



Statistik und Wahrscheinlichkeitsrechnung  
– Wintersemester 2019/20 –

# Kapitel 01: Deskriptive Statistik

Prof. Dr. Adrian Ulges

Angewandte Informatik (B.Sc.) /  
Informatik - Technische Systeme (B.Sc.) /  
Wirtschaftsinformatik (B.Sc.)

Fachbereich DCSM  
Hochschule RheinMain

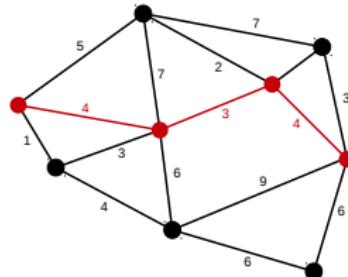


Im bisherigen Verlauf Ihres Studiums haben Sie sich primär mit **universalgültigen Fakten** befasst:

- ▶ “Jede konvergente Folge ist beschränkt.”
- ▶ “Eine Kante im Graph habe Gewicht  $g$ ” (Routenplaner)
- ▶ “ $(A \rightarrow B) \rightarrow (\neg B \rightarrow \neg A)$ ”

*“Wissen = sowohl subjektiv als auch objektiv  
zureichendes Fürwahrhalten”*

(Kant)



# Statistik: Motivation

In der Realität ist unser Wissen aber 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", ...



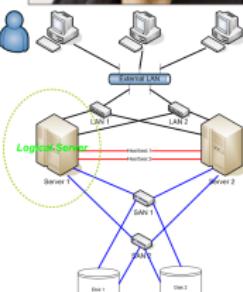
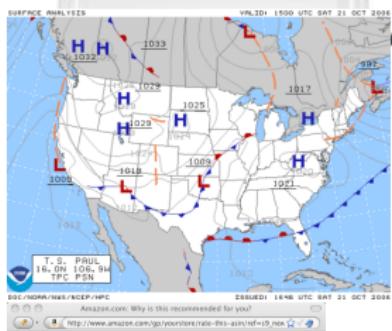
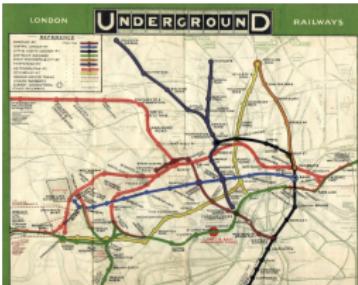
Was sind die Gründe für diese Unsicherheit?

- ▶ **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!**



# Anwendungsfelder

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



The screenshot shows a user profile with the following details:

- Recommended for You:**
  - I Just Want You to Know: Letters** by T.S. Paul
  - Our Price: \$9.99
  - Used & new from \$9.99
  - [See all buying options](#)
- Because you purchased...**
  - I Am Ozzy** (Kindle Edition)
  - This one is off
  - Don't see the recommendations

# 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.”*

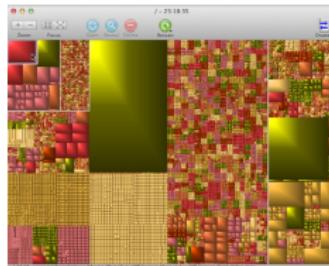
(Gabler Wirtschaftslexikon)



*“... 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*)

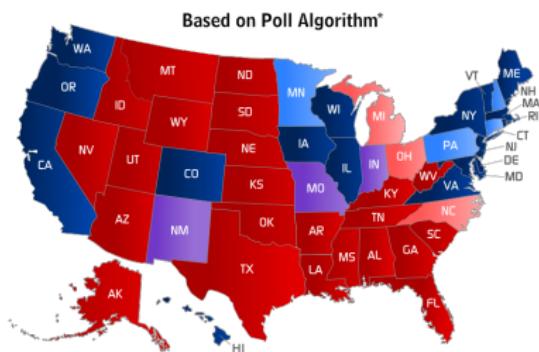


# Teilgebiet 2: Induktive Statistik

Bilder: [8] [11]



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



0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5  
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6  
7 7 7 7 7 7 7 7 7 7 7 7 7 7 1  
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8  
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9



Classifier

# Outline

1. Grundbegriffe
2. Lageparameter
3. Streuungsparameter
4. Zusammenhangsparameter
5. 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.

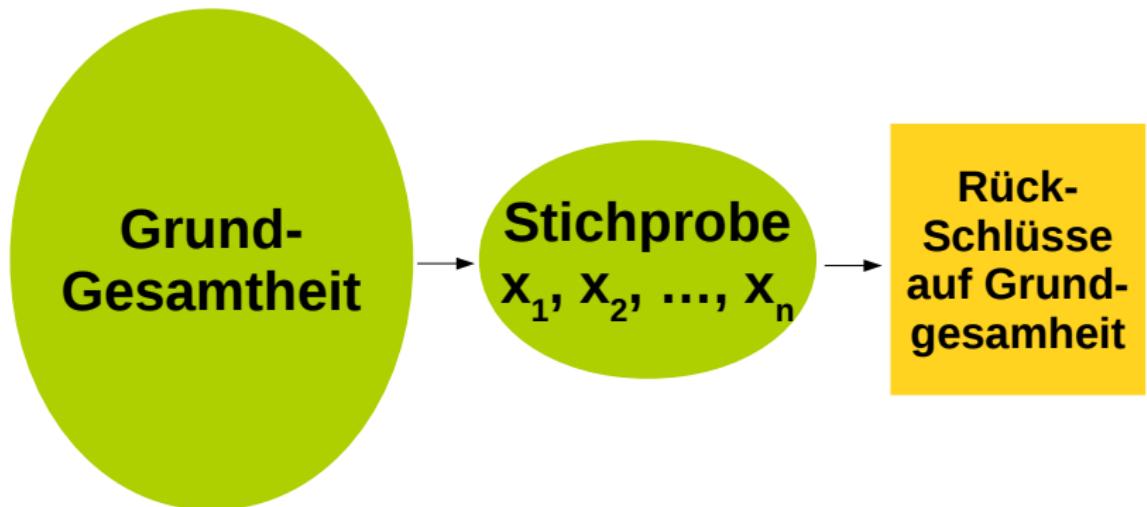
- ▶ *Beispiel: Eine Anzahl zufällig ausgewählter Personen*
- 

## Merkmale

Die Datenpunkte (engl. *Samples*) der Stichprobe werden durch dieselben **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:

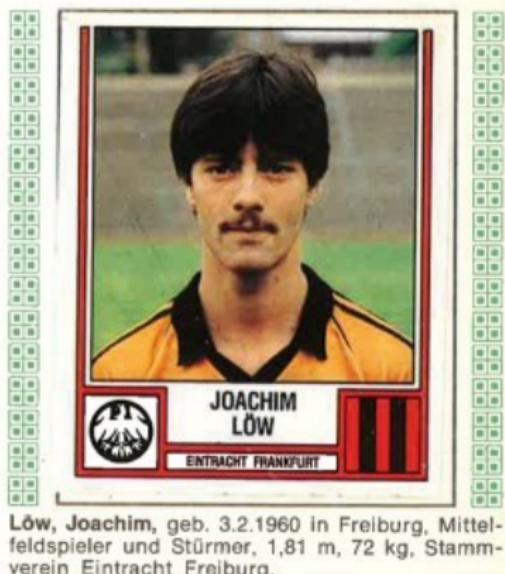
- ▶ **nominal**: Kategorien ohne natürliche Ordnung  
z.B. *Farbtöne (rot, grün, gelb...)*, *Geschlechter*
- ▶ **ordinalskaliert**: Kategorien mit Ordnung  
z.B. *Dienstgrade (Bachelor < Master < PhD )*
- ▶ **intervallskaliert**: Zahlen mit Abstandsmaß  
z.B. *Temperaturen, Kalendertage*
- ▶ **verhältnisskaliert**: 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”**.

