



Statistik und Wahrscheinlichkeitsrechnung

# Kapitel 01: Deskriptive Statistik

Prof. Dr. Adrian Ulges

B.Sc. \*Informatik\*  
Fachbereich DCSM  
Hochschule RheinMain



# Outline

1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression

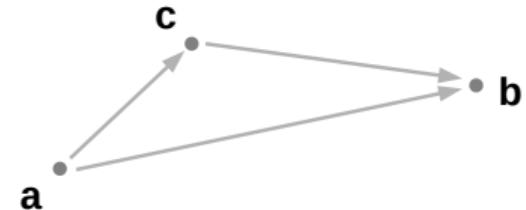
# Statistik: Motivation

In bisherigen Mathematik-Vorlesungen haben wir uns primär mit **universalgültigen** Sachverhalten befasst:

## Beispiele

*Jede konvergente Folge ist beschränkt.*

$$\|\mathbf{b} - \mathbf{a}\| \leq \|\mathbf{b} - \mathbf{c}\| + \|\mathbf{c} - \mathbf{a}\| \quad \forall \mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$$



Wie "sicher" ist unser Wissen in der Praxis ...?

*"Wissen = sowohl subjektiv als auch objektiv  
zureichendes Fürwahrhalten"*

(Immanuel Kant)



# Statistik: Motivation

In der Realität ist unser Wissen oft mit Unsicherheit behaftet

- ▶ “Bringt Route A mich schneller zum Ziel als Route B?”
- ▶ “Wird das Wetter morgen sonnig?”
- ▶ “Werde ich die Statistik-Klausur bestehen?”
- ▶ “Wird die Behandlung von Patient X erfolgreich sein?”
- ▶ “Wird Global Warming den Meeresspiegel um 50 cm ansteigen lassen?”



Im alltäglichen Sprachgebrauch...

“Vermutung”, “Zweifel”, “Risiko”, “Unsicherheit”,  
“Prognose”, “meistens”, “voraussichtlich”,  
“in der Regel”, “erwartungsgemäß”,  
“eventuell”, “selten”, ...



Was ist der Grund für diese Unsicherheit?

- ▶ **Unvollständigkeit von Information:** In der Praxis sind oft nur manche Variablen eines Problems bekannt, andere sind **latent** (*d.h., wir kennen ihren Wert nicht*).
- ▶ Wird ein Problembereich **maschinell abgebildet**, ist dies sehr häufig der Fall (Beispiel: *Routenplaner*).
- ▶ **Ziel:** Treffe **optimale Entscheidungen** bei **unvollständiger Information!**



# Definieren Sie “Statistik”!

“Die Wissenschaft von der zahlenmäßigen *Erfassung*, *Untersuchung* und *Auswertung* von *Massenerscheinungen*”

(duden.de)

“ ... mit *universellen Einsatzmöglichkeiten* in Politik, Wirtschaft und Gesellschaft und allen Geistes-, Sozial- und Naturwissenschaften.”

(Gabler Wirtschaftslexikon)

# Teilgebiet 1: Deskriptive Statistik

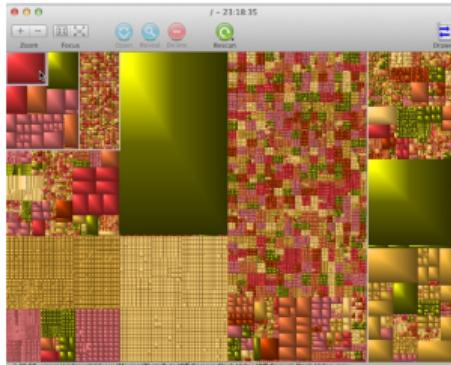
Bilder: [13] [6]



*“... quantitatively describing the main features  
of a data collection ...”*

(Mann: Introductory Statistics)

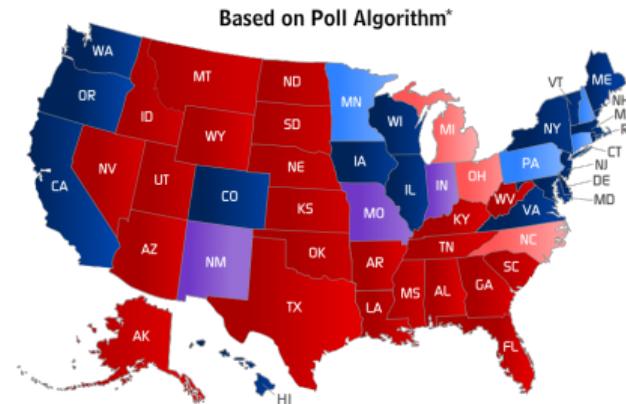
- ▶ Zielsetzung: Große, komplexe Stichproben übersichtlich darstellen/beschreiben
- ▶ Beschreibung mittels **Visualisierung**
- ▶ Beschreibung mittels **Kennzahlen** (*unser Schwerpunkt*)



## Teilgebiet 2: Induktive Statistik

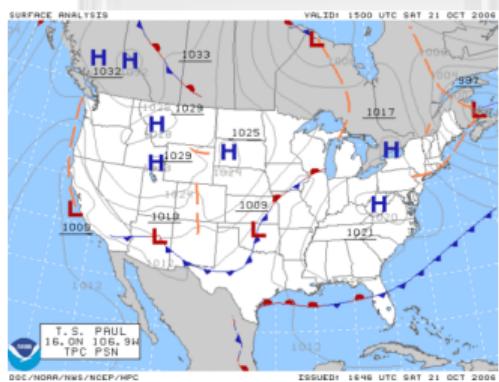
Bilder: [8] [11]

- ▶ Zielsetzung: **Schlussfolgerungen** ziehen von einer **Stichprobe** (*mit wenigen Objekten*) auf **alle Objekte**
- ▶ Verwendet **Wahrscheinlichkeitsrechnung**
- ▶ Grundlage intelligenter Systeme
- ▶ Hauptgegenstand dieser Vorlesung.



# Anwendungsfelder

Bilder: [9] [2] [14] [1] [16] [3] [12]



Amazon.com: Why is this recommended for you?

amazon.com

Recommended for You

I Just Want You to Know: Letters to My Kids on Love, Faith, and Family  
Our Price: \$9.99  
Used & new from \$9.99  
[See all buying options](#)

the camellizer

Because you purchased...

I Am Ozzy (Kindle Edition)

This was a gift  
Don't use for recommendations

Help | Close window

Done 575x521 FastProxy: Disabled

9



# Outline

1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression

# Grundbegriffe: “Grundgesamtheit”, “Stichprobe”

## Grundgesamtheit

Gegenstand statistischer Fragestellungen ist meist eine Gruppe (oder “Population”) ähnlicher Objekte, die **Grundgesamtheit**.

- ▶ *Beispiel: Die Bevölkerung Deutschlands*
- 

## Stichprobe

Wir erfassen (z.B. aus Kostengründen) Daten zu einer **Teilmenge** der Grundgesamtheit. Diese Daten nennen wir die **Stichprobe**. Die Anzahl der erfassten Objekte  $n$  ist der **Umfang** der Stichprobe.

- ▶ *Beispiel: Eine Anzahl zufällig ausgewählter Personen*
- 

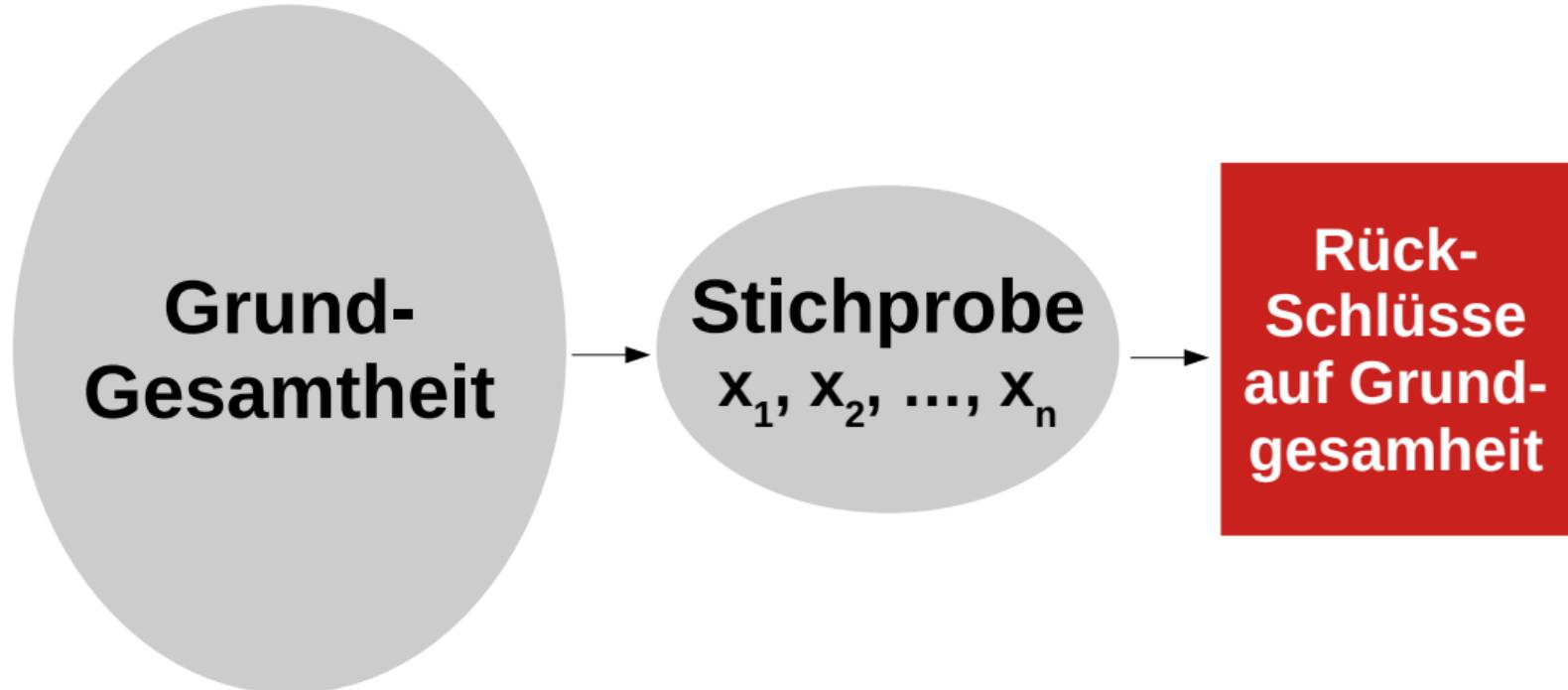
## Merkmale

Die Datenpunkte (engl. *Samples*) der Stichprobe werden durch **Merkmale** beschrieben.

- ▶ *Beispiel: Jede Person gibt an welche Partei sie wählen würde*  
 $x_1, \dots, x_n$  mit  $x_i \in \{CDU, Grüne, Linke, SPD, \dots\}$

## Grundbegriffe: "Grundgesamtheit", "Stichprobe"

\*





# Grundbegriffe: “Merkmal”

Wir unterscheiden **vier Arten** von Merkmalen:

- ▶ **nominal**: Kategorien ohne natürliche Ordnung  
z.B. *Farbtöne (rot, grün, gelb...)*, *Geschlechter*
- ▶ **ordinalskaliert**: Kategorien mit Ordnung  
z.B. *Dienstgrade (Bachelor < Master < PhD )*
- ▶ **intervallskaliert**: Zahlen mit Abstandsmaß  
z.B. *Temperaturen, Kalendertage*
- ▶ **verhältnisskaliert**: Skale besitzt zusätzlich einen Nullpunkt  
z.B. *Körpergröße, Alter, Einkommen, Preise*

## Anmerkungen

- ▶ Nominale und ordinalskalierte Merkmale nennen wir auch “**qualitativ**”. Sie geben das Vorhandensein einer Eigenschaft (Qualität) an, aber nicht deren *Ausmaß*.
- ▶ Intervallskalierte oder verhältnisskalierte Merkmale nennen wir auch “**quantitativ**”.



# Absolute Häufigkeit und Relative Häufigkeit

## Definition (Absolute und relative Häufigkeit)

In einer Stichprobe  $x_1, \dots, x_n$  kommen die Werte  $a_1, \dots, a_m$  vor.

