(\frac{1}{3}\right)^{1/3} \cdot \left(\frac{1}{3}\right)^{1/3} \cdot \left(\frac{1}{3}\right)^{1/3} = \left[\left(\frac{1}{3}\right)^{1/3}\right]^3 = \frac{1}{3}$

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

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

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

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

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

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

## 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:**

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

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

**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.)

# Streuungsdiagramm

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

## Beispiel:

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

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

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

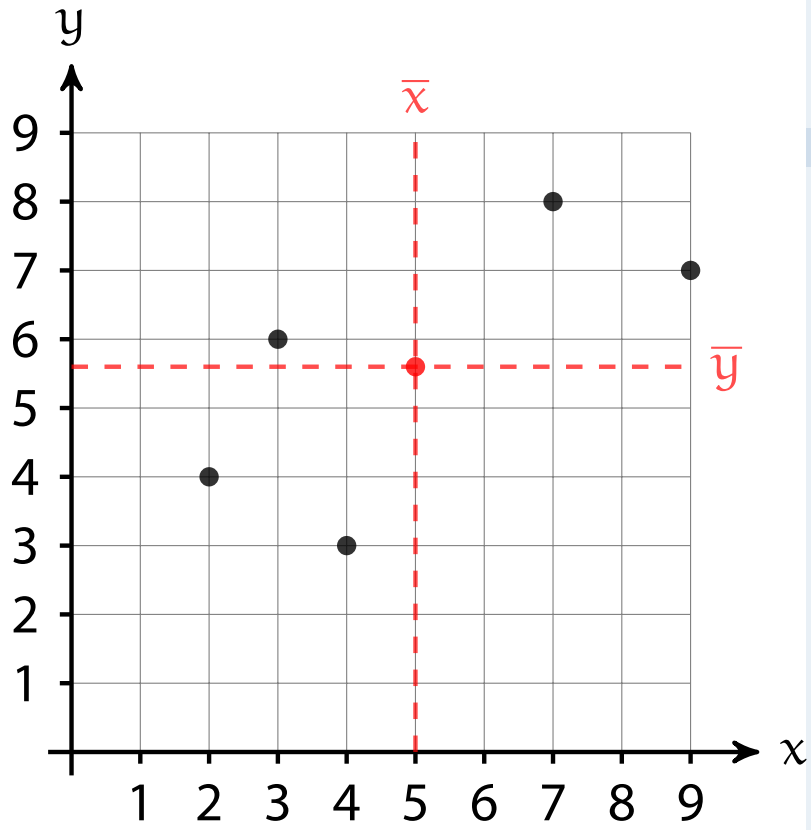
**Streuungsdiagramm** sinnvoll bei vielen verschiedenen Ausprägungen (z.B. stetige Merkmale)