# 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 beobachten wir nicht nur ein einziges Merkmal

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

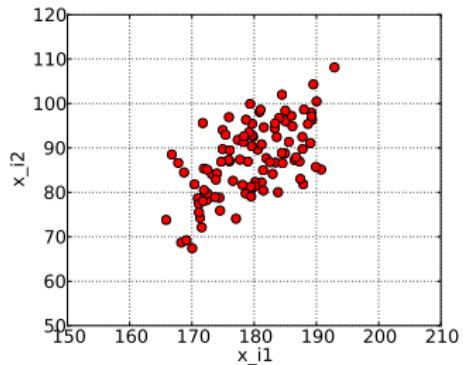
1

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

# Grundbegriffe: “multivariat”

```
# Größe/Gewicht von  
# Testpersonen (cm/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 sind 2D-Arrays / Matrizen
- $n$  Zeilen (Datenpunkte),  $m$  Spalten (Merkmale)
- Für quantitative Merkmale: Datenpunkt = Punkt in  $\mathbb{R}^m$
- Beispielvisualisierung: Scatterplot



# Absolute Häufigkeit und Relative Häufigkeit



## Definition (Absolute und relative Häufigkeit)

In einer univariaten 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  $a_j$ :

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

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

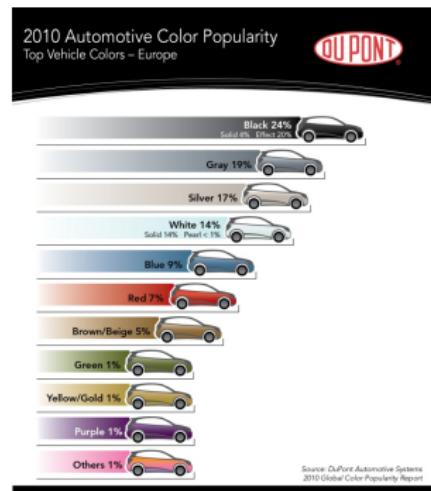
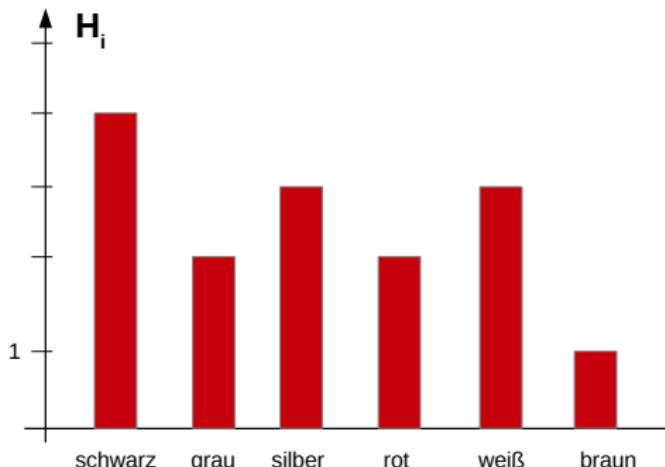
- ▶ Werte in der Stichprobe:  $(a_1, \dots, a_6) = (\text{schwarz}, \text{grau}, \text{silber}, \text{rot}, \text{weiß}, \text{braun})$
- ▶ Absolute Häufigkeiten:  
 $(H_1, \dots, H_6) = (4, 2, 3, 2, 3, 1)$
- ▶ Relative Häufigkeiten ( $n = 15$ ):  
 $(h_1, \dots, h_6) = \left(\frac{4}{15}, \frac{2}{15}, \frac{3}{15}, \frac{2}{15}, \frac{3}{15}, \frac{1}{15}\right)$

# 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 univariate 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,

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

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

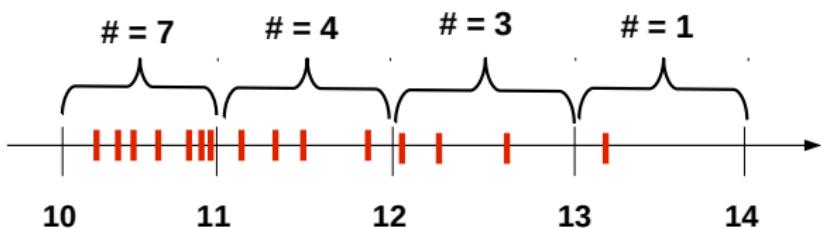
- $H_1^* = \# \text{ Werte zwischen } 10 \text{ und } 11 = 7$

- $H_2^* = \# \text{ Werte zwischen } 11 \text{ und } 12 = 4$

- $H_3^* = \# \text{ Werte zwischen } 12 \text{ und } 13 = 3$

- $H_4^* = \# \text{ Werte zwischen } 13 \text{ und } 14 = 1$

- Also:  $(H_1^*, \dots, H_4^*) = (7, 4, 3, 1)$



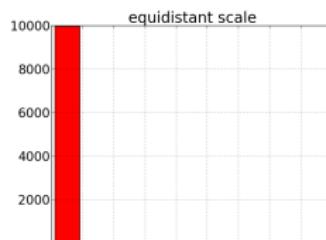
# Histogramme: Zerlegungen

- ▶ Oft wählen wir für unsere Histogramme **äquidistante Zerlegungen** (*mit gleich breiten Abschnitten*).
- ▶ Sind die Samples sehr ungleich verteilt, wählen für Bereiche mit wenig Samples gerne **breitere Bins**.

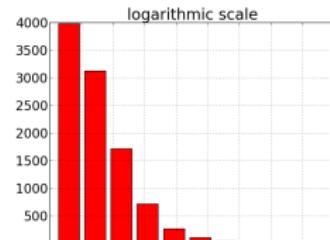
## Gegenbeispiel: Häufigkeiten englischer Wörter