Dann bezeichnen wir die **Anzahl der Vorkommen** eines Wertes  $a_j$  als die **absolute Häufigkeit** von  $a_j$ :

$$H_j := \#\left\{ i \mid i \in \{1, \dots, n\} \text{ und } x_i = a_j \right\}.$$

Wir erhalten die **relative Häufigkeit**  $h_j$ , indem wir durch die Stichprobengröße teilen:

$$h_j := H_j / n$$

```
# Autos
schwarz
grau
silber
rot
weiß
schwarz
silber
weiß
schwarz
braun
grau
rot
schwarz
silber
weiß
```

## Beispiel: Autofarben

- Werte in der Stichprobe:

$$(a_1, \dots, a_6) = (\text{schwarz}, \text{grau}, \text{silber}, \text{rot}, \text{weiß}, \text{braun})$$

- Absolute Häufigkeiten:

$$(H_1, \dots, H_6) = (4, 2, 3, 2, 3, 1)$$

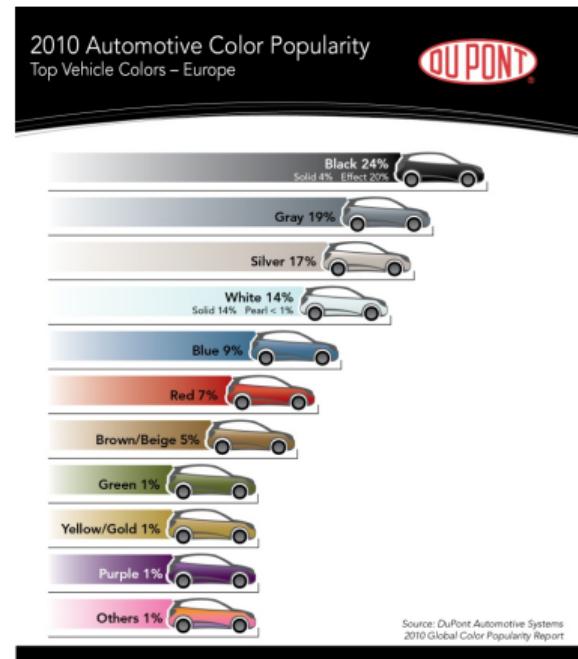
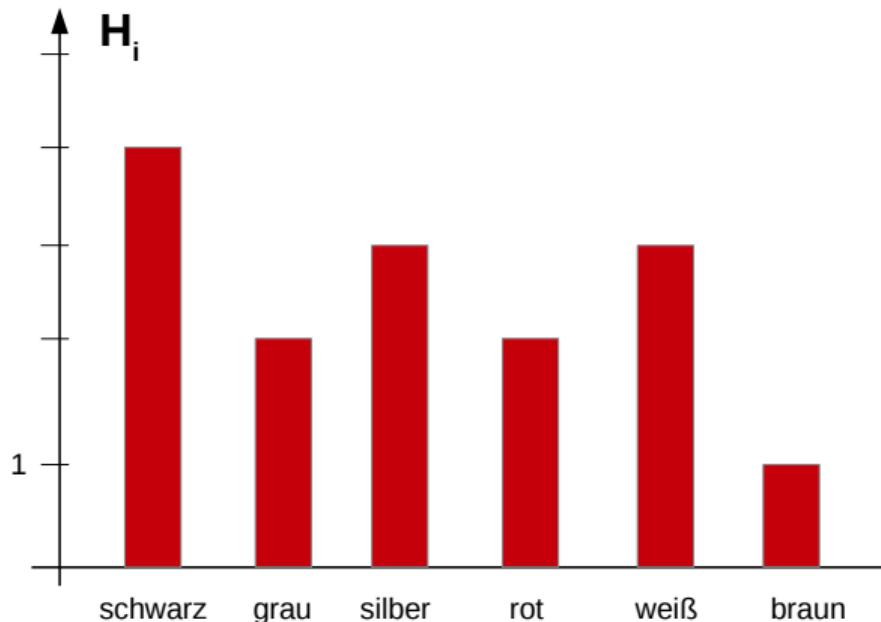
- Relative Häufigkeiten ( $n = 15$ ):

$$(h_1, \dots, h_6) = \left( \frac{4}{15}, \frac{2}{15}, \frac{3}{15}, \frac{2}{15}, \frac{3}{15}, \frac{1}{15} \right)$$

# Säulen-/Balkendiagramme

Bild: [15]

Wir stellen absolute und relative Häufigkeiten in Form von **Säulendiagrammen** (links) oder – falls um  $90^\circ$  gedreht – **Balkendiagrammen** (rechts) dar.





# Grundbegriffe: “Histogramm”

- ▶ Bei quantitativen Merkmalen macht es oft keinen Sinn, die Häufigkeit der einzelnen Werte direkt zu zählen (*warum?*).
- ▶ Wir nehmen an, die Samples liegen innerhalb eines Intervalls  $[a, b]$ . Üblicher Weise wählen wir als **Randpunkte**  $a, b$  das Minimum und Maximum der Stichprobe, plus etwas “Randabstand”.
- ▶ Wir wählen eine **Zerlegung** des Intervalls,  $Z_p = (y_0, y_1, \dots, y_p) \in \mathbb{R}^{p+1}$  mit

$$(a = )y_0 < y_1 < \dots < y_p (= b)$$

- ▶ Wir bezeichnen die einzelnen Abschnitte  $[y_0, y_1], [y_1, y_2], \dots, [y_{p-1}, y_p]$  als **Klassen** (engl. *bins*).

## Definition (Histogramm)

Gegeben eine Stichprobe  $x_1, \dots, x_n$  sowie eine Zerlegung  $Z_p = (y_0, y_1, \dots, y_p)$ , ermitteln wir die absoluten (bzw. relativen) Häufigkeiten der einzelnen **Klassen**,

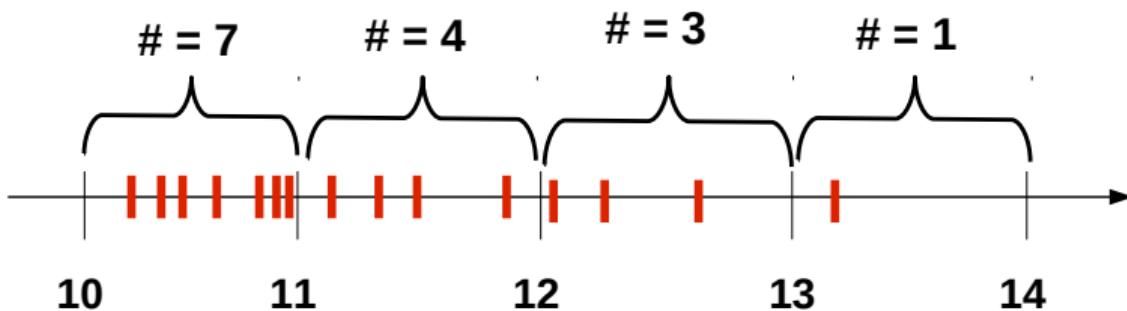
$$H_k^* := \#\{ i \mid x_i \in ]y_{k-1}, y_k] \}, \quad (\text{bzw. } h_k^* := H_k^*/n),$$

und bezeichnen das Ergebnis  $(H_1^*, \dots, H_p^*)$  (bzw.  $(h_1^*, \dots, h_p^*)$ ) als **Histogramm**.

## “Histogramm”: Beispiel

```
# n=15
12.03
13.14
10.85
10.94
11.32
11.14
10.67
12.34
11.56
10.37
11.89
10.47
12.63
10.23
10.94
```

- ▶ Wir wählen als Zerlegung:  $Z_4 = (10, 11, 12, 13, 14)$
- ▶ Anzahl der Samples pro Bin
  - ▶  $H_1^* = \# \text{ Werte zwischen } 10 \text{ und } 11 = 7$
  - ▶  $H_2^* = \# \text{ Werte zwischen } 11 \text{ und } 12 = 4$
  - ▶  $H_3^* = \# \text{ Werte zwischen } 12 \text{ und } 13 = 3$
  - ▶  $H_4^* = \# \text{ Werte zwischen } 13 \text{ und } 14 = 1$
- ▶ Also:  $(H_1^*, \dots, H_4^*) = (7, 4, 3, 1)$



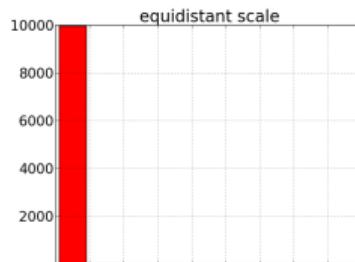
# Histogramme: Zerlegungen

- ▶ Oft wählen wir für unsere Histogramme **äquidistante** Zerlegungen, d.h. die Abschnitte sind *gleich breit*.
- ▶ Sind die Samples aber **ungleich** verteilt, wählen wir für Bereiche mit **wenig Samples** gerne **breitere Bins**.

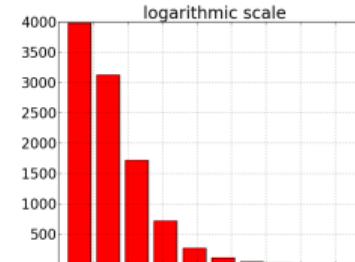
## Beispiel: Verteilung von Wörtern

```
# Vorkommen
the 56271872
of 33950064
and 29944184
to 25956096
in 17420636
i 11764797
that 11073318
was 10078245
...
purified 3924
sequel 3924
calves 3923
```

- ▶ Anzahl von Wortvorkommen in der (englischen) Wikipedia
- ▶ Die Verteilung ist **schief** (engl. *long-tail*): Die meisten Wörter sind selten, sehr wenige Wörter sind sehr häufig.



$$Z_{10} = (0, 5\text{mio.}, 10\text{mio.}, \dots, 30\text{mio.})$$

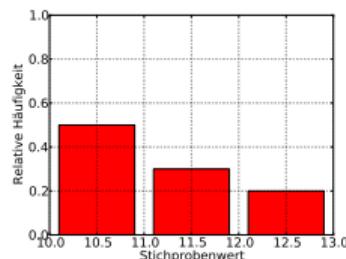


$$Z_{10} = (0, 10000, 20000, 50000, \dots, 30\text{mio.})$$

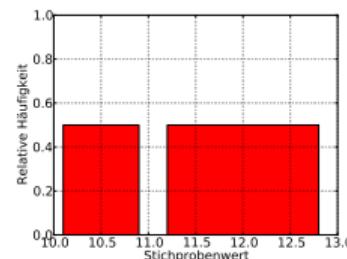
# Normierung von Histogrammen

- Histogramme sollen die “Dichte” der Datenpunkte je Bereich angeben. Sind Klassen **breiter**, müssen wir deshalb ihre Höhe im Balkendiagramm **normieren**.
  - Es sei  $\Delta_k := y_k - y_{k-1}$  die **Breite einer Klasse** und  $H_k^*$  (bzw.  $h_k^*$ ) die zugehörige Häufigkeit. Dann wählen wir die **Höhe des Rechtecks  $r_k$**  in der grafischen Darstellung als
- $$r_k := H_k^*/\Delta_k \quad (\text{bzw. } r_k := h_k^*/\Delta_k)$$

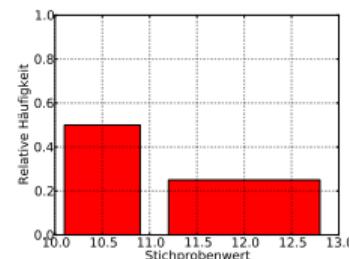
- Beispiel:** Verbreitern wir eine Klasse um den Faktor 2, halbieren wir die Höhe des zugehörigen Rechtecks!



$$Z = (10, 11, 12, 13)$$



$$Z = (10, 11, 13) \text{ (nicht normiert!)}$$



$$Z = (10, 11, 13) \text{ (normiert!)}$$

# Do-Histogramme-yourself

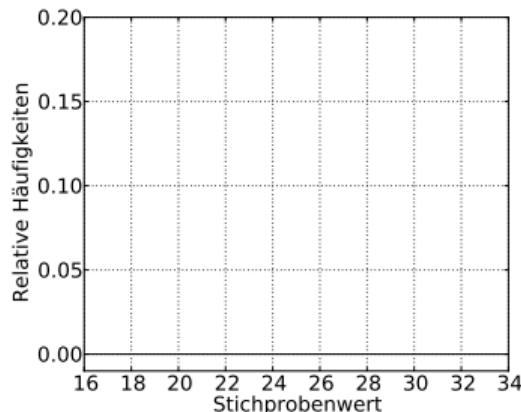
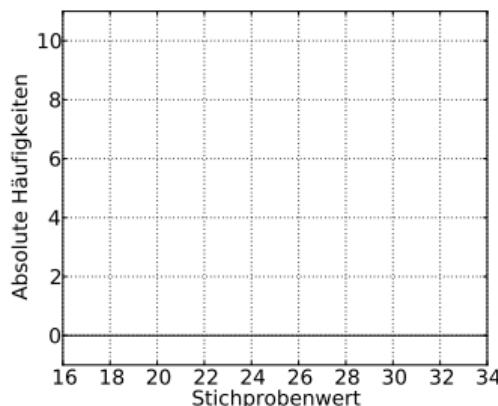


Gegeben ist die folgende (geordnete) Stichprobe:

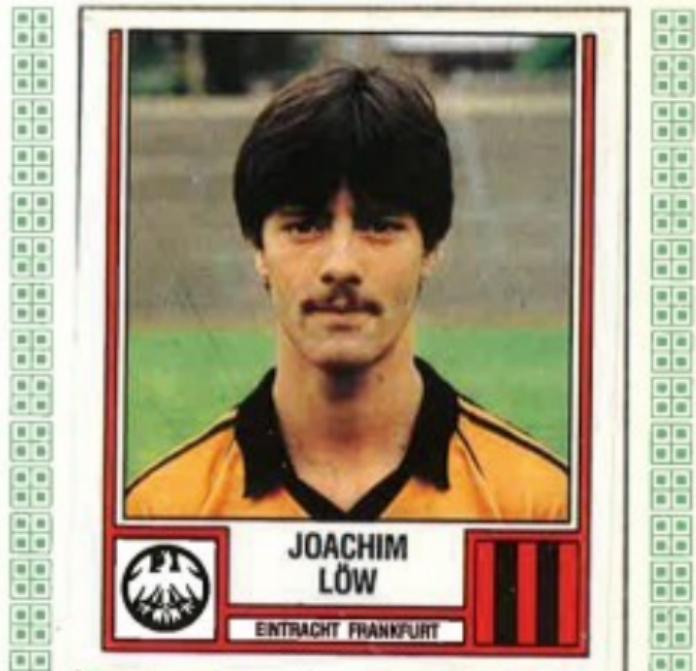
18, 18, 18, 19, 20, 20, 20, 20, 20, 21,  
21, 21, 22, 22, 23, 24, 24, 25, 28, 31.

Skizzieren Sie das Histogramm mit Zerlegung (17, 20, 22, 26, 32) ...

- a) nicht-normiert, mit absoluten Häufigkeiten (links)
- b) normiert, mit relativen Häufigkeiten (rechts)



## Grundbegriffe: “univariat” vs. “multivariat” Bild: [5]



Löw, Joachim, geb. 3.2.1960 in Freiburg, Mittelfeldspieler und Stürmer, 1,81 m, 72 kg, Stammverein Eintracht Freiburg.

