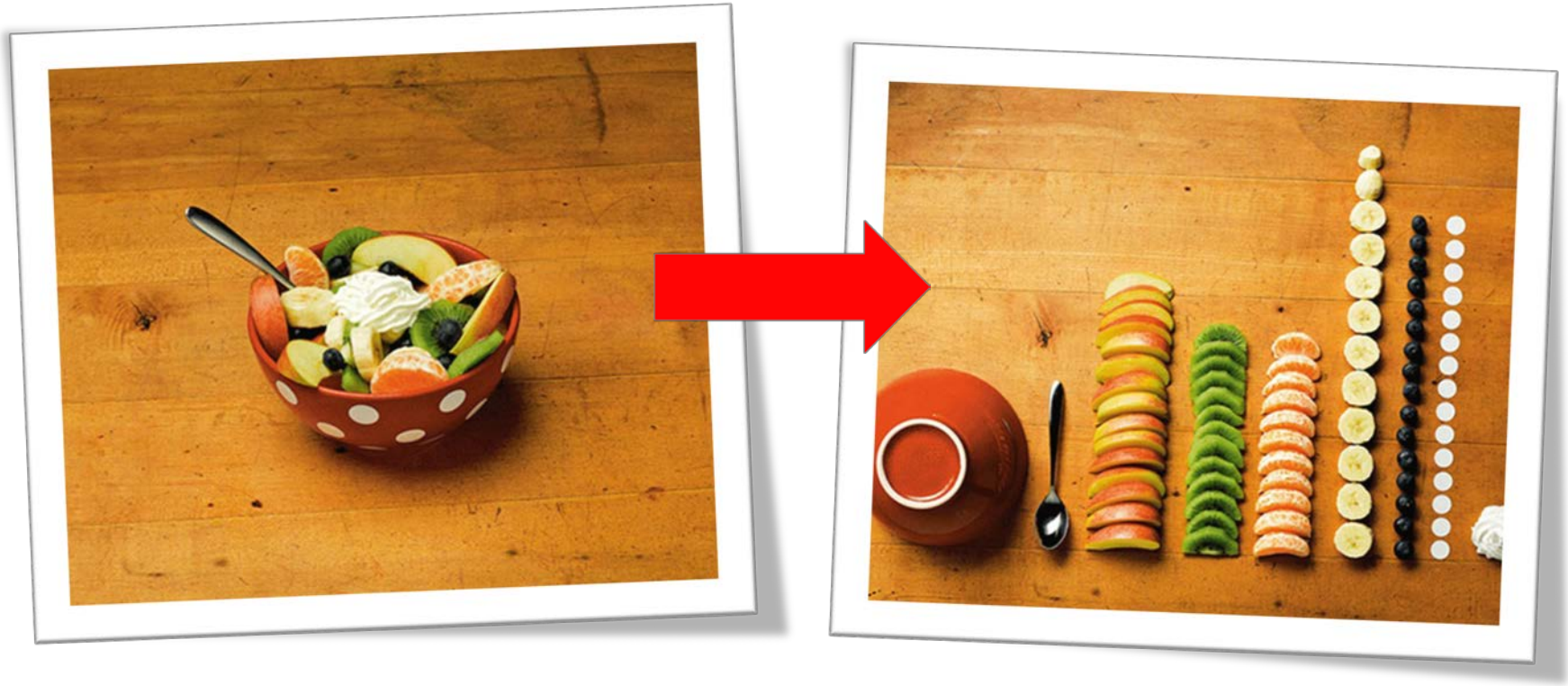
Deskriptive Statistik

(basierend auf Slides von Lukas Meier)

Deskriptive Statistik: Ziele

- Daten **zusammenfassen** durch **numerische Kennzahlen**.
- **Grafische Darstellung** der Daten.



Quelle: Ursus Wehrli, Kunst aufräumen

Modell vs. Daten

- Bis jetzt haben wir nur **Modelle** (**Verteilungen**) angeschaut.
- Jetzt betrachten wir (erstmals) **reale Daten**.
- Vorerst treffen wir aber **keine Annahmen**, dass diese von einer bestimmten Verteilung kommen! D.h. wir legen uns **nicht** auf ein Modell fest.
- Basierend auf den Daten können wir diverse **Kennzahlen** berechnen bzw. die Daten **grafisch darstellen**.

Kennzahlen: Überblick

Wir haben n beobachtete Datenpunkte $x_1, x_2, \ldots, x_n$ (z.B. das Verkehrsaufkommen an n verschiedenen Tagen oder Orten).