```
# frequency
# (n=10.000)
the 56271872
of 33950064
and 29944184
to 25956096
in 17420636
i 11764797
that 11073318
was 10078245
his 8799755
he 8397205
...
purified 3924
sequel 3924
calves 3923
```

- ▶ Samples = Vorkommens-Häufigkeiten Englischer Terme in Beispiel-Texten
- ▶ Die Daten sind **schief** (engl. *long-tail*)
  - ▶ die meisten Wörter sind selten
  - ▶ sehr wenige Wörter sind sehr häufig



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



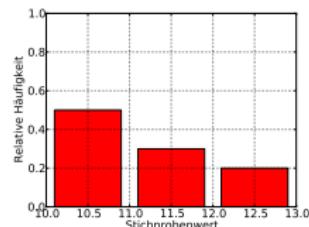
$$Z_{10} = (0, 10000, 20000, 50000, \dots, 30\text{mio.})$$

# Normierung von Histogrammen

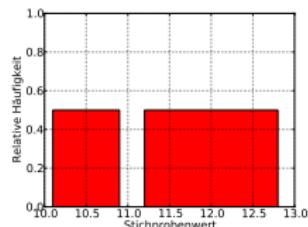
- Bei der grafischen Darstellung von Histogrammen ist zu beachten, dass die dargestellte **Gesamtfläche** sich nicht mit der Bin-Breite ändern sollte.
- Deshalb **normieren** wir Histogramme: Es sei  $\Delta_k := y_k - y_{k-1}$  und  $H_k^*$  (bzw.  $h_k^*$ ) die zugehörige Häufigkeit. Dann wählen wir 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 einen Bin um den Faktor 2, halbieren wir die Höhe des zugehörigen Rechtecks!

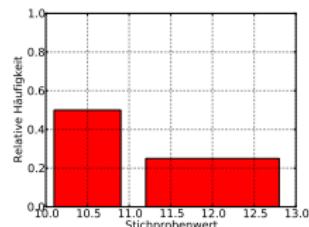
# n=10  
10.1  
10.3  
10.6  
10.7  
10.9  
11.2  
11.4  
11.8  
12.4  
12.7



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



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



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

# Do-Histogramme-yourself

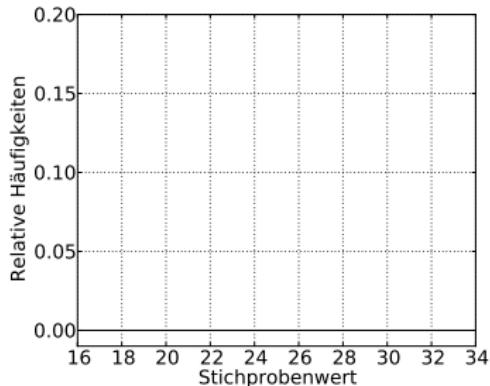
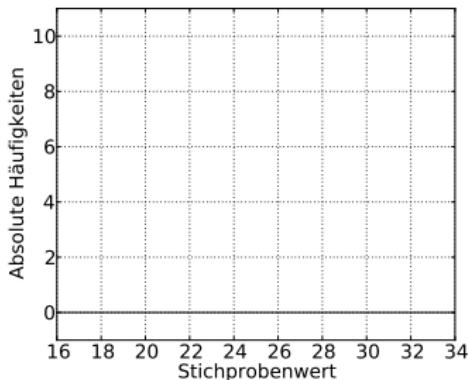


Gegeben ist die folgende (geordnete) Stichprobe:

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)



# Outline



1. Grundbegriffe
2. Lageparameter
3. Streuungsparameter
4. Zusammenhangsparameter
5. Lineare Regression



# Kennwerte von Stichproben: Motivation

Ein **Grundproblem**: Die Beschreibung der Daten mittels der Verteilung von Häufigkeiten ist oft zu detailliert / unhandlich.

- ▶ **Beispiel:** Verteilung der Produktpreise zweier Supermärkte – Welcher ist der günstigere?
- ▶ **Ansatz:** Charakterisiere die Stichprobe durch **Kennwerte**
  - ▶ Mittelwert
  - ▶ Median
  - ▶ Modalwert
  - ▶ 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 ((Arithmetischer) 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.

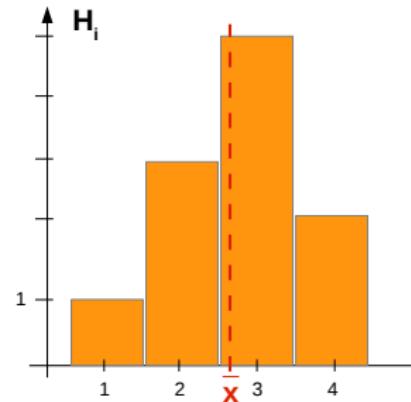
Beispiel:  $x_1, \dots, x_{11} = 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4$

$$\begin{aligned}\bar{x} &= \frac{1}{11} \cdot (1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 4) \\ &= \frac{1}{11} \cdot (1 \cdot 1 + 3 \cdot 2 + 5 \cdot 3 + 2 \cdot 4) \quad // \text{ mit absoluten Häufigkeiten} \\ &\approx 2.7\end{aligned}$$

# Mittelwert: Anschauliche Interpretation

## Univariate Stichproben

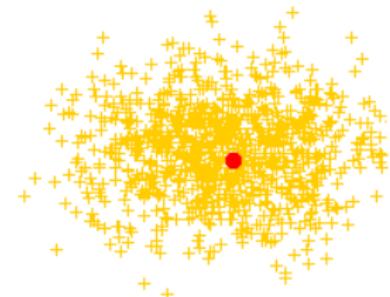
- ▶ **Beispiel:**  $x_1, \dots, x_{11} = 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4$
- ▶  $\bar{x} \approx 2.7$
- ▶ Der Mittelwert entspricht dem **Schwerpunkt** des Säulendiagramms



## Multivariate Stichproben

- ▶ **Beispiel:**  $x_1, \dots, x_n \in \mathbb{R}^2$
- ▶ Der Mittelwert entspricht dem **Schwerpunkt** der Punktewolke:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix}$$



# 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

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

$$\tilde{x} = x_6 = 3$$

- $x_1, \dots, x_{10} = 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4$

$$\tilde{x} = \frac{1}{2}(x_5 + x_6) = 2.5$$

# Kennwerte: Median (cont'd)

## Anmerkungen

- ▶ Die Gaussklammern  $[x]$  bezeichnen das **Aufrunden** von  $x$ :  