Oft beschreiben wir unsere Objekte mit mehreren Merkmalen

- ▶ Name: ordinal
- ▶ Stammverein: ordinal
- ▶ Geburtsdatum: intervallskaliert
- ▶ Gewicht: verhältnisskaliert
- ▶ ...

1

Wir nennen Stichproben mit nur einem Merkmal **univariat** und Stichproben mit mehreren Merkmalen **multivariat**.

# Grundbegriffe: “multivariat”

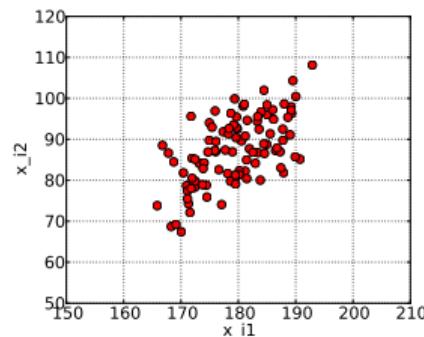
```
# Multivariate Stichprobe: Personen  
# Größe(cm) Gewicht(kg)  
172.51054875 89.81090441  
178.5491509 94.14875995  
160.19982596 77.61266935  
185.33582886 99.26311152  
173.42373218 78.43528082  
178.07276393 85.89384238  
171.39415797 81.06861227  
163.07000132 79.57634485  
178.97868362 87.06345319  
181.77268699 76.21846529  
165.12776354 78.48439432  
180.1249523 100.79476513  
160.15953819 76.28881635  
186.62205244 98.02854219  
178.1006582 94.72277617  
182.8521624 89.55199009  
..
```

## Multivariate Datensätze = Matrizen

- ▶  $n$  Datenpunkte, d.h.  $n$  Zeilen
- ▶ je Datenpunkt  $m$  Merkmale,  
d.h.  $m$  Spalten.

## Visualisierung: Scatterplots

- ▶ Bei quantitativen Daten ist jeder Datenpunkt ein **Punkt** in  $\mathbb{R}^m$
- ▶ Die Stichprobe ist eine “Punktwolke”:





# Outline

1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression



# Kennwerte von Stichproben: Motivation

**Grundproblem:** Die Beschreibung der Daten mittels Histogrammen ist oft noch **zu detailliert** / unhandlich.

*Beispiel: Verteilung der Produktpreise zweier Supermärkte –  
Welcher ist der günstigere?*

**Ansatz:** Charakterisiere die Stichprobe durch **Kennwerte**

*Mittelwert, Modalwert, Median, Varianz, Quantile, ...*

Wir unterteilen die Kennwerte in verschiedene Typen:

- ▶ **Lageparameter:** beschreiben die generelle Lage der Werte
- ▶ **Streuungsparameter:** beschreiben, wie stark Werte variieren
- ▶ **Zusammenhangsparameter:** beschreiben Abhängigkeiten zwischen Merkmalen

## Kennwerte: Mittelwert

### Definition ((Arithmetisches) Mittelwert)

Gegeben eine univariate Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$ , nennen wir

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$

den **Mittelwert** (oder das Arithmetische Mittel) der Stichprobe.

Beispiel:  $x_1, \dots, x_{11} = 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4$

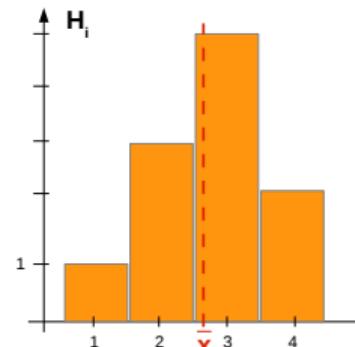
$$\begin{aligned}\bar{x} &= \frac{1}{11} \cdot (1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 4) \\ &= \frac{1}{11} \cdot 1 + \frac{3}{11} \cdot 2 + \frac{5}{11} \cdot 3 + \frac{2}{11} \cdot 4 \quad // \text{ mit relativen Häufigkeiten}\end{aligned}$$

$$\approx 2.7$$

# Mittelwert: Anschauliche Interpretation

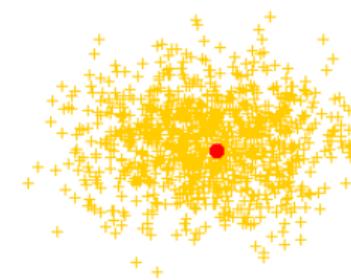
## Univariate Stichproben

- Der Mittelwert entspricht dem Schwerpunkt des Säulendiagramms



## Multivariate Stichproben

- Der Mittelwert entspricht dem Zentrum der Punktwolke.



## Berechnung (exemplarisch)

- Eine Stichprobe besitze zwei Merkmale, d.h. sie lautet  $(x_1, y_1), \dots, (x_n, y_n)$ .

$$\text{Mittelwert: } \frac{1}{n} \sum_{i=1}^n (x_i, y_i) = \left( \frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n y_i \right) = (\bar{x}, \bar{y})$$

# Kennwerte: Median

## Definition (Median)

Eine gegebene Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$  sei **sortiert** (d.h.,  $x_i \leq x_{i+1}$  für alle  $i = 1, \dots, n - 1$ ). Dann nennen wir den Wert in der Mitte der Stichprobe

$$\tilde{x} := \begin{cases} x_{\lceil \frac{n}{2} \rceil} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{sonst} \end{cases}$$

den **Median** der Stichprobe.

## Beispiele

- ▶  $x_1, \dots, x_{11} = 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4 \rightarrow \tilde{x} = x_6 = 3$
- ▶  $x_1, \dots, x_{10} = 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4 \rightarrow \tilde{x} = \frac{1}{2}(x_5 + x_6) = 2.5$

# Kennwerte: Median (cont'd)

## Anmerkungen

- ▶ Die Gaussklammern  $[x]$  bezeichnen das **Aufrunden** von  $x$ :  $[2.1] = 3$ .
- ▶ Ist die Stichprobe **unsortiert**, müssen wir sie vor der Berechnung des Medians sortieren.
- ▶ Der Median ist deshalb im Allgemeinen aufwändiger zu berechnen als der Mittelwert: Eine **(Teil-)sortierung** der Daten ist erforderlich.





## Median: Robustheit

- ▶ Wie verhalten sich Mittelwert und Median im Fall von **Ausreißern** (engl. *outliers*) in den Daten?
- ▶ Hier ist oft **Robustheit** erwünscht, d.h. der Kennwert sollte von Ausreißern nicht zu stark beeinflusst werden.

### Beispiel

Wo liegen bei dieser Stichprobe Mittelwert und Median?

1, 2, 2, 2,  
2, 3,  
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 573, 1224

## Lageparameter: Quantile

- ▶ Genau 50% der Samples einer Stichprobe liegen unterhalb des Medians.
- ▶ Dies können wir auf beliebige Prozentsätze  $\alpha$  erweitern!  
Wir erhalten die sogenannten  $\alpha$ -Quantile der Stichprobe.

Definition:  $\alpha$ -Quantil

Es sei  $\alpha \in ]0, 1[$  und  $x_1, \dots, x_n \in \mathbb{R}$  eine aufsteigend sortierten Stichprobe. Dann nennen wir

$$\tilde{x}_\alpha := \begin{cases} x_{\lceil \alpha \cdot n \rceil}, & \text{falls } \alpha \cdot n \notin \mathbb{N} \\ \frac{1}{2} \cdot (x_{\alpha \cdot n} + x_{\alpha \cdot n + 1}), & \text{sougt.} \end{cases}$$

das  $\alpha$ -Quantil der Stichprobe.

## Lageparameter: Quantile



# Quantile: Beispiel

$$\tilde{x}_\alpha := \begin{cases} x_{[\alpha n]} & \text{falls } \alpha n \notin \mathbb{N} \\ \frac{1}{2}(x_{\alpha n} + x_{\alpha n+1}) & \text{sonst} \end{cases} \quad (\star) \quad (\star\star)$$

$x_1, \dots, x_{14} = 10, 13, 14, 16, 18, 19, 19, 21, 23, 25, 28, 30, 36, 41$

(a)  $\alpha = 0,25 : \alpha \cdot n = 0,25 \cdot 14 = 3,5$  ( $\star$ )

$$\Rightarrow \tilde{x}_{0,25} = x_{\lceil 3,5 \rceil} = x_4 = 16$$

(b)  $\alpha = 0,5 : \alpha \cdot n = 0,5 \cdot 14 = 7$  ( $\star\star$ )

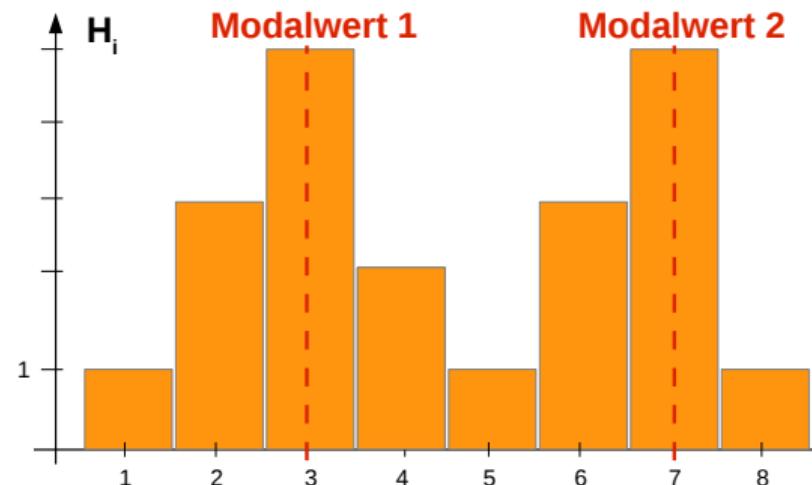
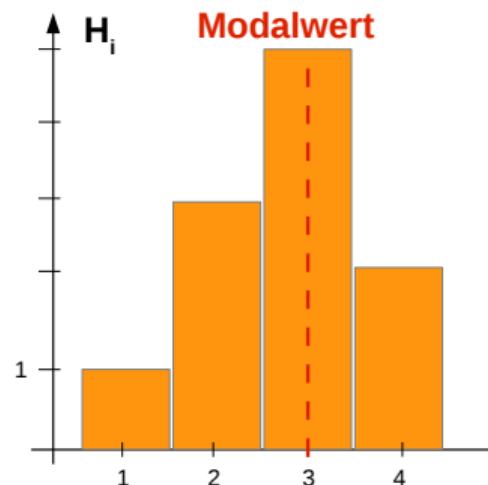
$$\Rightarrow \tilde{x}_{0,5} = \frac{1}{2} \cdot (x_7 + x_8) = \frac{1}{2} (19 + 21) = 20$$

# Lageparameter: Modalwert

Wir nennen den **häufigsten Wert** einer Stichprobe den **Modalwert**.

- Beispiel:  $x_1, \dots, x_{11} = 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4$
- Modalwert: 3 (kommt 5x vor)

Existieren mehrere Werte mit maximaler Häufigkeit, besitzt die Stichprobe mehrere Modalwerte. Wir sprechen von einer **multi-modalen** Stichprobe.



## Do-Lageparameter-Yourself Bild: [10]

Wir befragen 500 Einwohner Berlins nach ihrem **Jahreseinkommen** (in 1,000 EUR). Was ist (sehr wahrscheinlich) höher: Der **Mittelwert** oder der **Median**?





# Outline

1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression

# Mittelwert: Wiederholung

#	Stadt A	#	Stadt B
	57.7		51.3
	33.7		24.1
	85.1		86.9
	31.2		37.2
	66.9		110.7
	46.9		95.6
	57.2		23.2
	68.7		64.8
	50.2		15.2
	38.3		26.9

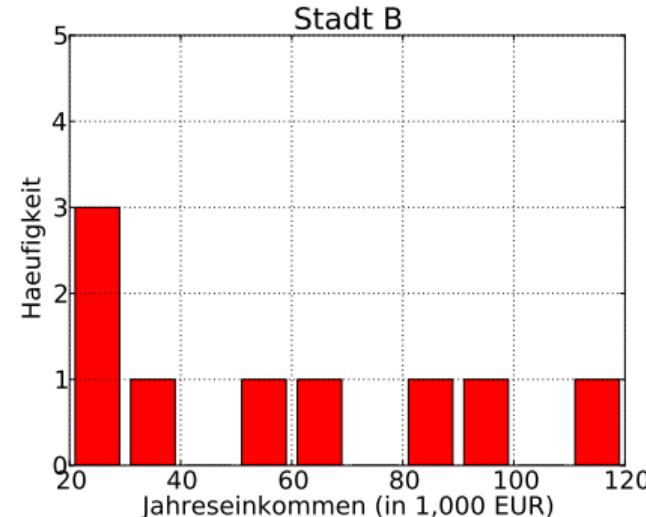
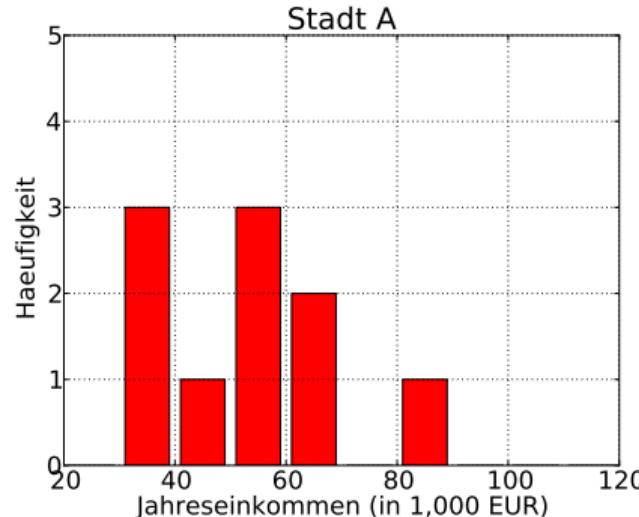
- ▶ Diese beiden Stichproben stellen das Jahreseinkommen der Einwohner zweier Deutscher Städte dar (in 1,000 EUR). Es wurden jeweils  $n = 10$  Einwohner befragt.