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

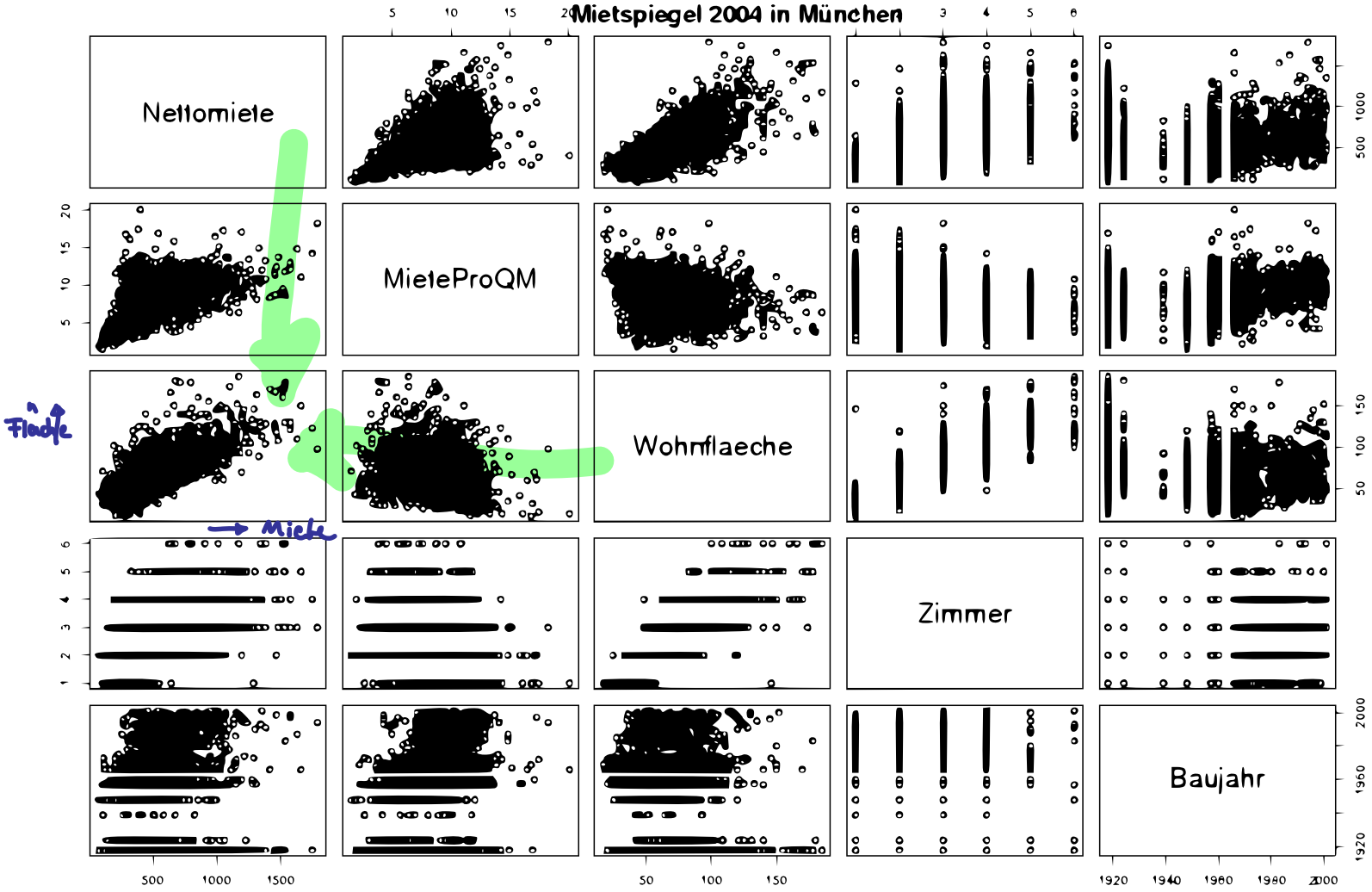
**Beispiel:**

| $i$ | 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: **fahrmeir2009** )



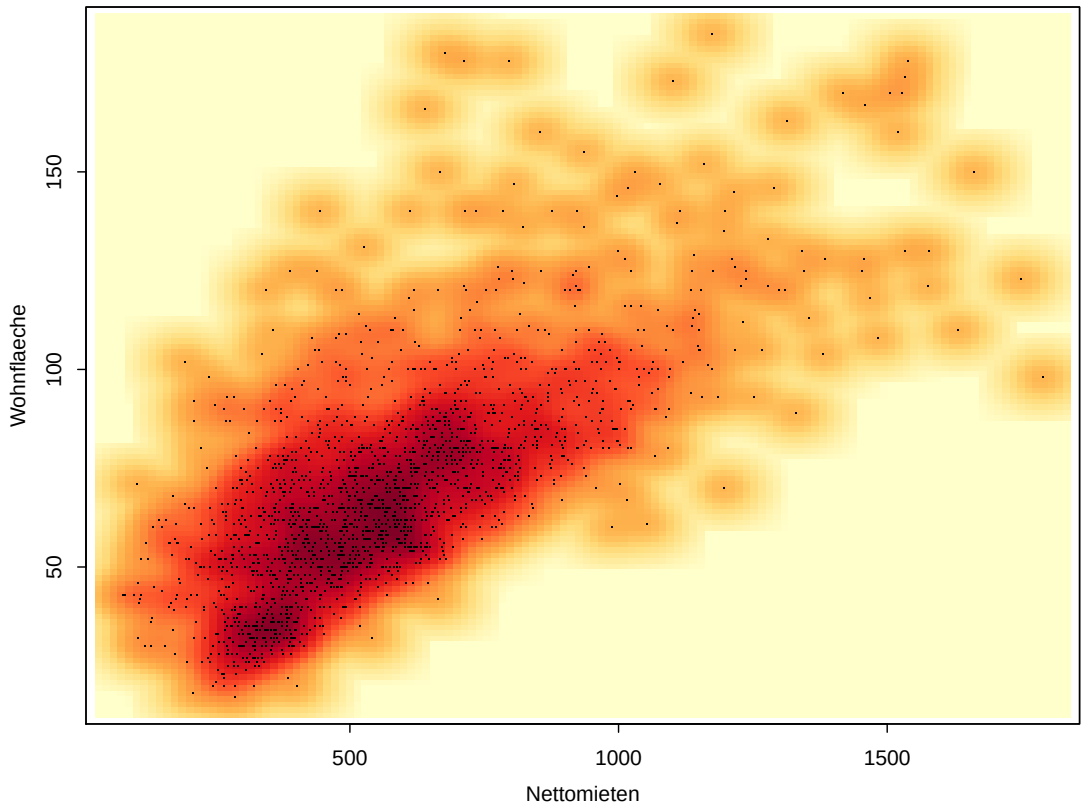
1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