Beispiel:  $[2.1] = 3$ .
- ▶ 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, 4, 4, 4, 5, 5, 573, 1224

## Lageparameter: Quantile



..  
U  
a

- Der Median entspricht der Grenze, unter der genau 50% der Samples liegen.
- Dies können wir auf beliebige Prozentsätze  $\alpha$  erweitern! Wir erhalten die sogenannten  $\alpha$ -Quantile der Stichprobe.

Es sei  $\alpha \in ]0,1[$  und  $x_1, \dots, x_n \in \mathbb{R}$   
eine aufsteigend sortierte Stichprobe

Dann nennen wir:

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

das  $\alpha$ -Quantil der Stichprobe.

## Lageparameter: Quantile



# Quantile: Beispiel

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

$\downarrow$

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

$$\alpha = 0,25 : \alpha \cdot n = 0,25 \cdot 14 = 3,5$$

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

$$\alpha = 0,5 : \alpha \cdot n = 0,5 \cdot 14 = 7$$

$$\tilde{x}_\alpha = \frac{1}{2} \cdot (x_7 + x_8) = 20$$

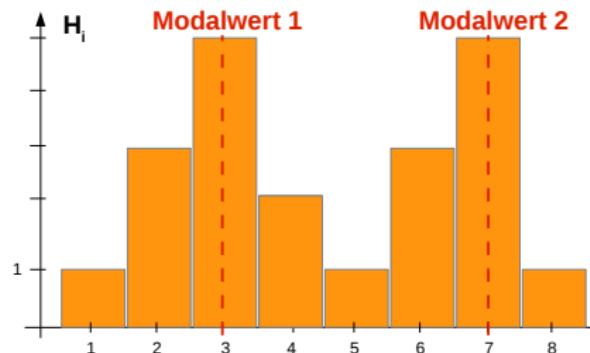
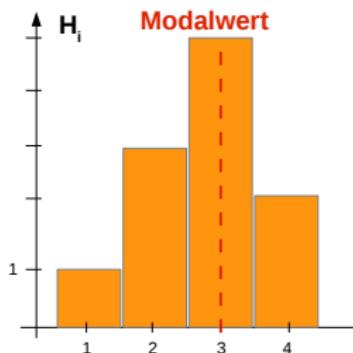
# Lageparameter: Modalwert



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  $5 \times$  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. Grundbegriffe
2. Lageparameter
3. Streuungsparameter
4. Zusammenhangsparameter
5. Lineare Regression

# Mittelwert: Wiederholung

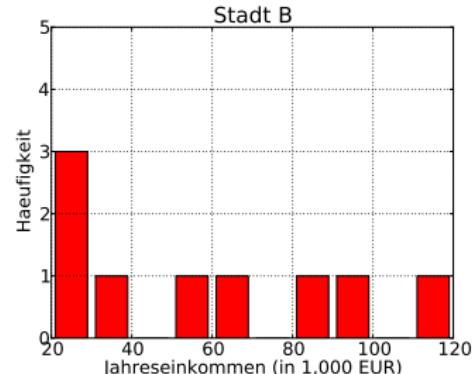
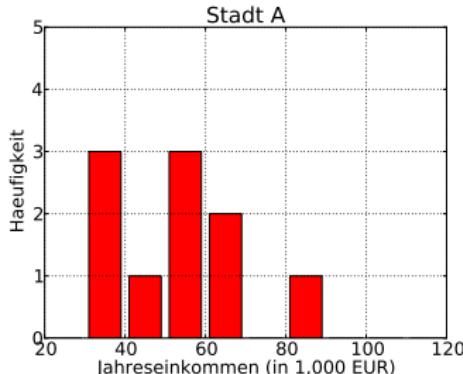
# Stadt A	# Stadt B
57.7	51.3
33.7	24.1
85.1	86.9
31.2	37.2
66.9	110.7
46.9	95.6
57.2	23.2
68.7	64.8
50.2	15.2
38.3	26.9

- ▶ 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

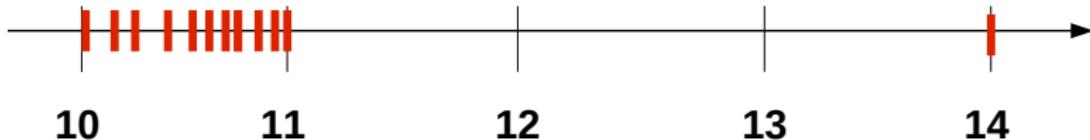
$$R := x_n - x_1$$

(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?





## 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.

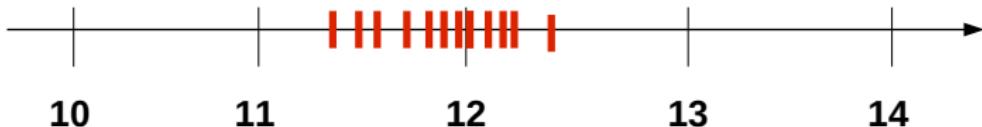
## Anmerkungen

- ▶ Umgangssprachlich: “Der durchschnittliche 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>*.

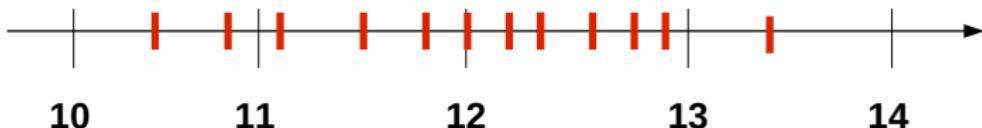
# Varianz und Standardabweichung: Illustration



Geringere Varianz



Höhere Varianz



# Varianz und Standardabweichung: Beispiel



# 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:

$$\begin{aligned}s^2 &= \frac{1}{n} \sum_i (x_i - \bar{x})^2 \\ &= \frac{1}{14} \left( (10 - 22)^2 + (13 - 22)^2 + \dots \right. \\ &\quad \left. \dots + (36 - 22)^2 + (41 - 22)^2 \right) \\ &\approx 76.29\end{aligned}$$

► 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**:

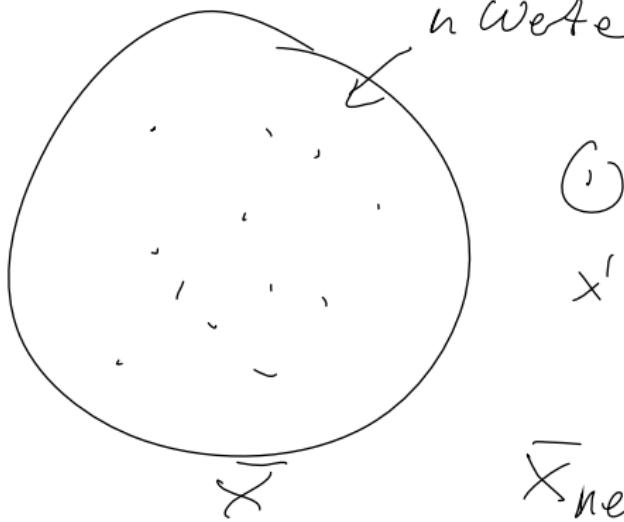
## Varianz: Verschiebungssatz (Beweis)



$$\begin{aligned}
 S^2 &= \frac{1}{n} \sum_i (x_i - \bar{x})^2 \quad // 2. Binomische Formel \\
 &= \frac{1}{n} \sum_i (x_i^2 - 2 \cdot x_i \bar{x} + \bar{x}^2) \\
 &= \frac{1}{n} \sum_i x_i^2 - 2 \cdot \bar{x} \underbrace{\frac{1}{n} \sum_i x_i}_{\text{}} + \frac{1}{n} \left( \underbrace{\sum_{i=1}^n x_i}_{n \bar{x}} - n \bar{x} \right)^2 \\
 &= " - 2 \cdot \bar{x} \cdot \bar{x} + \cancel{\frac{1}{n} \cdot n \cdot \bar{x}^2} \\
 &= \frac{1}{n} \sum_i x_i^2 - 2 \cdot \bar{x}^2 + \bar{x}^2 \\
 &= \frac{1}{n} \sum_i x_i^2 - \underbrace{\bar{x}^2}_{\text{Verschiebungssatz}}
 \end{aligned}$$

Mittelwert über die quadratierte Stichprobenwerte.

## Varianz: Verschiebungssatz (Beweis)



$$\bar{X}_{\text{neu}} = \frac{n}{n+1} \cdot \bar{X} + \frac{1}{n+1} x'$$

# Varianz: Verschiebungssatz (Diskussion)



# Varianz: Do-it-yourself