- ▶ Wir berechnen den **Mittelwert** der beiden Stichproben:

	Mittelwert $\bar{x}$
Stadt A	53.59
Stadt B	53.59

- ▶ Welcher Aspekt des Wohlstandes in den Städten ist **nicht berücksichtigt**?

# Streuungsparameter



- ▶ Lageparameter wie der Mittelwert geben Auskunft darüber wo die Daten liegen.
- ▶ Ein wichtiger Aspekt fehlt: Wie stark sind die Daten gestreut?
- ▶ Dies messen wir mit Hilfe von **Streuungsparametern**.

# Kennwerte: Spannweite

## Definition (Spannweite)

Gegeben eine sortierte Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$ , nennen wir

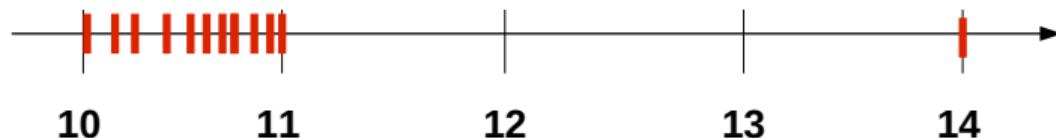
$$R := x_n - x_1$$

(d.h., Maximum – Minimum) die **Spannweite** (engl. “Range”) der Stichprobe.

## Beispiel

- ▶  $x_1, \dots, x_{14} = 10, 13, 14, 16, 18, 19, 19, 21, 23, 25, 28, 30, 36, 41$
- ▶  $R = x_n - x_1 = 41 - 10 = 31$

## Problem?



Die Spannweite ist sensitiv gegen Ausreißer. Hier ist die “eigentliche” Spannweite:  $R = 11 - 10 = 1$ , wir messen aber  $R = 14 - 10 = 4$ .

# Varianz und Standardabweichung

## Definition (Varianz)

Gegeben eine univariate Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$  mit Mittelwert  $\bar{x}$ , nennen wir

$$s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

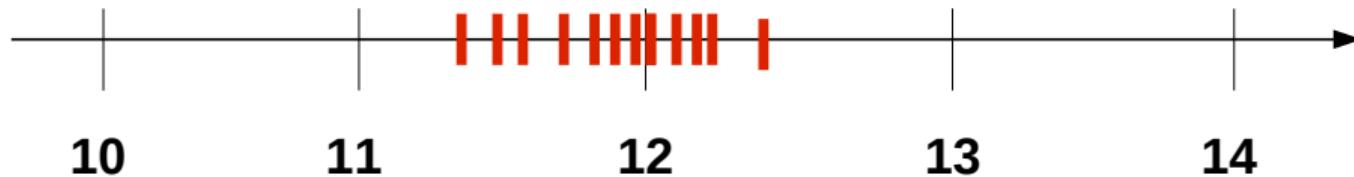
die **Varianz** (oder Stichprobenvarianz) der Stichprobe.

## Anmerkungen

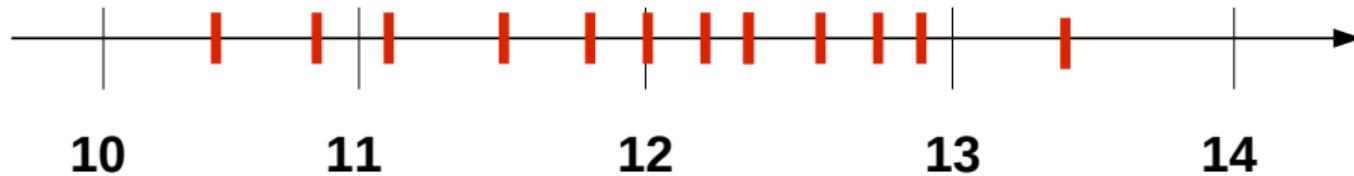
- ▶ Umgangssprachlich: "Varianz = durchschnittlicher quadratische Abstand vom Mittelwert".
- ▶ Wir nennen die Wurzel der Varianz,  $s$ , die **Standardabweichung** der Stichprobe.
- ▶ Die Standardabweichung liegt in **derselben Einheit** vor wie die Stichprobendaten. Sind die Daten z.B. in Metern erfasst, so liegt  $s$  ebenfalls in der Einheit *Meter* vor, aber  $s^2$  in *Meter<sup>2</sup>*.

# Varianz und Standardabweichung: Illustration

Geringere Varianz



Hohe Varianz



# Varianz und Standardabweichung: Beispiel

$$s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

# n=14

10

13

14

15

17

17

19

20

23

25

28

30

36

41

► Mittelwert:  $\bar{x} = \frac{1}{n} \sum_i x_i = 22$

► Varianz:

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

$$= \frac{1}{14} \left( (10 - 22)^2 + (13 - 22)^2 + \dots \right.$$

$$\left. \dots + (36 - 22)^2 + (41 - 22)^2 \right)$$

$$\approx 76.29$$

► Standardabweichung:

$$s \approx \sqrt{76.29} \approx 8.73$$



# Korrigierte Stichprobenvarianz

Eine alternative Definition der Varianz ist die **korrigierte** (Stichproben-)varianz:

## Definition (Korrigierte Stichprobenvarianz)

Gegeben eine univariate Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$  mit Mittelwert  $\bar{x}$   
und  $n > 1$ , nennen wir

$$s^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

die **korrigierte** (Stichproben-)Varianz.

## Anmerkungen

- Die korrigierte Stichprobenvarianz unterscheidet sich von der vorherigen Definition  $s^2$  lediglich im Normalisierungsfaktor ( $\frac{1}{n-1}$  statt  $\frac{1}{n}$ ).
- Sie bietet gegenüber  $s^2$  den Vorteil der **Erwartungstreue** (mehr hierzu später).

# Korrigierte Standardabweichung: Do-it-Yourself



# n=14

10

13

14

15

17

17

19

20

23

25

28

30

36

41

- ▶ Berechnen Sie  $s^*$  !
- ▶ **Tip:** Diese Stichprobe haben wir schon im Beispiel weiter oben betrachtet. Verwenden Sie Zwischenergebnisse soweit möglich!

## Varianz: Verschiebungssatz

In der obigen Formel benötigen wir zur Berechnung der Varianz **zwei Durchläufe** durch die Stichprobe:

1. Einen zur Bestimmung des Mittelwertes  $\bar{x}$
2. Einen (gegeben  $\bar{x}$ ) zur Bestimmung des mittleren quadratischen Abstands  $\frac{1}{n} \sum_i (x_i - \bar{x})^2$

### Problem

Diese Berechnung der Varianz ist ineffizient wenn...

- ▶ ... die Stichprobe groß (und/oder **verteilt**) ist
- ▶ ... die Stichprobe **dynamisch** ist  
(d.h., wenn ständig neue Werte hinzukommen)

Eine alternative Formel zur Berechnung der Varianz

bietet der **Verschiebungssatz**:

$$\left( \frac{1}{n} \sum_i (x_i - \bar{x})^2 \right) = S^2 = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2$$

*Verschiebungssatz*

# Varianz: Verschiebungssatz (Beweis)

$$S^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \quad // \text{ 2. binomische Formel}$$

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

$$= \frac{1}{n} \sum_i x_i^2 - 2 \cdot \bar{x} \cdot \underbrace{\frac{1}{n} \sum_i x_i}_{\bar{x}} + \frac{1}{n} \sum_{i=1}^n \bar{x}^2$$

$$\begin{aligned} &= \bar{x}^2 + \bar{x}^2 + \bar{x}^2 + \dots + \bar{x}^2 \\ &\quad \underbrace{\bar{x}^2 + \bar{x}^2 + \bar{x}^2 + \dots + \bar{x}^2}_{i=1 \quad i=2 \quad i=3 \quad \dots \quad i=n} \\ &= n \cdot \bar{x}^2 \end{aligned}$$

$$= \quad // \quad - 2 \cdot \bar{x} \cdot \bar{x} + \cancel{\frac{1}{n} \cdot n \cdot \bar{x}^2}$$

$$= \quad // \quad - \underbrace{2\bar{x}^2}_{\cancel{2\bar{x}^2}} + \cancel{\bar{x}^2}$$

$$= \frac{1}{n} \sum_i x_i^2 - \bar{x}^2$$

Mittelwert über  $x_i^2$ 
Mittelwert über  $x_i$



## Varianz: Verschiebungssatz (Beweis)



## Varianz: Verschiebungssatz (Diskussion)

# Varianz: Do-it-yourself



```
# Multivariate Stichprobe: Personen  
# Größe(cm) Gewicht(kg)  
172.51054875 89.81090441  
178.5491509 94.14875995  
160.19982596 77.61266935  
185.33582886 99.26311152  
173.42373218 78.43528082  
178.07276393 85.89384238  
171.39415797 81.06861227  
163.07000132 79.57634485  
178.97868362 87.06345319  
181.77268699 76.21846529  
165.12776354 78.48439432  
180.1249523 100.79476513  
160.15953819 76.28881635  
186.62205244 98.02854219  
178.1006582 94.72277617  
182.8521624 89.55199009  
..
```

Die Kennwerte dieses Datensatzes lauten:

	$\bar{x}$	$s^2$	$s$
Größe	174.8	70.2	8.4
Gewicht	86.7	71.4	8.5

Wie **verändern** sich die Werte, wenn wir ...

- ▶ die Größe in Metern (statt Zentimetern)
- ▶ das Gewicht in Gramm (statt Kilogramm)

angeben?

# Mittelwert und Varianz: Transformation

Im obigen Beispiel entspricht das Umrechnen der Einheiten einer **linearen Transformation** der Stichprobe. Wie verhalten sich **Mittelwert und Varianz** im Falle solcher Transformationen?

## Theorem (Mittelwert und Varianz linear transformierter Stichproben)

Es sei  $x_1, \dots, x_n \in \mathbb{R}$  eine univariate Stichprobe mit Mittelwert  $\bar{x}$  und Varianz  $s^2$ .

Außerdem seien  $\alpha, \beta \in \mathbb{R}$ . Wir definieren eine **linear transformierte Stichprobe**  $x'_1, x'_2, \dots, x'_n$  mit:

$$x'_i := \alpha \cdot x_i + \beta \quad \text{für alle } i = 1, \dots, n$$

Dann gilt für Mittelwert und Varianz dieser linear transformierten Stichprobe:

$$(a) \bar{x}' = \alpha \cdot \bar{x} + \beta$$

$$(b) s'^2 = \alpha^2 \cdot s^2$$

## Mittelwert und Varianz: Transformation

(Teil-) beweis (zu (b)) :

$$\begin{aligned}
 s'^2 &= \frac{1}{n} \cdot \sum_i (x'_i - \bar{x}')^2 \\
 &= \frac{1}{n} \cdot \sum_i (\cancel{\alpha \cdot x_i + \beta} - \cancel{(\alpha \cdot \bar{x} + \beta)})^2 \\
 &= \frac{1}{n} \cdot \sum_i (\alpha \cdot x_i - \alpha \cdot \bar{x})^2 \\
 &= \frac{1}{n} \sum_i (\alpha \cdot (x_i - \bar{x}))^2 \\
 &= \frac{1}{n} \sum_i \underbrace{\alpha^2}_{\text{K}} \cdot (x_i - \bar{x})^2 \\
 &= \alpha^2 \cdot \underbrace{\frac{1}{n} \sum_i (x_i - \bar{x})^2}_{S^2} \quad \checkmark
 \end{aligned}$$

## Mittelwert und Varianz: Transformation



## Mittelwert und Varianz: Transformation



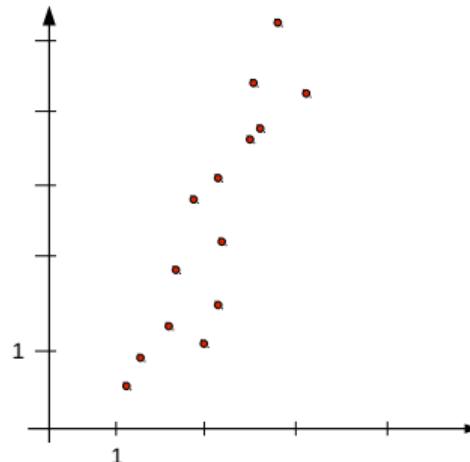


# Outline

1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression

# Varianz multivariater Stichproben?

Was können wir über die Streuung dieser **bivariaten** Stichprobe  $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$  aussagen?



## Zusammenhangsparameter

... helfen uns festzustellen ob eine **Abhängigkeit** zwischen zwei Merkmalen besteht.



# Kovarianz

Wir erweitern unsere Definition der Varianz, um die “**gemeinsame**” Streuung von  $x$  und  $y$  zu erfassen:

## Definition (Kovarianz)

Gegeben sei eine **bivariate Stichprobe**  $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$  mit Mittelwert  $(\bar{x}, \bar{y})$ .

Dann nennen wir

$$s_{xy} := \frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

die (Stichproben-) **Kovarianz**.