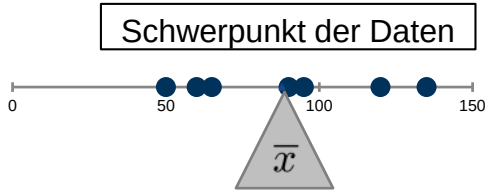
Wir unterscheiden zwischen

- **Lageparameter** («Wo liegen die Beobachtungen auf der Mess-Skala?»)
  - arithmetisches Mittel («Durchschnitt»)
  - empirischer Median
  - empirische Quantile
- **Streuungsparameter** («Wie streuen die Daten um ihre mittlere Lage?»)
  - empirische Varianz
  - empirische Standardabweichung
  - empirische Quartilsdifferenz

Arithmetisches Mittel und empirische Varianz

- Arithmetisches Mittel
  (emp. Pendant des Erwartungswerts μ)

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



- Empirische Varianz
  (emp. Pendant der Varianz σ^2)

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

- Empirische Standardabweichung
  (emp. Pendant der Standardabweichung σ)

$$s = \sqrt{s^2}$$

- Siehe Beispiel Wandtafel

Geordnete Stichprobe

- Wir ordnen unseren Datensatz in aufsteigender Reihenfolge und bezeichnen die **geordneten Daten** mit $x_{(i)}$, d.h.

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

- Die Position einer Beobachtung in der geordneten Stichprobe bezeichnet man als **Rang** (die kleinste Beobachtung hat also Rang 1, die grösste Beobachtung Rang n)

- Sind Beobachtungen gleich gross, so teilt man ihnen in der Regel ihren durchschnittlichen Rang zu

- Siehe Beispiel Wandtafel

Empirische Quantile

- Das empirische $(\alpha \times 100)\%$-Quantil $(0 < \alpha < 1)$ ist ein Wert q_α, so dass etwa $\alpha \times 100\%$ der Datenpunkte kleiner sind als q_α.

- Genau:
  - Falls $\alpha n \notin \mathbb{N}$, dann: $q_\alpha = x_{(\lceil \alpha n \rceil)}$, wobei $\lceil \alpha n \rceil$ die kleinste ganze Zahl grösser als αn ist
  - Falls $\alpha n \in \mathbb{N}$, dann: $q_\alpha = \frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)})$

- Es gibt (viele) Variationen für die genaue Definition. Für grosse n ist der Unterschied aber vernachlässigbar.

Empirische Quantile: Beispiel

i	1	2	3	4	5	6	7	8	9	10	11	12
$x_{(i)}$	79.97	79.98	80.04	80.08	80.12	80.23	80.35	80.38	80.39	80.44	80.45	80.48

Datensatz ist schon geordnet

- 90%-Quantil:

  $0.9 \cdot 12 = 10.8 \rightarrow 10.8 \notin \mathbb{N}$ und $\lceil 10.8 \rceil = 11 \rightarrow$ 90%-Quantil $= x_{(11)} = 80.45$

- 25%-Quantil:

  $0.25 \cdot 12 = 3 \rightarrow 3 \in \mathbb{N} \rightarrow$ 25%-Quantil $= \frac{1}{2}(x_{(3)} + x_{(4)}) = 80.06$

Ausgewählte Quantile

- **Median** (Zentralwert): 50%-Quantil $q_{0.5}$
- **Unteres Quartil**: 25%-Quantil $q_{0.25}$
- **Oberes Quartil**: 75%-Quantil $q_{0.75}$
- Die Differenz der Quartile $q_{0.75} - q_{0.25}$ bezeichnet man als **Quartilsdifferenz**, bzw. **Interquartile Range (IQR)**. Diese ist ein **Streuungsmass**.

- Bsp.

i	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$	6.2	6.3	7.0	7.1	9.6	9.9	10.8	11.8	12.5	14.4	16.2

Unteres Quartil

Median

Oberes Quartil

IQR = 12.5-7.0 = 5.5

$n = 11$:

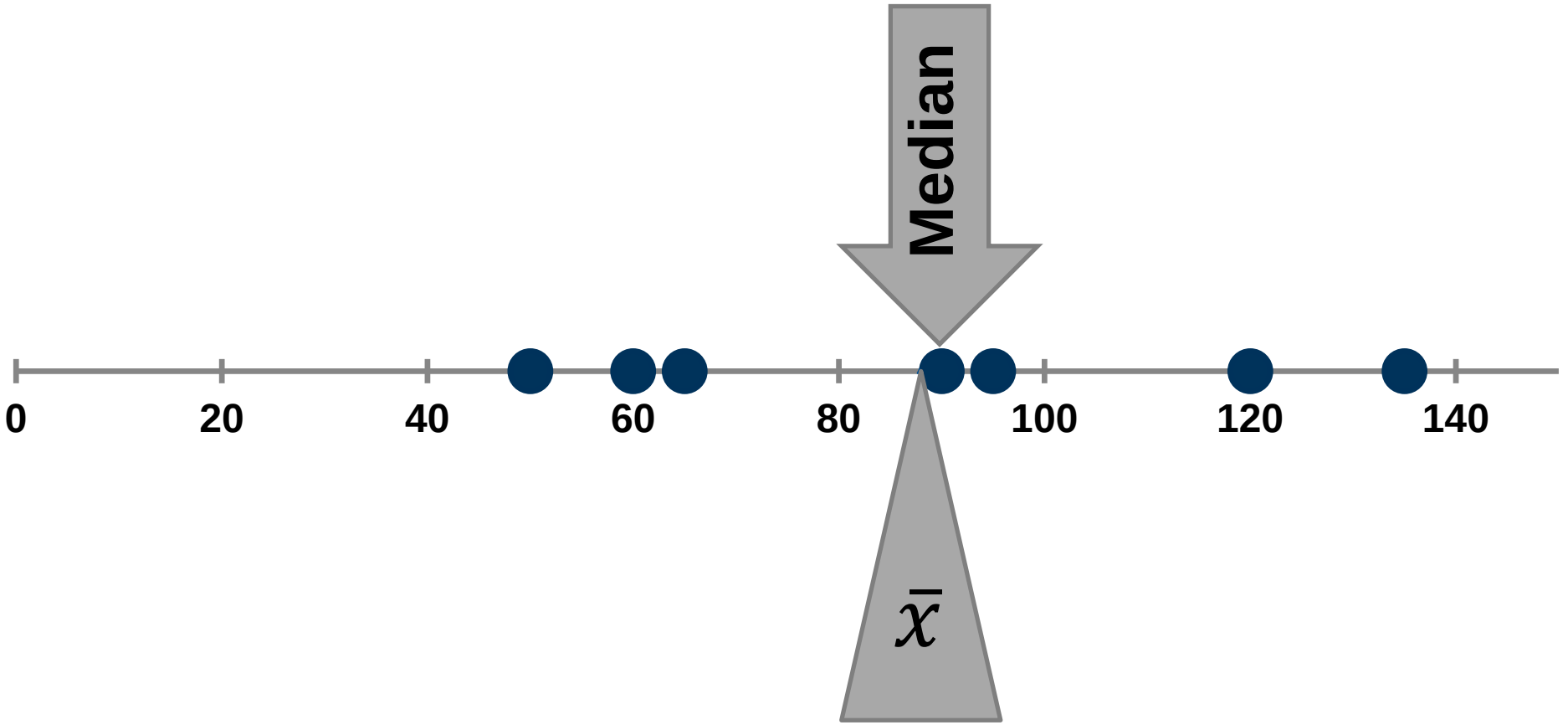
$0.25 \cdot 11 = 2.75 \rightarrow q_{0.25} = x_{(3)}$