```
# Größe/Gewicht von  
# Testpersonen (cm/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
```

- ▶ Kennwerte dieses Datensatzes:

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

- ▶ Wie verändern sich die Kennwerte, wenn wir die Größe in  $m$  und das Gewicht in  $g$  angeben?

$$\begin{array}{c} \text{Größe } \bar{x}' = 1748 \quad s'^2 = 714 \quad s' = 8,5 \cdot 1000 \\ \text{Gewicht } 867 \cdot 100 \quad 714 \cdot 1000^2 \quad 8,5 \cdot 1000 \\ : 100 \end{array}$$

# 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?

## Satz (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:

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

und berechnen Mittelwert und Varianz dieser linear transformierten Stichprobe,  $\bar{x}'$  und  $s'^2$ . Es gilt:

$$1) \bar{x}' = \alpha \cdot \bar{x} + \beta$$

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



## Mittelwert und Varianz: Transformation



(Teil-)Beweis zu (2)

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

## Mittelwert und Varianz: Transformation



## Mittelwert und Varianz: Transformation



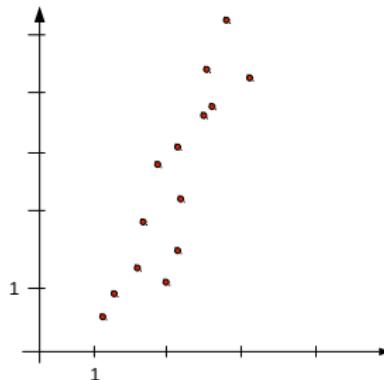
# Outline



1. Grundbegriffe
2. Lageparameter
3. Streuungsparameter
4. Zusammenhangsparameter
5. Lineare Regression

# Varianz multivariater Stichproben?

Gegeben sei eine **bivariate** Stichprobe  $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$



Was können wir über die Streuung dieser Stichprobe aussagen?

## Zusammenhangsparameter

Mit Zusammenhangsparametern stellen wir fest, ob eine *Abhangigkeit* zwischen zwei (oder mehr) 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

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

die (empirische) **Kovarianz** der Stichprobe.

# Kovarianz: Do-it-yourself

#  $x_i$   $y_i$ 

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:

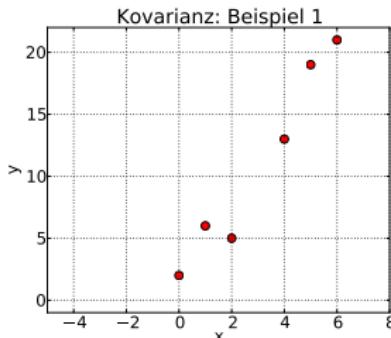
$$S_{xy} = \frac{1}{6} \cdot [(0-3) \cdot (2-11) + (1-3) \cdot (6-11) + (2-3) \cdot (5-11) + \dots]$$

$$\approx 15,17 > 0$$



# 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

- ▶  $x_1, \dots, x_n$  = Körpergröße,  $y_1, \dots, y_n$  = Gewicht
- ▶  $x_1, \dots, x_n$  = tägliche Regenmenge,  
 $y_1, \dots, y_n$  = tägliche 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

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

$$\begin{aligned} & \approx \frac{1}{6} \left( (1 - 5)(6 - 5.2) + (2 - 5)(2 - 5.2) + \dots \right. \\ & \quad \left. \dots + (10 - 5)(5 - 5.2) \right) \end{aligned}$$

$$\approx \frac{1}{6} \left( -3.3 + 9.5 - 3.8 - 1.2 - 0.3 - 0.8 \right)$$

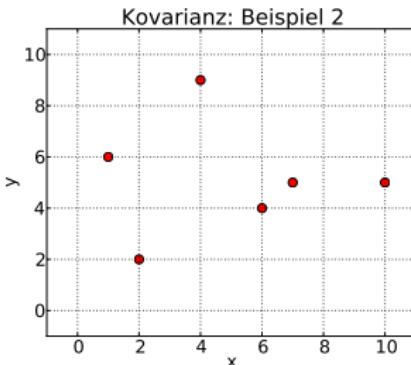
$$= 0.08$$

$$\approx 0$$

## Beispiel 2: Kovarianz $\approx 0$



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



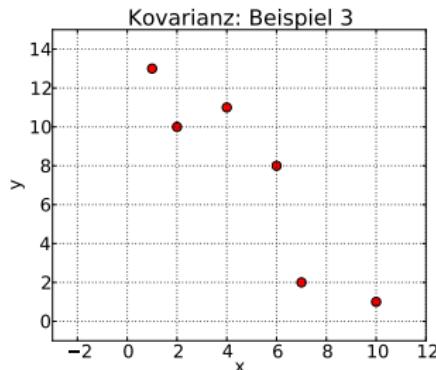
- Mit wachsendem  $x$ -Wert ändert sich der  $y$ -Wert **nicht** (bzw. nicht *stark*).
- Dies betrifft die **lineare Änderung** (*hierzu gleich mehr*).

### 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:  $x_1, \dots, x_n =$  investierter Zeitaufwand,  $y_1, \dots, y_n =$  Note
- ▶ Newton-Verfahren:  $x_1, \dots, x_n =$  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

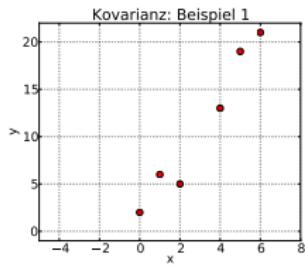
$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

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

# Korrelation: Beispiele (siehe eben)

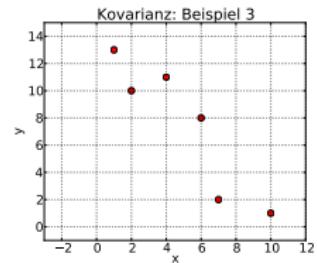
#	$x_i$	$y_i$
0	2	