# Beispiel Streuungsdiagramm



```
mieten <- read.table('http://goo.gl/jhpJW4', 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))
```

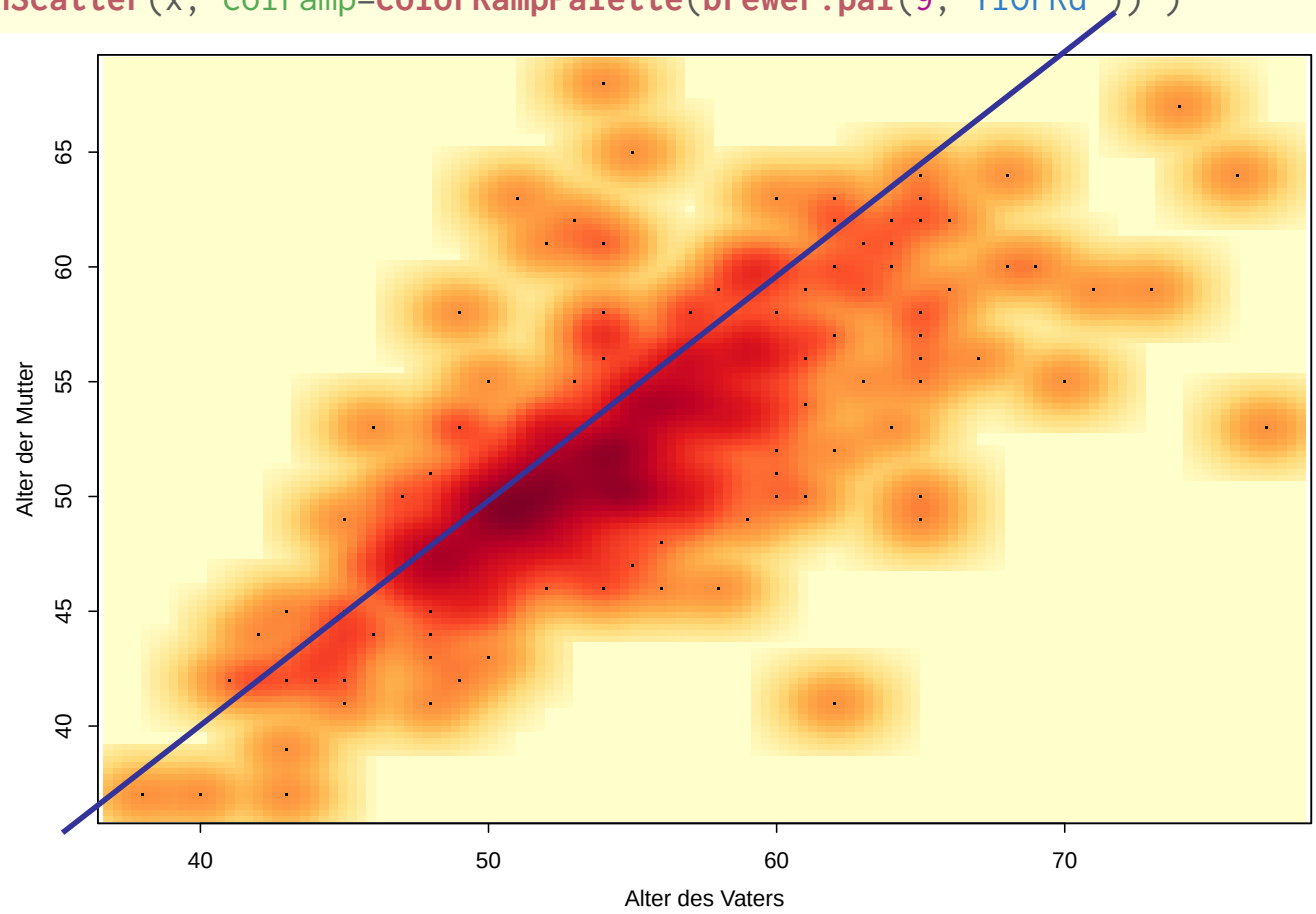



1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

# Beispiel Streuungsdiagramm



```
x = cbind("Alter des Vaters"=AlterV, "Alter der Mutter"=AlterM)
require("geneplotter") ## from BioConductor
smoothScatter(x, colramp=colorRampPalette(brewer.pal(9,"YlOrRd")) )
```




1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

```r
require(GGally)
ggpairs(MyData[, c("Alter", "AlterV", "AlterM", "AnzSchuhe", "Geschlecht")],
        upper = list(continuous = "density", combo = "box"),
       color='Geschlecht', alpha=0.5)
```