## Kovarianz: Do-it-yourself

$$s_{xy} := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

#	$x_i$	$y_i$
0	2	
1	6	
2	5	
4	13	
5	19	
6	21	

► Mittelwerte:

$$\blacktriangleright \bar{x} = \frac{1}{6}(0 + 1 + 2 + 4 + 5 + 6) = 3$$

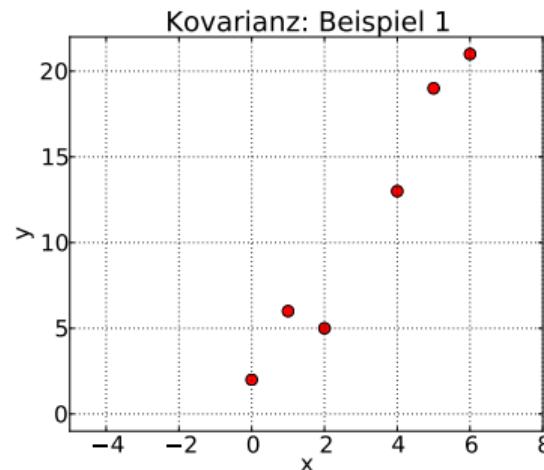
$$\blacktriangleright \bar{y} = \frac{1}{6}(2 + 6 + 5 + 13 + 19 + 21) = 11$$

► Kovarianz:

$$S_{xy} = \frac{1}{6} \cdot \left( (0-3) \cdot (2-11) + (1-3) \cdot (6-11) + (2-3) \cdot (5-11) + \dots + (6-3) \cdot (21-11) \right) \approx 15,17 > 0$$

# Beispiel 1: Kovarianz > 0

#	$x_i$	$y_i$
0	2	
1	6	
2	5	
4	13	
5	19	
6	21	



$s_{xy} > 0$  bedeutet: Mit wachsendem  $x$ -Wert wächst (tendenziell) auch der  $y$ -Wert.

## Beispiele

- ▶  $x_1, \dots, x_n = \text{Körpergröße}$ ,  $y_1, \dots, y_n = \text{Gewicht}$
- ▶  $x_1, \dots, x_n = \text{tägl. Regenmenge}$ ,  $y_1, \dots, y_n = \text{Schirmverkäufe}$

## Beispiel 2: Kovarianz $\approx 0$

#	$x_i$	$y_i$
1	6	
2	2	
4	9	
6	4	
7	5	
10	5	

### ► Mittelwerte

$$\blacktriangleright \bar{x} = \frac{1}{6}(1 + 2 + 4 + 6 + 7 + 10) = 5$$

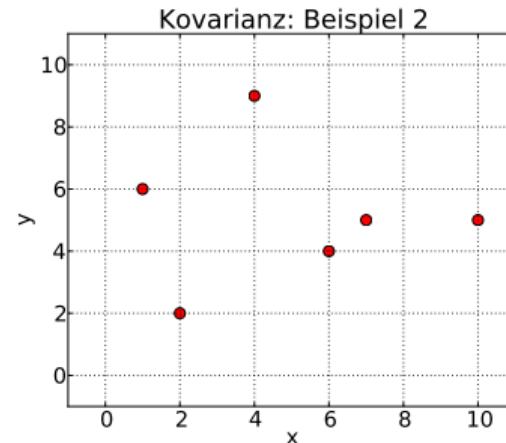
$$\blacktriangleright \bar{y} = \frac{1}{6}(6 + 2 + 9 + 4 + 5 + 5) \approx 5.2$$

### ► Kovarianz

$$\begin{aligned}
 s_{xy} &= \frac{1}{n} \cdot \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
 &\approx \frac{1}{6} \left( (1 - 5)(6 - 5.2) + (2 - 5)(2 - 5.2) + \dots \right. \\
 &\quad \left. \dots + (10 - 5)(5 - 5.2) \right) \\
 &\approx \frac{1}{6} \left( -3.3 + 9.5 - 3.8 - 1.2 - 0.3 - 0.8 \right) \\
 &= 0.08 \\
 &\approx 0
 \end{aligned}$$

## Beispiel 2: Kovarianz $\approx 0$

#	$x_i$	$y_i$
1	6	
2	2	
4	9	
6	4	
7	5	
10	5	



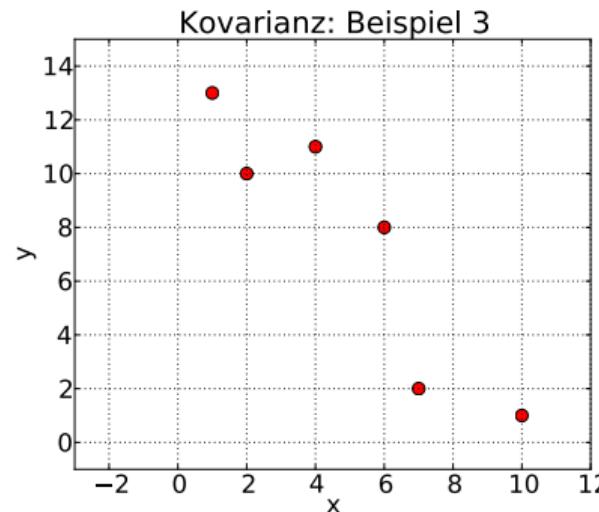
- ▶ Mit wachsendem  $x$ -Wert ändert sich der  $y$ -Wert **nicht** (bzw. nicht *stark*).

### Beispiele

- ▶ Würfelpaare:  $x_1, \dots, x_n = W1$ ,  $y_1, \dots, y_n = W2$
- ▶ Wirkungsloses Medikament:  $x_1, \dots, x_n = \text{Dosis}$ ,  $y_1, \dots, y_n = \text{Patientenbefinden}$

## Beispiel 3: Kovarianz < 0

#	$x_i$	$y_i$
1	13	
2	10	
4	11	
6	8	
7	2	
10	1	



$s_{xy} < 0$  bedeutet: Mit wachsendem  $x$ -Wert fällt der  $y$ -Wert

Beispiele

- ▶ Statistik-Klausur:  $x_1, \dots, x_n =$  Lernzeit,  $y_1, \dots, y_n =$  Note
- ▶ Gradientenabstieg:  $x_1, \dots, x_n =$  Anzahl der Iterationen,  $y_1, \dots, y_n =$  Fehler

# Von der Kovarianz zur Korrelation

- ▶ **Problem:** Die Kovarianz  $s_{xy}$  alleine ist nur bedingt aussagekräftig, da sie von der Größenordnung der Daten beeinflusst wird (vgl. Varianz).
- ▶ Um die Abhängigkeit zweier Variablen der Stichprobe generell zu untersuchen, **normalisieren** wir  $s_{xy}$  deshalb noch mit den Standardabweichungen  $s_x$  und  $s_y$ .

## Definition (Korrelation)

Gegeben sei eine bivariate Stichprobe  $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$ . Die Varianzen der  $x$ - und  $y$ -Werte seien  $s_x^2$  und  $s_y^2$ . Dann nennen wir

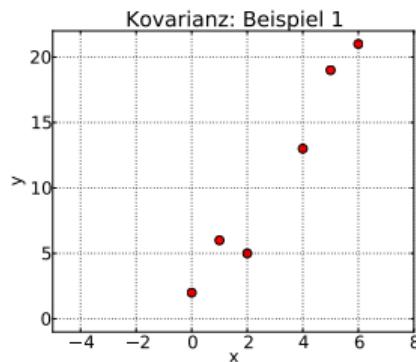
$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

die **Korrelation** zwischen  $x_1, \dots, x_n$  und  $y_1, \dots, y_n$ .

# Korrelation: Beispiele (siehe eben)

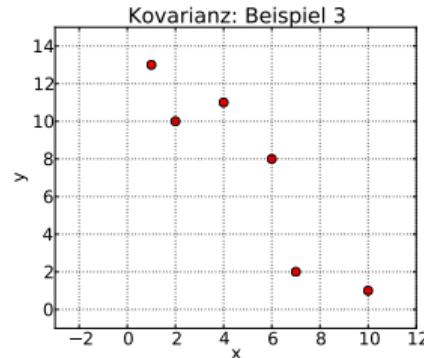
#	$x_i$	$y_i$
0	2	
1	6	
2	5	
4	13	
5	19	
6	21	

- ▶ Kovarianz:  $s_{xy} \approx 15.2$
- ▶ Varianzen:  $s_x^2 \approx 4.7, s_y^2 \approx 51.7$
- ▶ Standardabw.:  $s_x \approx 2.2, s_y \approx 7.2$
- ▶ Korrelation:  $r_{xy} \approx \frac{15.2}{2.2 \cdot 7.2} \approx \mathbf{0.96}$

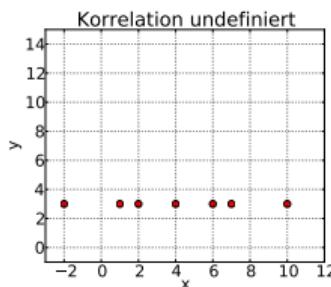
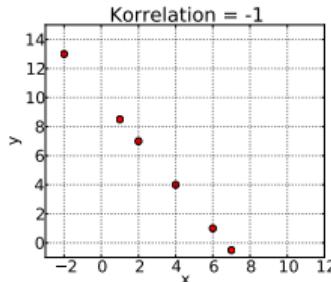
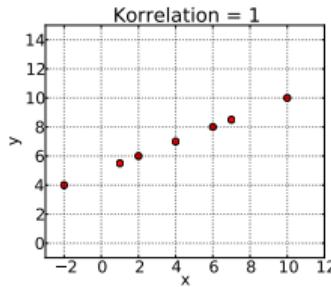


#	$x_i$	$y_i$
1	13	
2	10	
4	11	
6	8	
7	2	
10	1	

- ▶ Kovarianz:  $s_{xy} \approx -12.7$
- ▶ Varianzen:  $s_x^2 \approx 9.3, s_y^2 \approx 20.3$
- ▶ Standardabw.:  $s_x \approx 3.1, s_y \approx 4.5$
- ▶ Korrelation:  $r_{xy} \approx \frac{-12.7}{3.1 \cdot 4.5} \approx \mathbf{-0.91}$



# Korrelation: Eigenschaften



- ▶ Die Korrelation  $r_{xy}$  ist ein Maß für die **lineare Abhängigkeit** zwischen den Variablen einer Stichprobe.
- ▶ Die Korrelation liegt immer zwischen  $-1$  und  $1$ .
- ▶ Wenn  $|r_{xy}| = 1$ , liegen die Punkte der Stichprobe **exakt auf einer Geraden**, d.h. es gilt (mit  $a, b \in \mathbb{R}$ ):

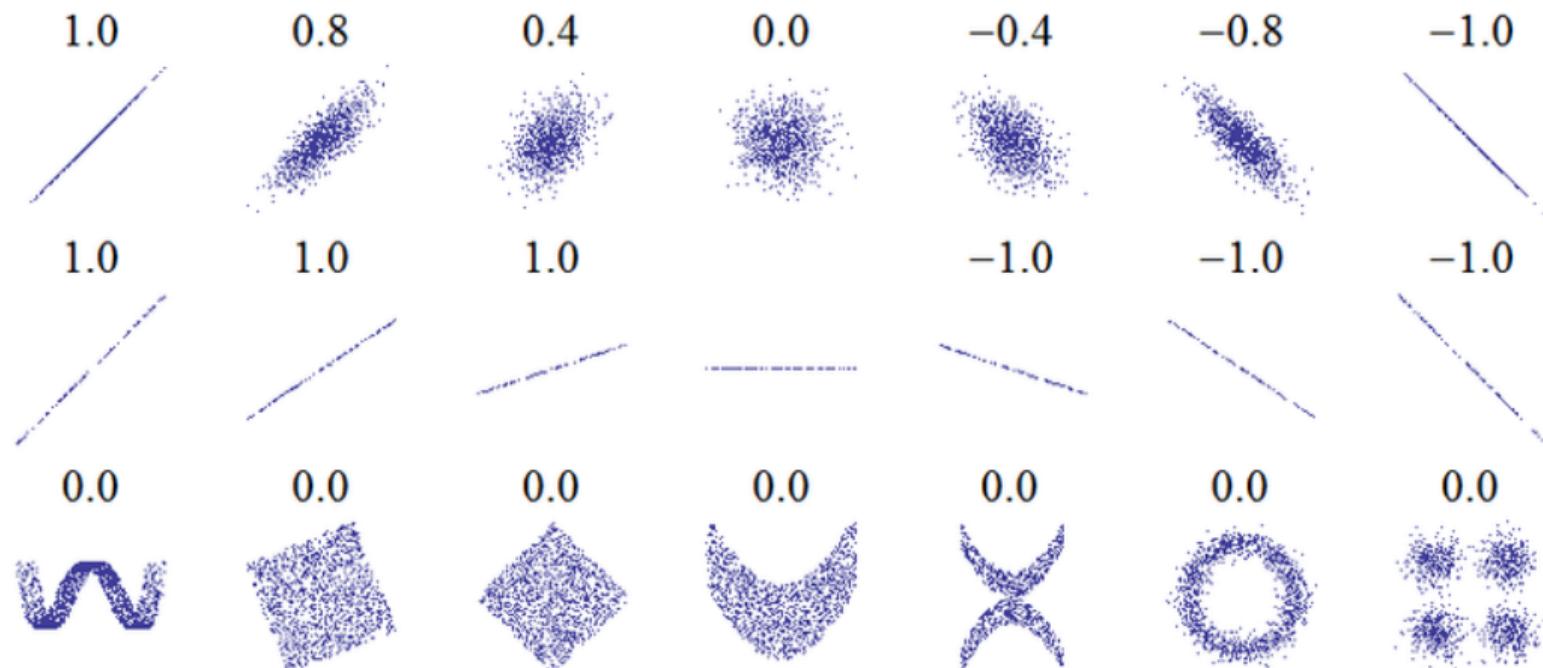
$$y_i = a \cdot x_i + b \quad \text{für alle } i = 1, \dots, n$$

Wir unterscheiden 3 Fälle:

- ▶  $a > 0 \Rightarrow r_{xy} = 1$
- ▶  $a < 0 \Rightarrow r_{xy} = -1$
- ▶  $a = 0 \Rightarrow r_{xy} = \uparrow$  (*man kann nicht durch  $s_y=0$  teilen*)

## Korrelation: Beispiele Bild: [17]

Einige Beispiele bivariater Stichproben mit zugehöriger Korrelation:

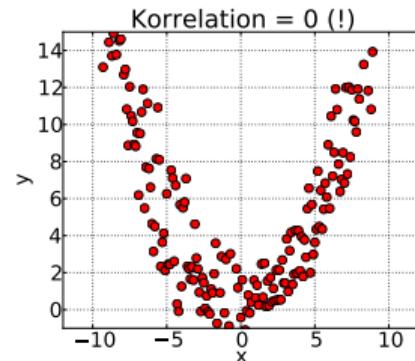


# Korrelation=0 bedeutet keine Unabhängigkeit

Falls  $r_{xy} = 0$ , sagen wir: "Es ist keine **lineare** Abhängigkeit nachweisbar".  
Bedeutet dies, dass **gar keine Abhängigkeit** besteht? **Nein!**

## Beispiel

- ▶ Im Plot unten liegt keine Korrelation vor,  
also keine **lineare Abhängigkeit**
- ▶ Dennoch besteht hier eine Abhängigkeit zwischen  
 $x$  und  $y$  (je größer  $|x|$ , desto größer  $y$ )!