1	6	
2	5	
4	13	
5	19	
6	21	

- ▶ 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:  $r_{xy} \approx \frac{-12.7}{3.1 \cdot 4.5} \approx \mathbf{-0.91}$



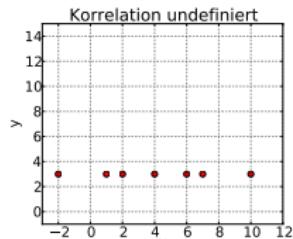
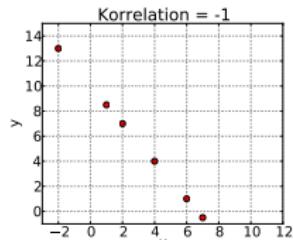
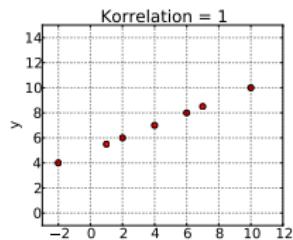
# Korrelation: Eigenschaften

- ▶ 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 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:

- ▶  $a > 0 \Rightarrow r_{xy} = 1$
- ▶  $a < 0 \Rightarrow r_{xy} = -1$
- ▶  $a = 0 \Rightarrow r_{xy} = \uparrow (\text{undefined})$



## Korrelation: Eigenschaften (cont'd)

- ▶ Falls  $r_{xy} = 0$ , sagen wir: "Es ist keine lineare Abhangigkeit nachweisbar".
- ▶ 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 Sortieralgorithmus  
 $y =$  Größe der Eingabe
  
- ▶  $x =$  Größe des Hauptspeichers  
 $y =$  Anzahl der Page Faults
  
- ▶  $x =$  Anzahl der Haare eines Mannes  
 $y =$  Einkommen des Mannes



# Korrelation $\not\leftrightarrow$ Kausalität



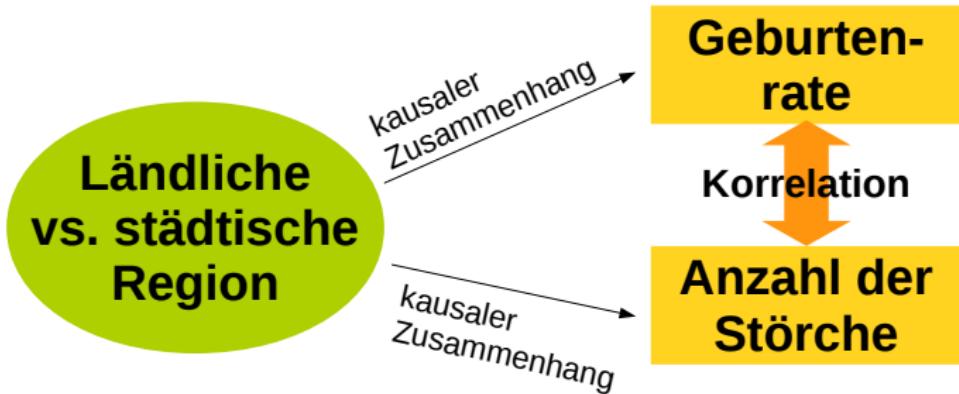
**Achtung:** Wenn wir eine hohe Korrelation messen, deutet dies noch nicht auf einen **kausalen Zusammenhang** hin. Das heisst: Weder muss Variable 1 die **Ursache** für das Verhalten von Variable 2 sein, noch Variable 2 die Ursache für das Verhalten von Variable 1.

## Beispiele

- ▶ Die Anzahl der Haare ist negativ korreliert mit dem Einkommen
- ▶ Die Geburtenquote von Gemeinden ist korreliert mit der Anzahl der ansässigen Storchenpaare

Warum?

# Korrelation $\not\leftrightarrow$ Kausalität



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

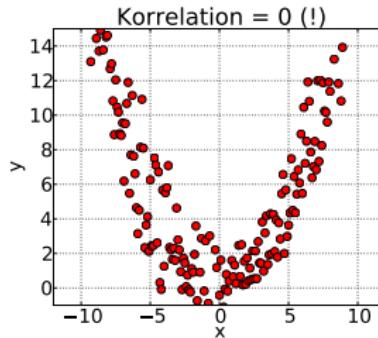
# Korrelation=0 $\Leftrightarrow$ Unabhängigkeit



Wir messen eine Korrelation von 0 zwischen zwei Variablen.  
Bedeutet dies, dass **keine Abhängigkeit** besteht? **Nein!**

## Beispiel

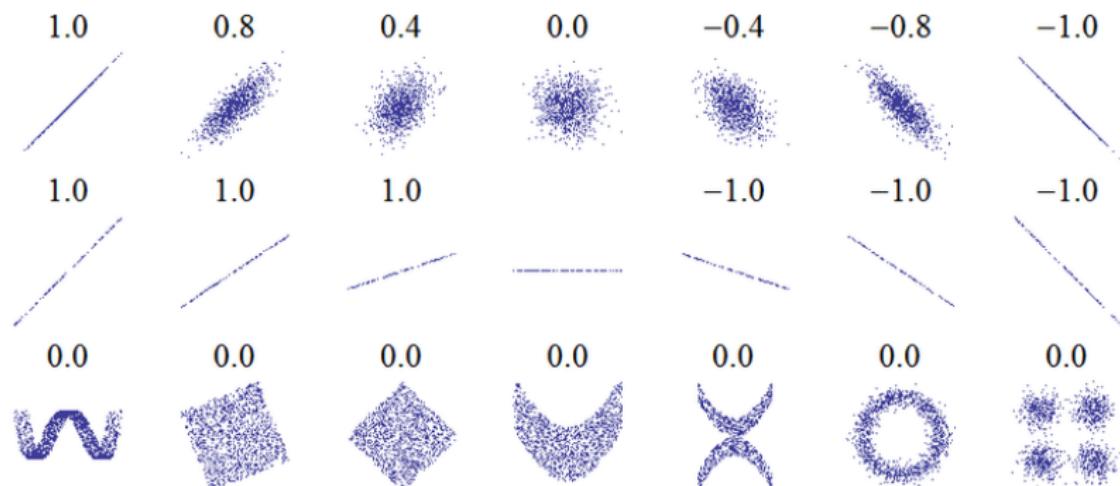
- ▶ Im Beispiel unten liegt keine Korrelation vor, also keine *lineare Abhängigkeit*
- ▶ Dennoch besteht hier eine Abhängigkeit zwischen  $x$  und  $y$  (je größer  $|x|$ , desto größer  $y$ )!



# Korrelation=0 $\Leftrightarrow$ Unabhängigkeit Bild: [17]



- ▶ Einige Beispiele bivariater Stichproben mit zugehöriger Korrelation
- ▶ In allen Beispielen der *unteren Reihe* liegt keine Korrelation vor, die Variablen  $x$  und  $y$  sind aber dennoch abhängig!



# Kovarianz multivariater Stichproben



Wir ignorieren nun Rangkorrelationen und rechnen wieder mit den **Ausgangswerten**. Hier widmen wir uns einer weiteren Fragestellung:  
Was wenn Stichproben **mehr als zwei** Merkmale aufweisen?

Hierfür **erweitern** wir die **Kovarianz** auf multivariate Stichproben:

## Definition (Kovarianz multivariater Stichproben)

Gegeben ist eine Stichprobe  $x_1, \dots, x_n \in \mathbb{R}^d$  mit **d Merkmalen** (wir notieren  $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ ). Der Mittelwert der Stichprobe sei  $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_d)$ . Dann nennen wir

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

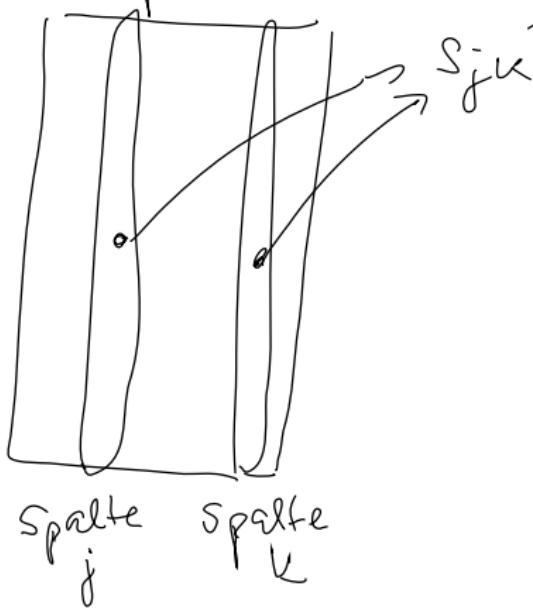
die (empirische) **Kovarianz** der Stichprobe zwischen den **Merkmälern/Spalten j und k**.

# Die Kovarianzmatrix



Wir können sämtliche Kovarianzen in einer Matrix zusammenfassen, der sogenannten **Kovarianzmatrix  $\Sigma$** .

Stichprobe



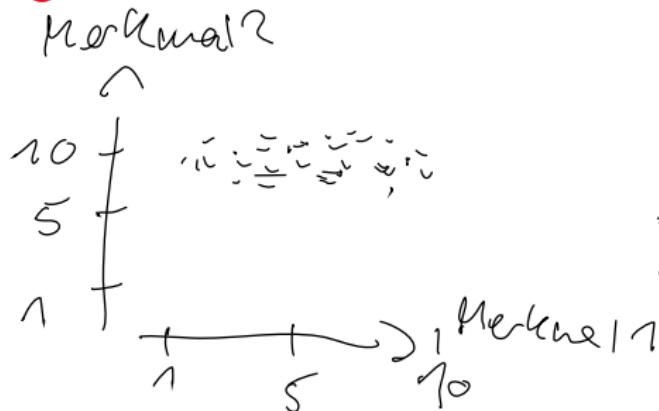