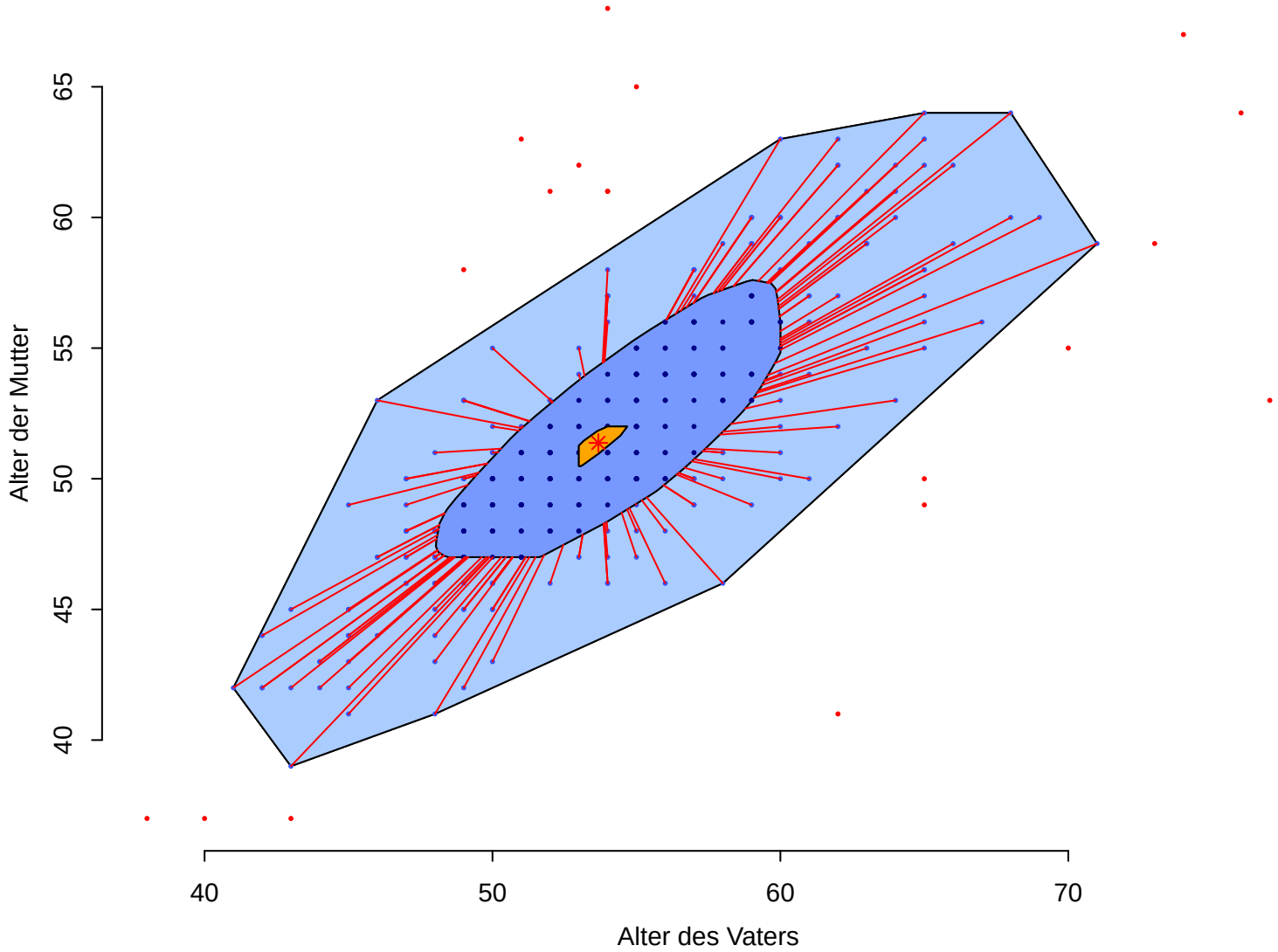



1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

# Bagplot: Boxplot in 2 Dimensionen



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



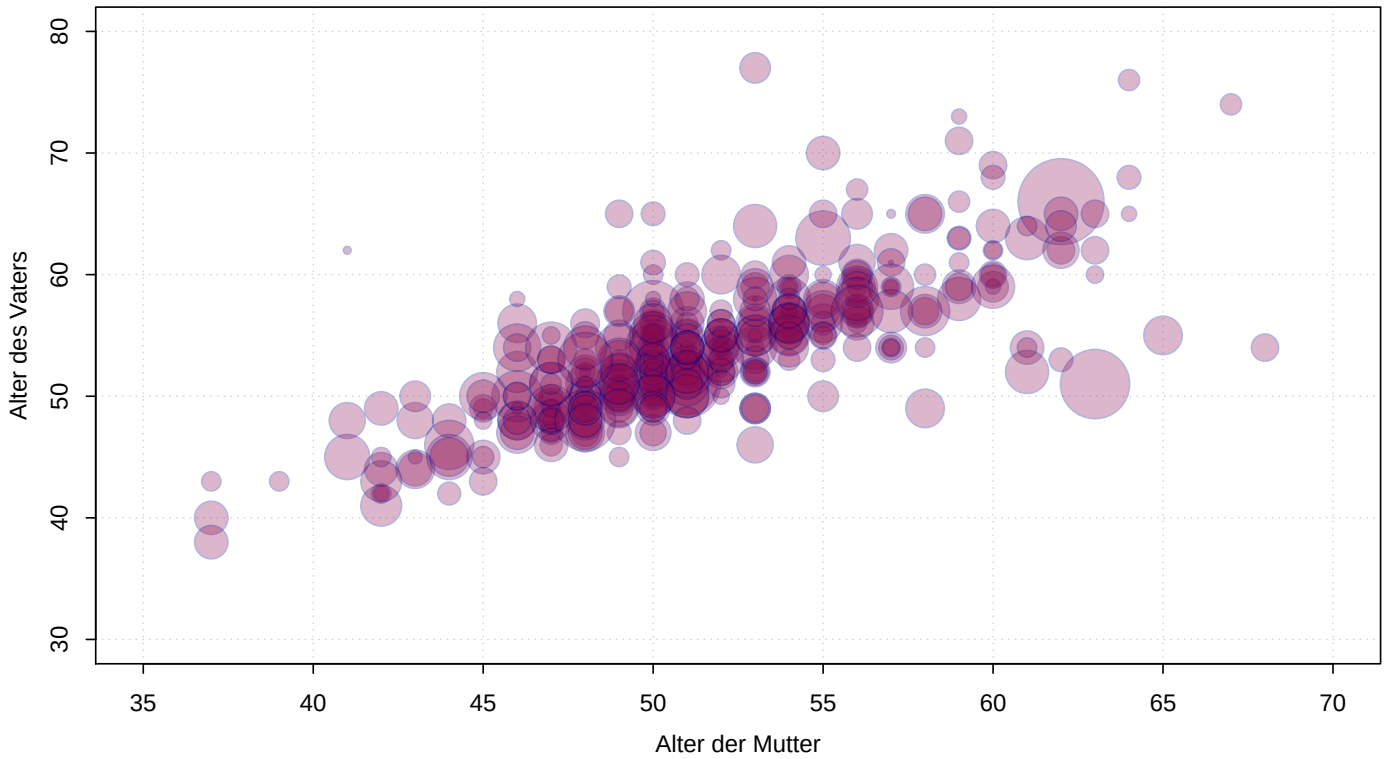

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

# Bubbleplot: 3 metrische Variablen



1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

```
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