## Korrelation: Eigenschaften (cont'd)

- Ab welchem Wert von  $|r_{xy}|$  gehen wir von einer Abhangigkeit zwischen den gemessenen Groen aus?

*“Ob ein gemessener Korrelationskoeffizient als gro oder klein interpretiert wird, hangt stark von der **Art der untersuchten Daten** ab. Bei psychologischen Fragebogendaten werden z. B. Werte bis ca. ±0.3 haufig als klein angesehen, ab ca. ±0.5 als gut, wahrend man ab ca. ±0.7 von einer (sehr) hohen Korrelation spricht.”*

(Wikipedia)

- Ob eine beobachtete Korrelation **signifikant** ist, hangt von weiteren Parametern der Stichprobenerhebung ab, z.B. von der **Stichprobengroe** (*Signifikanztests: spater*).

# Korrelation: Do-it-yourself



Welche der folgenden Größen sind wie korreliert?

- ▶  $x$  = Laufzeit eines Sortieralgoritmus  
 $y$  = Größe der Eingabe
  
- ▶  $x$  = Größe des Hauptspeichers  
 $y$  = Anzahl der Page Faults
  
- ▶  $x$  = Anzahl der Haare eines Mannes  
 $y$  = Einkommen des Mannes

# Korrelation $\iff$ Kausalität

- ▶ **Achtung:** Wenn wir eine **starke** Korrelation messen, deutet dies noch nicht auf einen **kausalen Zusammenhang** hin!
- ▶ Das heißt: Weder muss Variable 1 die **Ursache** für das Verhalten von Variable 2 sein, noch Variable 2 die Ursache für das Verhalten von Variable 1.

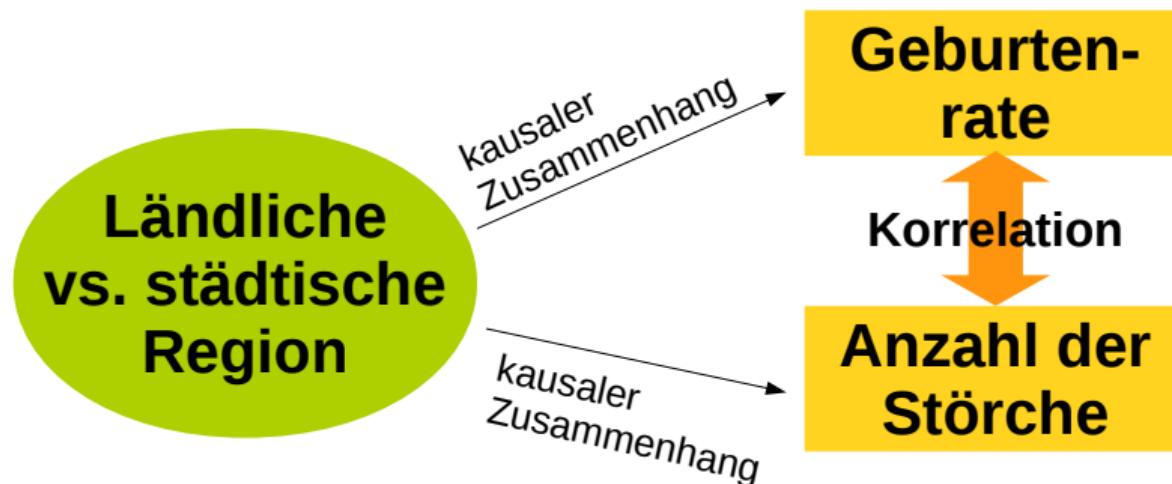
## Beispiele

- ▶ Die **Anzahl der Haare** eines Mannes ist negativ korreliert mit seinem **Einkommen**.
- ▶ Die **Geburtenquote** von Gemeinden ist positiv korreliert mit der Anzahl der ansässigen **Storchenpaare**.

Warum?

# Korrelation $\iff$ Kausalität

- ▶ Die Korrelation entsteht hier **nicht** durch einen kausalen Zusammenhang: Personen sind nicht wohlhabender **weil** sie wenige Haare haben.
- ▶ Grund ist eine andere, *latente* Variable – hier das Alter.



# Kovarianz multivariater Stichproben

- Wie messen wir Zusammenhänge in Stichproben mit **mehr als zwei Merkmalen**?
- Hierfür fassen wir **Kovarianzen** in der sogenannten **Kovarianzmatrix** zusammen.

## Definition (Kovarianz multivariater Stichproben)

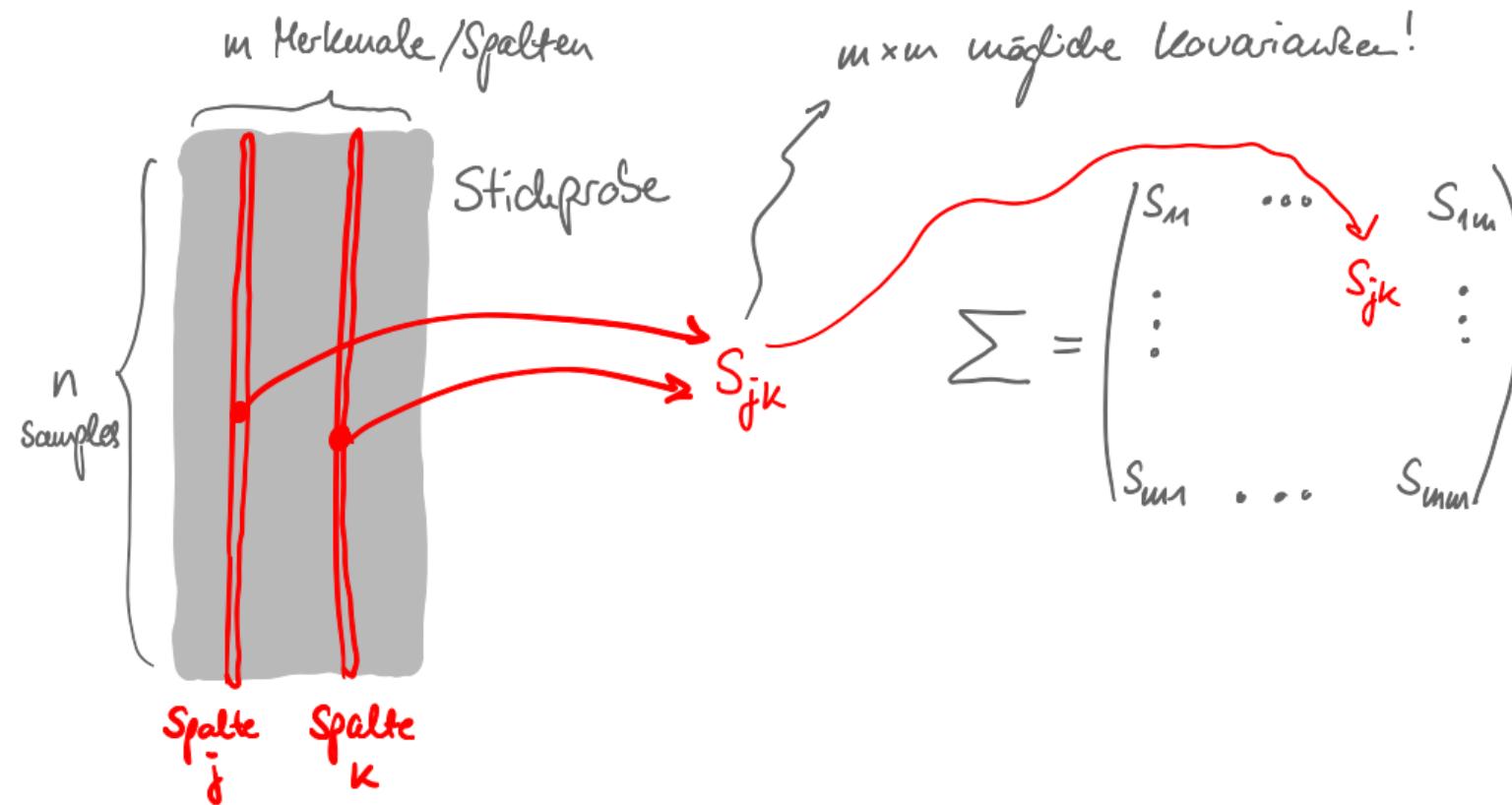
Gegeben ist eine Stichprobe  $x_1, \dots, x_n \in \mathbb{R}^m$  mit **m Merkmalen**. Wir notieren einen Datenpunkt als  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$ ). Der Mittelwert der Stichprobe sei  $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m)$ . Dann nennen wir

$$S_{kj} = s_{jk} = \frac{1}{n} \cdot \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad \text{mit } 1 \leq j, k \leq m$$


die **Kovarianz** der Stichprobe zwischen den Merkmalen/Spalten **j und k**.

# Die Kovarianzmatrix

Wir fassen alle Kovarianzen  $s_{jk}$  in der **Kovarianzmatrix  $\Sigma$**  zusammen:



## Eigenschaften der Kovarianzmatrix $\Sigma$

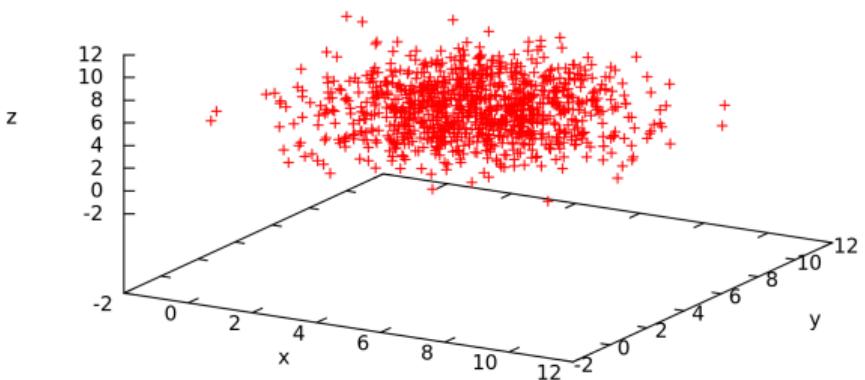
\*

- $\Sigma$  hat  $m$  Zeilen und  $m$  Spalten.
- $\Sigma$  ist symmetrisch (weil  $S_{jk} = S_{kj}$ )
- Auf der Diagonale von  $\Sigma$  stehen die Varianzen der einzelnen Merkmale:  $\sum_{jj} = s_j^{-2}$  für alle  $j = 1, \dots, m$



## Eigenschaften der Kovarianzmatrix

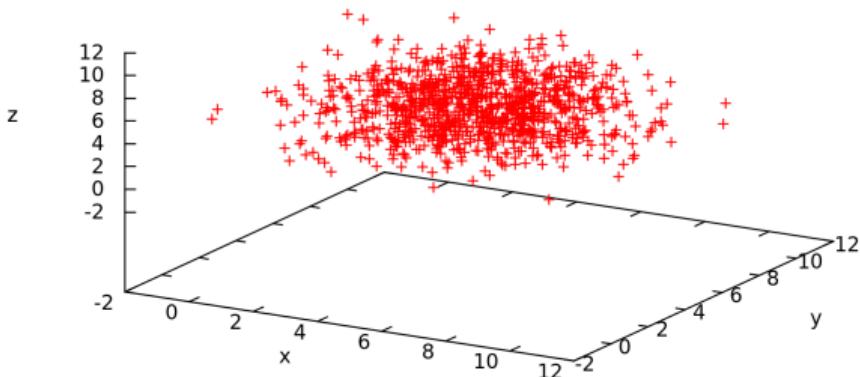
# Kennwerte multivariater Stichproben: Beispiel



Beispiel: Eine Stichprobe mit 3 Merkmalen

- ▶ Wie lautet der Mittelwert?
- ▶ Ordnen Sie die Varianzen  $s_1^2, s_2^2, s_3^2$  den Werten 2, 5, 2 zu!
- ▶ Ordnen Sie die Kovarianzen  $s_{12}, s_{23}, s_{13}$  den Werten 0, 0, 1.8 zu!

# Kennwerte multivariater Stichproben: Beispiel



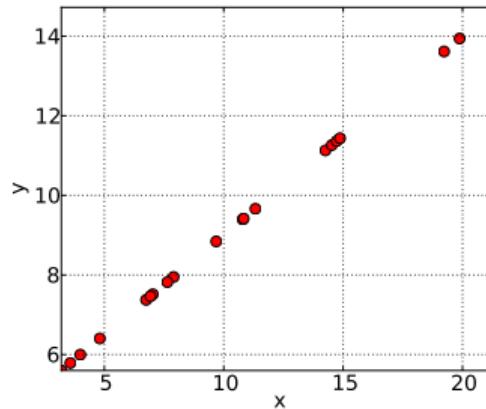
Wie lautet die **Kovarianzmatrix**?