$$\Sigma = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1d} \\ S_{21} & S_{22} & \dots & S_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ S_{d1} & \dots & S_{d2} & S_{dd} \end{pmatrix}$$

$$\Sigma \in \mathbb{R}^{d \times d}$$

## Eigenschaften der Kovarianzmatrix \*

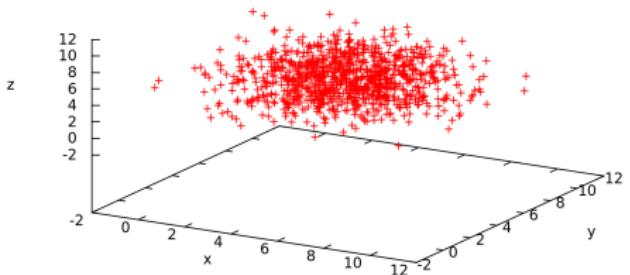
- $\Sigma$  ist  $d \times d$  (quadratisch)
- $\Sigma$  ist symmetrisch.
- Auf der Diagonale von  $\Sigma$  stehen die Varianzen der einzelnen Merkmale:  $\overbrace{\sum_{jj}} = s_j^2$

# Eigenschaften der Kovarianzmatrix



$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

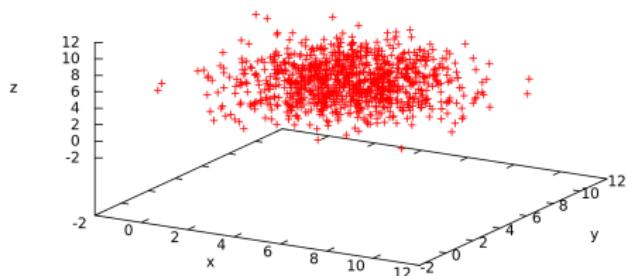
# Multivariate Stichproben: Do-it-yourself



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!

# Multivariate Stichproben: Do-it-yourself



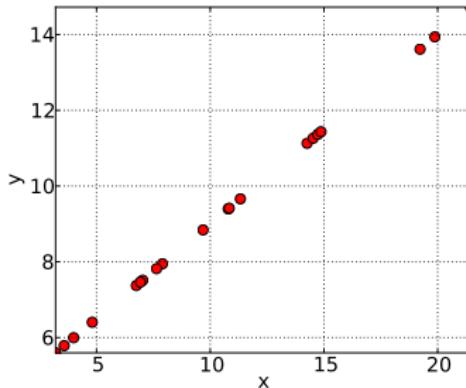
Wie lautet die **Kovarianzmatrix**?

# Outline



1. Grundbegriffe
2. Lageparameter
3. Streuungsparameter
4. Zusammenhangsparameter
5. Lineare Regression

# Regressionsprobleme

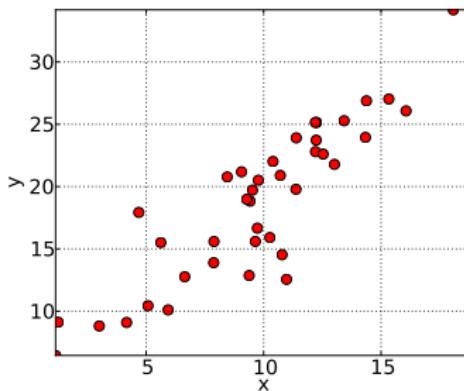


Unser Ziel im Folgenden ist es, eine **Funktion** (z.B. eine **Gerade**) zu ermitteln, die möglichst gut auf die Stichprobe  $x_1, \dots, x_n$  passt.

## Anwendungsfälle

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

# Regressionsprobleme



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

## “Least Squares”: Ansatz

- ▶ **Ansatz:** Wir erklären die Stichprobe  $(x_1, y_1), \dots, (x_n, y_n)$  durch eine **Funktion** (bzw. ein **Modell**)  $\mathcal{M}_\theta : \mathbb{R} \rightarrow \mathbb{R}$ .
- ▶ Die Funktion ordnet einem x-Wert einen y-Wert zu, und besitzt **Parameter**  $\theta$ :

$$y_i = \mathcal{M}_\theta(x_i) \quad \text{für } i = 1, \dots, n$$

- ▶ In unserem Fall ist das Modell  $\mathcal{M}_\theta$  eine **Gerade**. Die Parameter sind *Steigung*  $a$  und *Achsenabschnitt*  $b$ , d.h.  $\theta = (a, b)$  und

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

- ▶ **Anmerkung:** Least Squares kann auch mit anderen Modellen (Parabeln, Polynomen, ...) angewandt werden – *siehe Übung*.
- ▶ Unser Ziel ist es, die “besten” **Parameter**  $\theta$  zu ermitteln, d.h.  $a$  und  $b$  auf die Stichprobe zu **fitten**.

## “Least Squares”: Ansatz

- Wir modellieren die Variation in den Messwerten mittels Fehlervariablen. Diese nennen wir  $\epsilon_1, \dots, \epsilon_n$ . Es gilt:

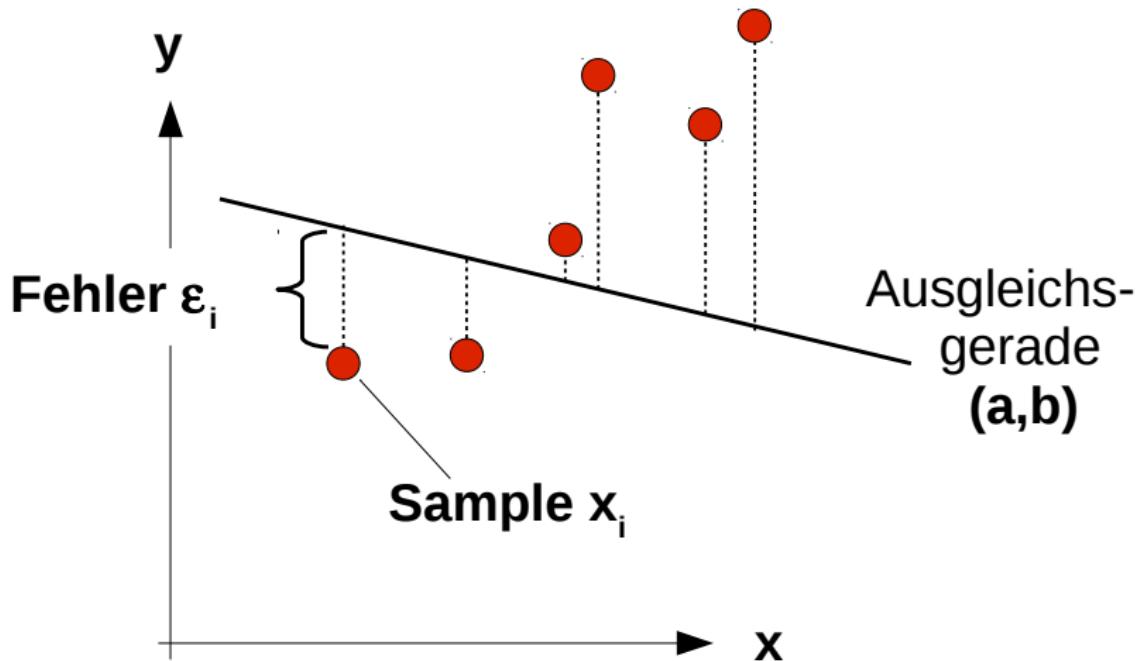
$$y_i = \underbrace{a \cdot x_i + b}_{\mathcal{M}_\theta(x_i)} + \epsilon_i \quad \text{für } i = 1, \dots, n$$

- 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

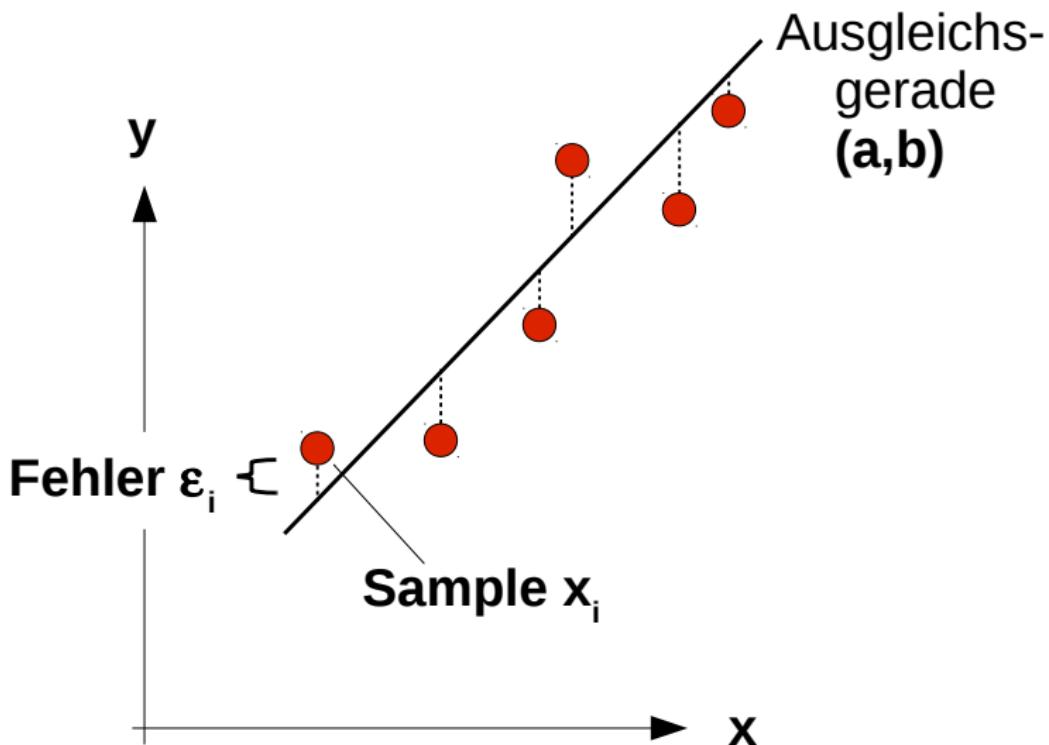
Die Fehler  $\epsilon_1, \dots, \epsilon_n$  sind nicht nahe null → schlechte Gerade.



## “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:

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

minf 

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

$$\begin{aligned} a^*, b^* &= \underset{a, b}{\operatorname{argmin}} E(a, b) \\ \text{Lösung, Ergebnis} &\equiv " \quad \sum_i \epsilon_i^2 \\ &= " \quad \sum_i ((ax_i + b) - y_i)^2 \end{aligned}$$

## “Least Squares”: Herleitung



$f_x$

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

$$E_a = \sum_i 2 \cdot ((\underline{ax}_i + \underline{b}) - y_i) \cdot x_i \stackrel{!}{=} 0 \quad (\text{I})$$

$$E_b = \sum_i 2 \cdot ((\underline{ax}_i + \underline{b}) - y_i) \cdot 1 \stackrel{!}{=} 0 \quad (\text{II})$$