# Outline

1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression

# Regressionsprobleme

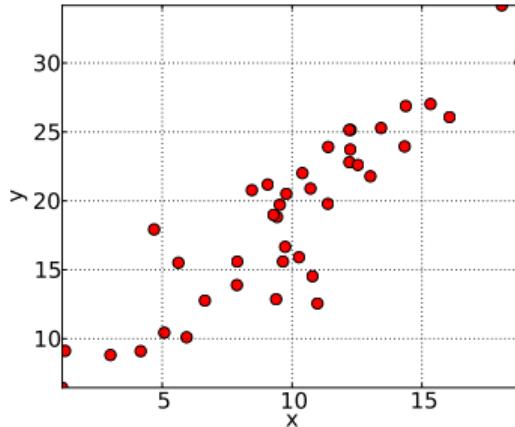


Unser Ziel im Folgenden ist es, eine **Funktion** (z.B. eine **Gerade**) zu ermitteln, die möglichst gut auf die **Stichprobe  $x_1, \dots, x_n$**  passt.

## Anwendungsfälle

- ▶ Trends vorhersagen (z.B. Preise, Aktienkurse)
- ▶ Fehlende Werte der Stichprobe füllen (*engl. imputation*)
- ▶ Ausreißer identifizieren.

# Regressionsprobleme



- ▶ **Problem:** In der Praxis liegen die Samples **nicht genau** auf einer Geraden.
- ▶ **Grund:** Messfehler, Rauschen, Varianz in den Daten, ...
- ▶ *Wie können wir in diesen Fällen eine passende Gerade (eine sogenannte "**Ausgleichsgerade**") ermitteln?*
- ▶ **Ansatz:** **Lineare Regression** mit der sog. Methode der kleinsten Quadrate (*engl. "**Least Squares**"*).

## “Least Squares”: Ansatz

- ▶ **Ansatz:** Wir nähern die Stichprobe  $(x_1, y_1), \dots, (x_n, y_n)$  durch eine **Funktion** (ein **Modell**)  $\mathcal{M}_\theta : \mathbb{R} \rightarrow \mathbb{R}$  an.
- ▶ Die Funktion ordnet jedem  $x$  einen  $y$ -Wert  $\mathcal{M}_\theta(x)$  zu. Sie besitzt **Parameter**  $\theta$ .
- ▶ Wir wählen als Modell  $\mathcal{M}_\theta$  eine **Gerade**. Die Parameter sind *Steigung*  $a$  und *Achsenabschnitt*  $b$ , d.h.  $\theta = (a, b)$ .
- ▶ **Anmerkung:** Least Squares kann auch mit anderen Modellen (z.B. Parabeln ...) angewandt werden – *siehe Übung*.
- ▶ Unser Ziel ist es, die “besten” **Parameter  $\theta$  zu ermitteln**, d.h.  $a$  und  $b$  auf die Stichprobe zu **fitten**, so dass:

$$y_i \approx \underbrace{a \cdot x_i + b}_{\mathcal{M}_\theta(x_i)} \quad \text{für } i = 1, \dots, n$$

## “Least Squares”: Ansatz

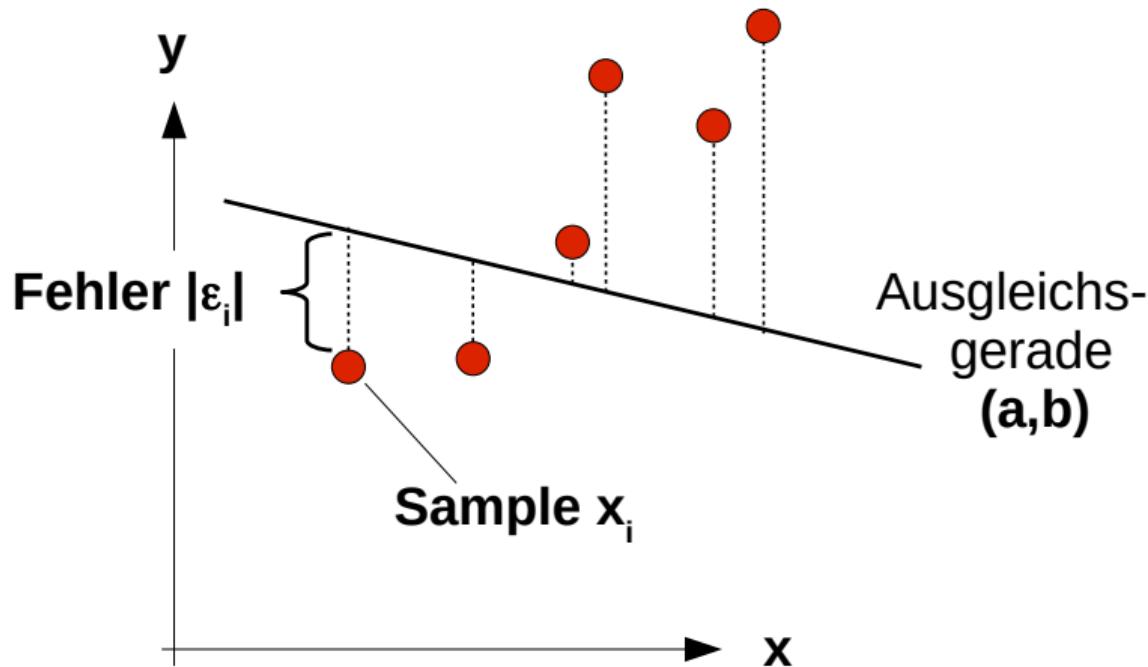
- ▶ **Idee:** Die besten Parameter  $\theta$  sind jene, für welche die Gerade **möglichst nahe an den Datenpunkten** liegt.
- ▶ Um diese “Nähe” zu messen, führen wir **Fehlervariablen**  $\epsilon_1, \dots, \epsilon_n$  ein.  
Diese definieren wir folgendermaßen:

$$y_i = \underbrace{a \cdot x_i + b}_{\mathcal{M}_\theta(x_i)} + \epsilon_i \quad \text{für } i = 1, \dots, n$$

- ▶ Unser Ziel ist es, eine Ausgleichsgerade zu finden, so dass die Fehler  $\epsilon_1, \dots, \epsilon_n$  **möglichst nahe null** sind!

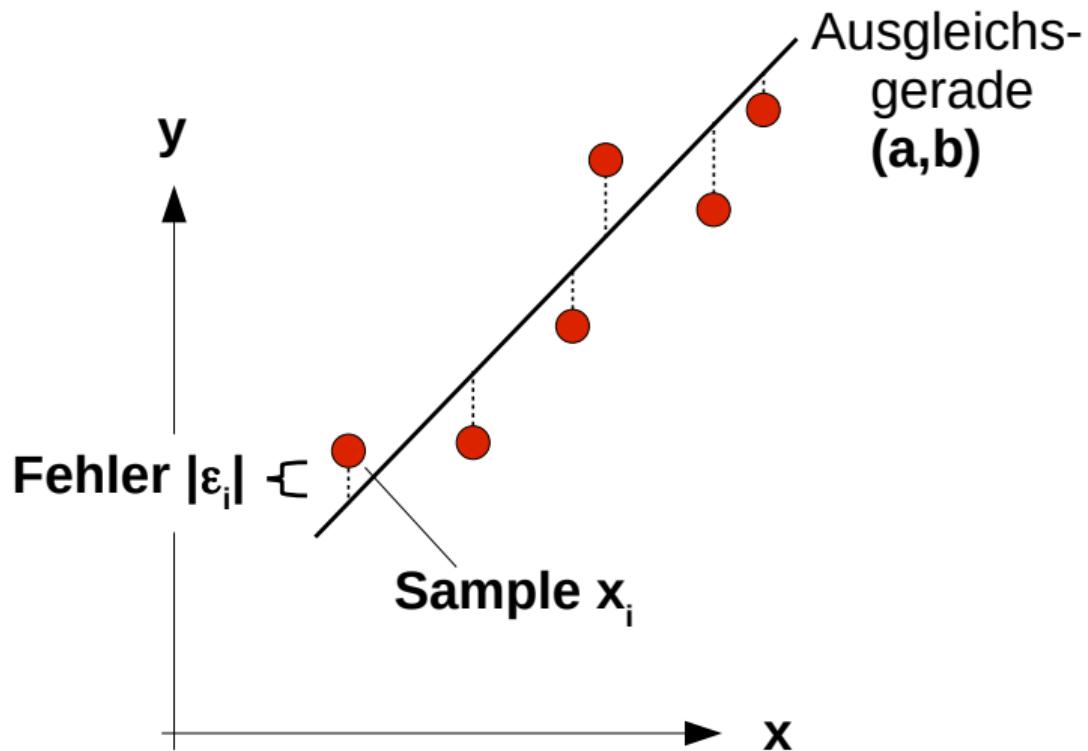
## "Least Squares": Beispiel 1

Die Fehler  $\epsilon_1, \dots, \epsilon_n$  sind nicht nahe null  $\rightarrow$  schlechte Gerade.



## "Least Squares": Beispiel 2

Die Fehler  $\epsilon_1, \dots, \epsilon_n$  sind nahe null  $\rightarrow$  gute Gerade.



## “Least Squares”: Genereller Ansatz

Wir definieren eine **Fehlerfunktion**, die die **quadratischen Fehler** der Messwerte enthält:

$$E(a, b) := \sum_i \epsilon_i^2$$

Als Ausgleichsgerade wählen wir die Gerade, die  $E(a, b)$  **minimiert** !  
 (deshalb der Name “least squares”):

$$\begin{aligned} *^{\text{Lösung, Ergebnis}} \hat{a}, \hat{b}^* &= \underset{a, b}{\operatorname{arg\min}} E(a, b) \\ &= " \sum_i \epsilon_i^2 \\ &= " \sum_i (\underbrace{ax_i + b}_{M_\theta(x_i)} - y_i)^2 \end{aligned}$$

# "Least Squares": Herleitung

\*

→ Minimum?

$$E(a, b) = \sum_i ((ax_i + b) - y_i)^2$$

(I)  $E_a = \sum_i \frac{1}{2} \cdot ((ax_i + b) - y_i) \cdot x_i \stackrel{!}{=} 0 \quad /:2$

$$a \cdot \sum_i x_i^2 + b \cdot \sum_i x_i - \sum_i x_i y_i = 0$$

→ Lineare Gleichung  
mit a, b als Unbekannte

(II)  $E_b = \sum_i \frac{1}{2} \cdot ((ax_i + b) - y_i) \cdot 1 \stackrel{!}{=} 0 \quad /:2$

$$a \cdot \sum_i x_i + b \cdot \sum_i 1 - \sum_i y_i = 0$$

... bekannt!

2x2-LGS mit Unbe-  
kannten a, b → lösen!

## “Least Squares”: Herleitung



## “Least Squares”: Herleitung



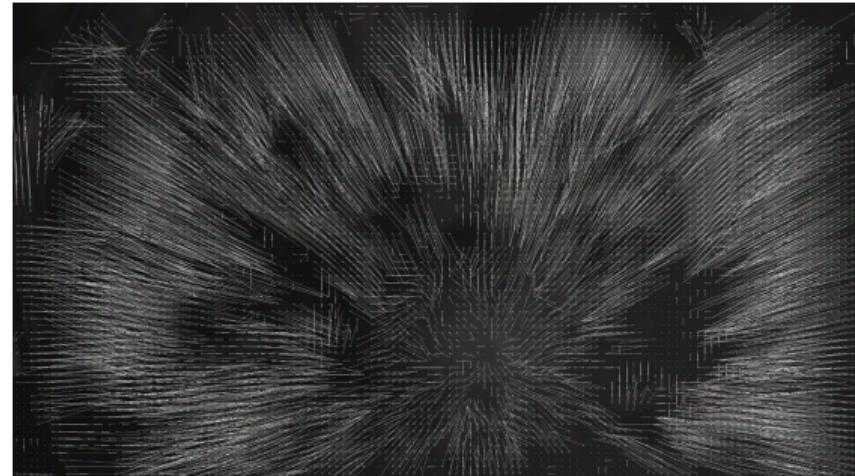
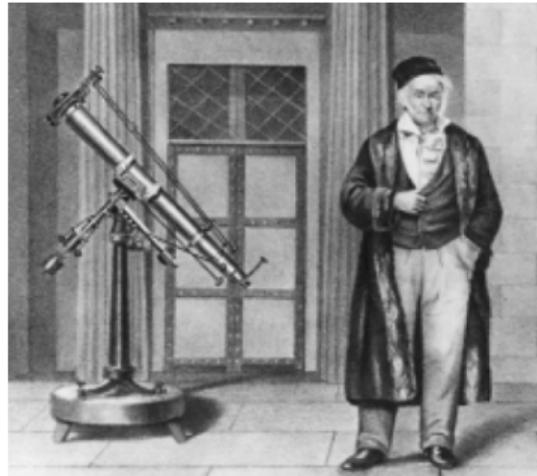
## “Least Squares”: Genereller Ansatz Bilder: [7] [4]

Wir erhalten also zwei lineare Gleichungen mit zwei Unbekannten  $a$ ,  $b$ .

Es ergibt sich (nach einigen Umformungen<sup>2</sup>):

$$a^* = s_{xy} / s_x^2$$

$$b^* = \bar{y} - a^* \cdot \bar{x}$$



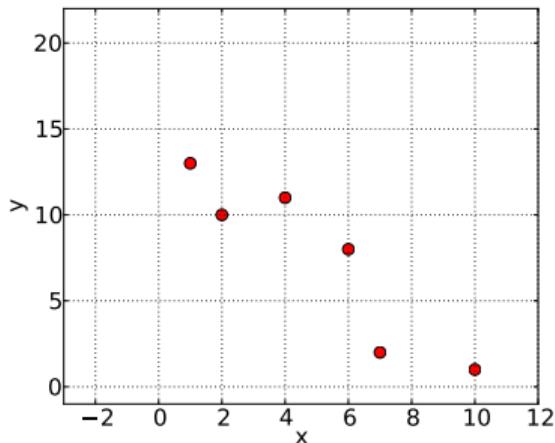
<sup>2</sup>siehe Teschl, Teschl: Mathematik für Informatiker, Band 2, 164ff

# Do-Lineare Regression-Yourself

Ermitteln und skizzieren Sie die Ausgleichsgerade!