$$\begin{aligned} (\text{I}) \quad \dots \times a + \dots \times b &= \dots \\ (\text{II}) \quad \dots \times a + \dots \times b &= \dots \end{aligned} \quad \left. \begin{array}{l} \\ \end{array} \right\} 2 \times 2 \text{ LGS}$$

$\Rightarrow$  Löse  $\Rightarrow a, b$



# “Least Squares”: Herleitung

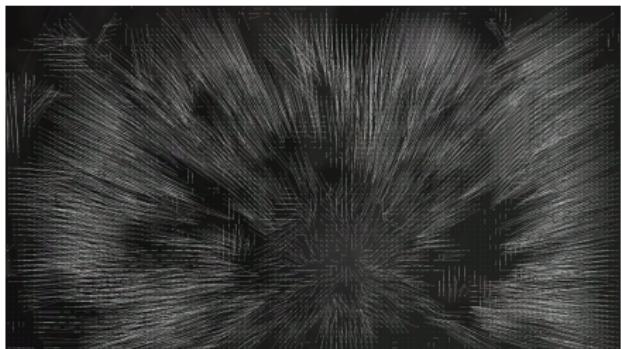
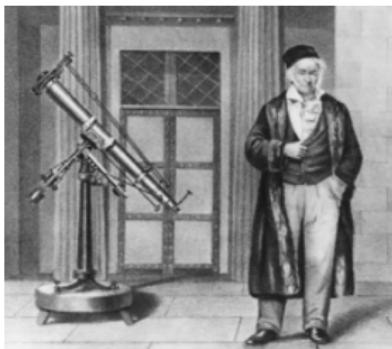
# “Least Squares”: Genereller Ansatz

Bilder: [7] [4]



Wir erhalten also **zwei lineare Gleichungen mit zwei Unbekannten**  $a, b$ . Es ergibt sich (nach einigen Umformungen<sup>2</sup>):

$$\begin{aligned}a^* &= s_{xy}/s_x^2 \\b^* &= \bar{y} - a^* \cdot \bar{x}\end{aligned}$$



---

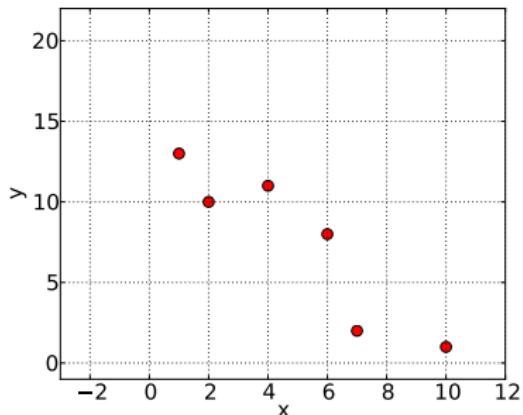
<sup>2</sup>siehe Teschl, Teschl: Mathematik für Informatiker, Band 2, 164ff

# 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

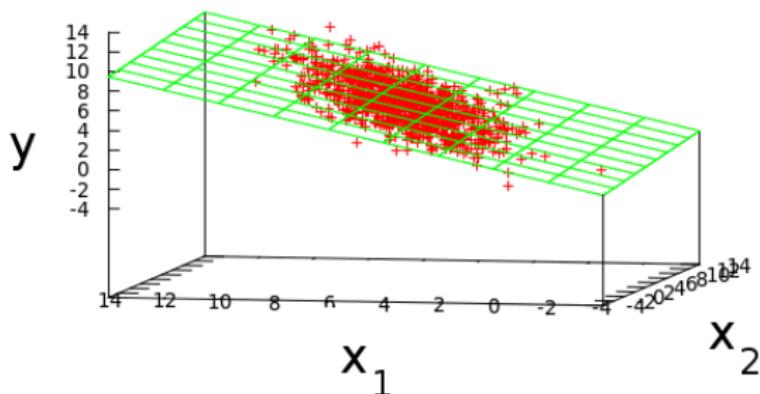


# Multiple Lineare Regression

- ▶ 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. der Preis),  
gegeben **mehrere** andere Variablen  $x_1, \dots, x_{m-1}$ .



# Multiple Lineare Regression: Herleitung

- Wir gehen (erneut) von einem linearen Zusammenhang aus. Gegeben einen **Eingabevektor**  $\mathbf{x} = (x_1, \dots, x_{m-1})$ , sagen wir den Wert  $y$  voraus:

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

- Zur Vereinfachung fügen wir dem Eingabevektor eine 1 hinzu ( $b$  verschwindet, es bleibt ein **Gewichtsvektor**  $\mathbf{w}$  der Länge  $m$ ):

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

- Einziger Parameter des Modells: Der **Gewichtsvektor**  $\mathbf{w}$ . Dieser definiert eine **(Hyper-)Ebene**.

# 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 Samples 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 Kriteriumswerte  $y_1, \dots, y_n \in \mathbb{R}$ . 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 (siehe oben) – eine **Fehlerfunktion**  $E$ :

$$E(\mathbf{w}) = \sum_i (\mathbf{w} \cdot \mathbf{x}_i - y_i)^2$$

- $E$  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:

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

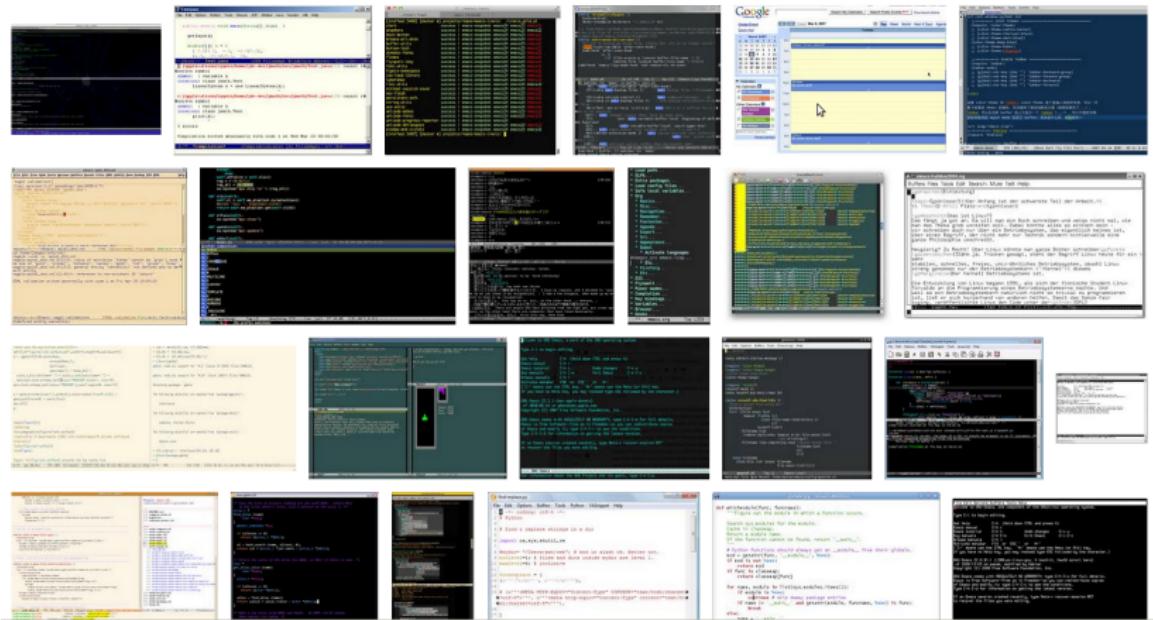
# Multiple Lineare Regression: Herleitung



# Multiple Lineare Regression: Herleitung



# Multiple Lineare Regression: Code-Beispiel



# Multiple Lineare Regression: Anmerkungen



- Die Regressionsgewichte geben an, wie groß der Einfluss eines einzigen Merkmals auf  $y$  ist!
- In der Praxis versuchen wir, *so viele “nützliche” Merkmale* wie möglich zu akquirieren.

## Beispiel

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

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

Feature $x_i$	weight $w_i$
<b>INDESCR_CNT_GEBRAUCHT</b>	<b>-2.475</b>
RELEASE AGE MONTHS	-5.107
...	

Wir trainieren ein Modell für die Vorhersage von **Auktionspreisen auf eBay**. Über die eBay-API erhalten wir Merkmale zur Beschreibung einer Auktion. Wir inspizieren deren **Gewichte**:

- Die Bewertung des Verkäufers ist wichtig für einen hohen Preis.
- Ist der Versand inclusive, hat dies einen leicht positiven Einfluss.
- Kommt das Wort “gebraucht” im Text vor, ist der Preis niedrig.



# References |

- [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).

# References II



- [10] Studie: "Kreditschwemme" kommt beim Mittelstand nicht an.  
<http://www.wirtschaft.com/studie-kreditschwemme-kommt-beim-mittelstand-nicht/> (retrieved: Oct 2016).
- [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).