#	$x_i$	$y_i$
1	13	
2	10	
4	11	
6	8	
7	2	
10	1	



# Do-Least-Squares-Yourself



# Do-Least-Squares-Yourself





# Outline

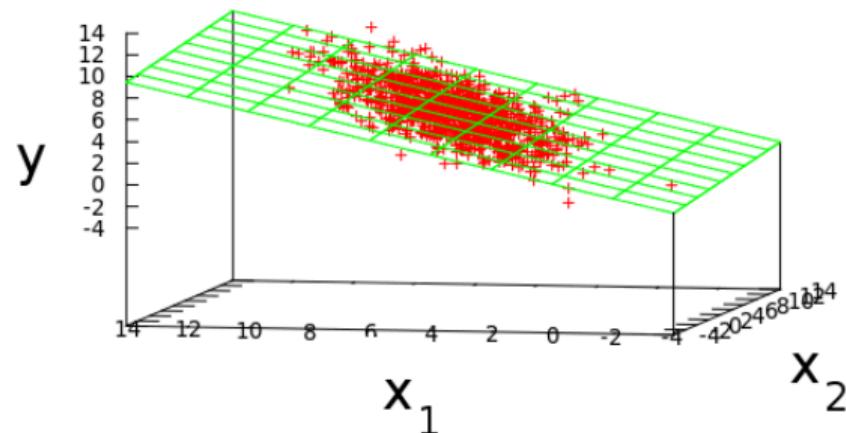
1. Motivation
2. Grundbegriffe
3. Lageparameter
4. Streuungsparameter
5. Zusammenhangsparameter
6. Lineare Regression
7. Multiple Lineare Regression

# Multiple Lineare Regression

- ▶ Stichproben/Probleme sind in der Praxis meist **multivariat!**
- ▶ **Beispiel:** Sage den Endpreis einer Auktion auf eBay voraus  
*(verkauft wird ein Auto mit Baujahr, Marke, Verkäuferbewertungen, ...)*

## Ziel

Vorhersage einer Variable  $y$  (z.B. Preis), gegeben **mehrere** andere Variablen  $x_1, \dots, x_{m-1}$ .



# Multiple Lineare Regression: Herleitung

Gegeben einen **Eingabevektor**  $\mathbf{x} = (x_1, \dots, x_{m-1})$ , sagen wir den Wert  $y$  voraus:

$$y := \underbrace{w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_{m-1} \cdot x_{m-1}}_{\mathcal{M}_w(\mathbf{x})} + b.$$

Unser Modell ist eine **Hyperebene** mit Parametern  $w_1, \dots, w_{m-1}, b$ .

## Vereinfachung der Notation

- ▶ Damit der Parameter  $b$  entfällt, fügen wir dem Eingabevektor einen zusätzlichen Wert  $x_m := 1$  hinzu.
- ▶ Als einziger Parameter bleibt ein **Gewichtsvektor**  $\mathbf{w} = (w_1, \dots, w_m)$ :

$$y := w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_{m-1} \cdot x_{m-1} + \underbrace{w_m \cdot x_m}_{\text{entspricht } b \cdot 1} = \mathbf{w} \cdot \mathbf{x}.$$

# Multiple Lineare Regression: Herleitung

- ▶ Gegeben ist eine multivariate Stichprobe  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . Jedes Sample der Stichprobe besteht aus  $m$  Merkmalen:  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$ .
- ▶ Wir fassen die Daten zu einer **Matrix  $X$**  zusammen  
*(mit den Samples als Zeilen):*

$$X := \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

- ▶ Zu den Samples existieren auch  $y$ -Werte  $y_1, \dots, y_n \in \mathbb{R}$ .  
Unser Ziel ist es, diese anzunähern. Wir fassen sie in einem Vektor  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  zusammen.

# Multiple Lineare Regression: Herleitung

- Wir formulieren – analog zur Ausgleichsgerade (s.o.) – eine Fehlerfunktion  $E$ :

$$E(\mathbf{w}) = \sum_{i=1}^n (\underbrace{\mathbf{w} \cdot \mathbf{x}_i - y_i}_{\mathcal{M}_{\mathbf{w}}(\mathbf{x}_i)})^2 = \sum_{i=1}^n (\underbrace{w_1 \cdot x_{i1} + w_2 \cdot x_{i2} + \dots + w_m \cdot x_{im} - y_i}_{\mathcal{M}_{\mathbf{w}}(\mathbf{x}_i)})^2$$

- $E$  bewertet den Fehler einer Hyperebene  $\mathbf{w}$ . Wir bestimmen das Minimum, indem wir nach  $w_1, w_2, \dots, w_m$  ableiten (*hier die Ableitung für  $w_k$* ) ...

$$\frac{\partial E}{\partial w_k} = \sum_{i=1}^n 2 \cdot (\mathbf{w} \cdot \mathbf{x}_i - y_i) \cdot x_{ik}$$

- ... und die Ableitung gleich null setzen:

$$\sum_{i=1}^n 2 \cdot (\mathbf{w} \cdot \mathbf{x}_i - y_i) \cdot x_{ik} \stackrel{!}{=} 0$$

## Multiple Lineare Regression: Herleitung

Für alle  $k = 1, \dots, m$  gilt:

$$\sum_{i=1}^n (wx_i - y_i) \cdot x_{ik} = 0$$

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1k} & \dots & x_{1m} \\ x_{21} & \dots & x_{2k} & \dots & x_{2m} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nk} & \dots & x_{nm} \end{pmatrix}$$

K-te Spalte =  $x_k$

$$\sum_i w_1 \cdot x_{i1} \cdot x_{ik} + w_2 \cdot x_{i2} \cdot x_{ik} + \dots + w_m \cdot x_{im} \cdot x_{ik} = \sum_i y_i \cdot x_{ik}$$

$$w_1 \cdot \sum_i x_{i1} \cdot x_{ik} + w_2 \cdot \sum_i x_{i2} \cdot x_{ik} + \dots + w_m \cdot \sum_i x_{im} \cdot x_{ik} = \sum_i y_i \cdot x_{ik}$$

$$w_1 \cdot (x_k \cdot X_1) + w_2 \cdot (x_k \cdot X_2) + \dots + w_m \cdot (x_k \cdot X_m) = (x_k \cdot y)$$

## Multiple Lineare Regression: Herleitung

Wir fassen die Gleichungen für  $k=1, \dots, m$  in einem linearen Gleichungssystem zusammen:

$$\begin{matrix}
 k=1 & X_1 \cdot X_1 & X_1 \cdot X_2 & \dots & X_1 \cdot X_m \\
 k=2 & X_2 \cdot X_1 & X_2 \cdot X_2 & \dots & X_2 \cdot X_m \\
 & \vdots & & & \\
 k=m & X_m \cdot X_1 & X_m \cdot X_2 & \dots & X_m \cdot X_m
 \end{matrix} \cdot W = \begin{pmatrix} X_1 \cdot y \\ X_2 \cdot y \\ \vdots \\ X_m \cdot y \end{pmatrix}$$

$$X^T \cdot X \cdot W = X^T \cdot y$$

$\Rightarrow$  LGS lösen  $\Rightarrow W \quad \ddot{u}$

# Multiple Lineare Regression: Code-Beispiel



The image displays a 4x5 grid of screenshots from various software environments, likely demonstrating different approaches or tools used in the development of a multiple linear regression model. The screenshots include:

- Top row:
  - NetBeans IDE showing Java code for a regression model.
  - Ubuntu terminal showing command-line execution of a regression script.
  - IntelliJ IDEA showing Java code for a regression model.
  - Google Sheets showing a data analysis interface.
  - Ubuntu terminal showing command-line execution of a regression script.
- Second row:
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
- Third row:
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
- Bottom row:
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.
  - Ubuntu terminal showing command-line execution of a regression script.

# Multiple Lineare Regression: Anmerkungen

- Die **Regressionsgewichte**  $w_k$  zeigen den Einfluss jedes Merkmals  $k$ !
- In der Praxis versuchen wir, **so viele “nützliche” Merkmale** wie möglich zu finden.

## Beispiel: eBay Preisprognose

Über die eBay-API erhalten wir Merkmale zur Beschreibung von Auktionen. Wir trainieren ein Modell für die **Vorhersage von Auktionspreisen**, und inspizieren die resultierenden **Gewichte**:

<b>Feature <math>x_i</math></b>	<b>weight <math>w_i</math></b>
<b>SELLER_RATING</b>	4.572
VENDOR_ID_ONEHOT	3.812
CONDITION_ID	3.674
...	

<b>Feature <math>x_i</math></b>	<b>weight <math>w_i</math></b>
...	
SELLER_REG_STATE_ADDR	0.024
<b>SHIPPING_INCL_FLAG</b>	0.001
DISPLAY_IS_GLARE_FLAG	-0.019
...	

<b>Feature <math>x_i</math></b>	<b>weight <math>w_i</math></b>
...	
<b>INDESCR_CNT_GEBRAUCHT</b>	<b>-2.475</b>
RELEASE_AGE_MONTHS	-5.107
...	

- **Links:** Die Bewertung des Verkäufers ist wichtig für einen hohen Preis.
- **Mitte:** Ist der Versand inklusive, steigt die Preisprognose minimal.
- **Rechts:** Kommt das Wort “gebraucht” im Text vor, ist der Preis niedrig.



# References I

- [1] A surface weather analysis for the United States on October 21, 2006.  
[https://en.wikipedia.org/wiki/Weather\\_map](https://en.wikipedia.org/wiki/Weather_map) (retrieved: Oct 2016).
- [2] Aaron Parecki: Face Detection.  
<https://www.flickr.com/photos/aaronpk/6706242723> (retrieved: Oct 2016).
- [3] Ars Electronica: ADM8.  
<https://www.flickr.com/photos/arselectronica/7650332104> (retrieved: Oct 2016).
- [4] Blender Foundation / Netherlands Media Art Institute: Bewegungsvektoren, die eine schnelle Kamerafahrt auf ein Ziel unten-mittig im Bild verursacht hat.  
[https://de.wikipedia.org/wiki/Optischer\\_Fluss](https://de.wikipedia.org/wiki/Optischer_Fluss) (retrieved: Oct 2016).
- [5] Euro 2012 : Le bal des entraîneurs.  
<http://www.oldschoolpanini.com/2012/06/euro-2012-le-bal-des-entraineurs.html> (retrieved: Oct 2016).
- [6] Hospital Dashboard / Clinical Dashboard Metrics.  
<http://www.dashboardzone.com/hospital-dashboard-clinical-dashboard-metrics> (retrieved: Oct 2016).
- [7] Le Corvec et al.: How Gauss Determined the Orbit of Ceres.  
[https://math.berkeley.edu/~mgu/MA221/Ceres\\_Presentation.pdf](https://math.berkeley.edu/~mgu/MA221/Ceres_Presentation.pdf) (retrieved: Oct 2016).
- [8] Nationwide Poll Results for 2008 Presidential Election.  
[https://commons.wikimedia.org/wiki/File:Nationwide\\_Poll\\_Results\\_for\\_2008\\_Presidential\\_Election.svg](https://commons.wikimedia.org/wiki/File:Nationwide_Poll_Results_for_2008_Presidential_Election.svg) (retrieved: Oct 2016).
- [9] OpenStreetMap-Stadtplan von Hamburg auf einem Smartphone.  
[https://de.wikipedia.org/wiki/Stadtplan#/media/File:Osmand\\_auf\\_Samsung\\_Galaxy\\_S\\_Advance\\_\(Hamburg\).jpg](https://de.wikipedia.org/wiki/Stadtplan#/media/File:Osmand_auf_Samsung_Galaxy_S_Advance_(Hamburg).jpg) (retrieved: Oct 2016).
- [10] Studie: "Kreditschwemme" kommt beim Mittelstand nicht an .  
<http://www.wirtschaft.com/studie-kreditschwemme-kommt-beim-mittelstand-nicht/> (retrieved: Oct 2016).

# References II



- [11] **The MNIST Database of Handwritten Digits.**  
<http://yann.lecun.com/exdb/mnist/> (retrieved: Oct 2016).
- [12] **Torley: Improving Amazon's Recommendation System... heh...**  
<https://www.flickr.com/photos/torley/4551424756> (retrieved: Oct 2016).
- [13] **Trends Of Grandperspective Images.**  
40cg.com (retrieved: Oct 2016).
- [14] **"Underground"-branded Tube map from 1908.**  
[https://en.wikipedia.org/wiki/Timeline\\_of\\_the\\_London\\_Underground#/media/File:Tube\\_map\\_1908-2.jpg](https://en.wikipedia.org/wiki/Timeline_of_the_London_Underground#/media/File:Tube_map_1908-2.jpg) (retrieved: Oct 2016).
- [15] **Which color car was most popular in 2010?**  
<http://carinsurance.arrivealive.co.za/which-color-car-was-most-popular-in-2010.php> (retrieved: Oct 2016).
- [16] **Wikipedia: High-availability cluster.**  
[https://en.wikipedia.org/wiki/High-availability\\_cluster](https://en.wikipedia.org/wiki/High-availability_cluster) (retrieved: Oct 2016).
- [17] **Wikipedia: Verschiedene Punktwolken zusammen mit dem für sie jeweils berechenbaren Pearson'schen Korrelationskoeffizienten.**  
[https://de.wikipedia.org/wiki/Datei:Correlation\\_examples.png](https://de.wikipedia.org/wiki/Datei:Correlation_examples.png) (retrieved: Oct 2016).