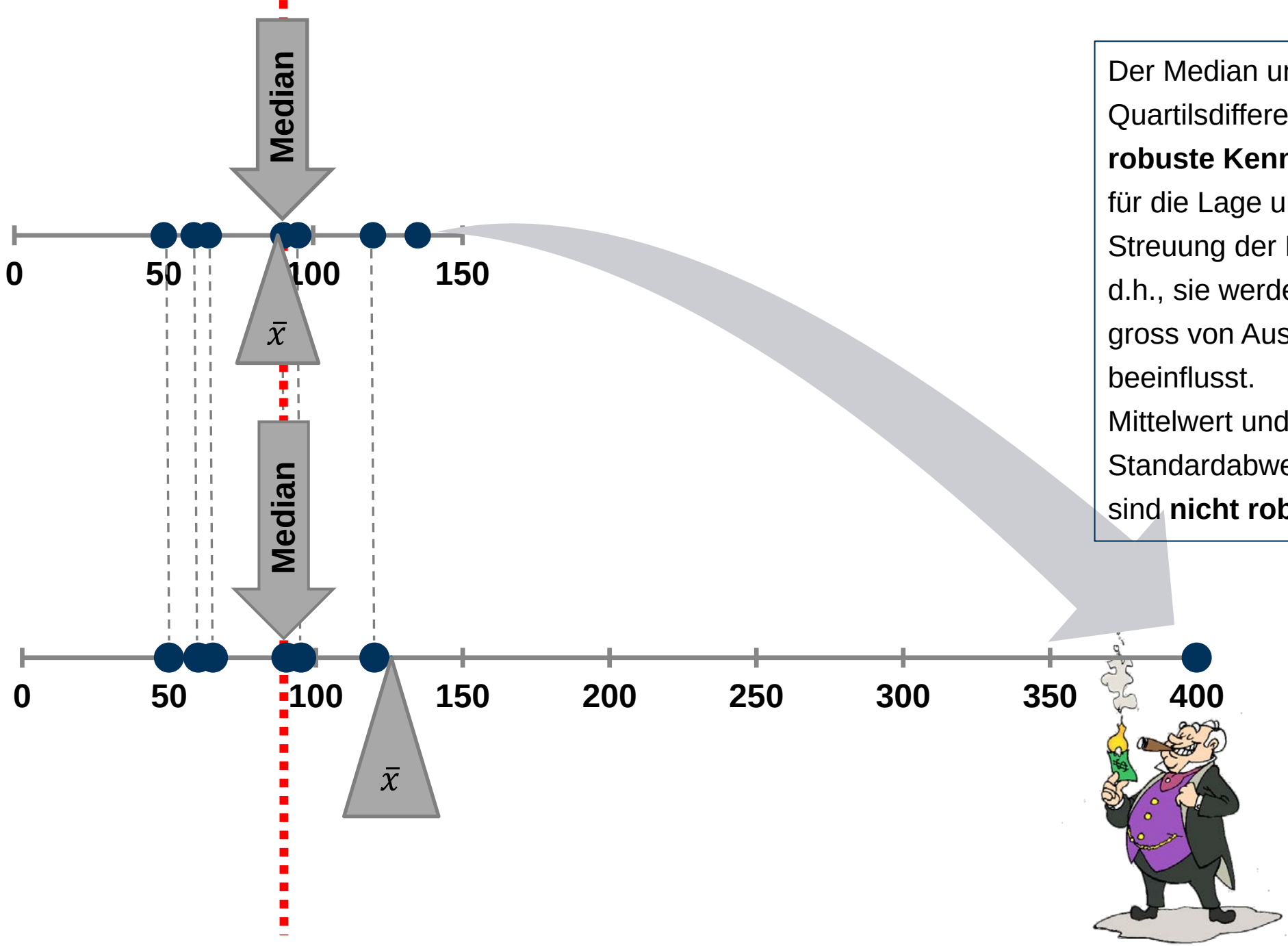
$0.5 \cdot 11 = 5.5 \rightarrow q_{0.5} = x_{(6)}$

$0.75 \cdot 11 = 8.25 \rightarrow q_{0.75} = x_{(9)}$

Arithmetisches Mittel vs. Median: Einkommen [k CHF]

7 Beobachtungen





Der Median und die Quartilsdifferenz sind **robuste Kennzahlen** für die Lage und die Streuung der Daten, d.h., sie werden nicht gross von Ausreissern beeinflusst.
Mittelwert und Standardabweichung sind **nicht robust**.

Arithmetisches Mittel vs. Median

»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

Grafische Darstellungen: Überblick

Wir behandeln folgende Darstellungen:

- Histogramm
- Boxplot
- empirische kumulative Verteilungsfunktion

Histogramm

- Aufteilung des Wertebereichs in Intervalle der Breite h.
- Zähle Anzahl Beobachtungen in jedem Intervall.
- Graphische Darstellung mit Balken. Höhe der Balken ist

$$\frac{\#(Beobachtungen\ im\ Intervall)}{nh}$$

- Die Gesamtfläche unter dem Histogramm ist 1.
  Die Fläche über einem Intervall entspricht die relative Häufigkeit (vgl Dichte).

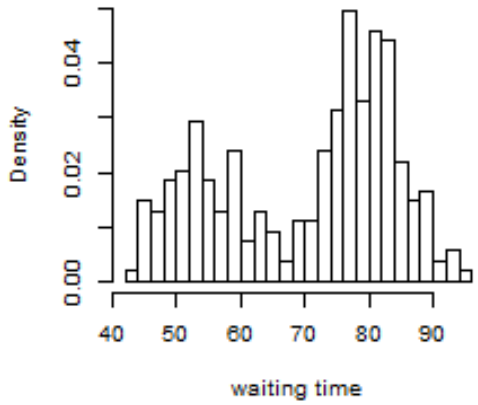
Old Faithful Geysir (Yellowstone): Daten

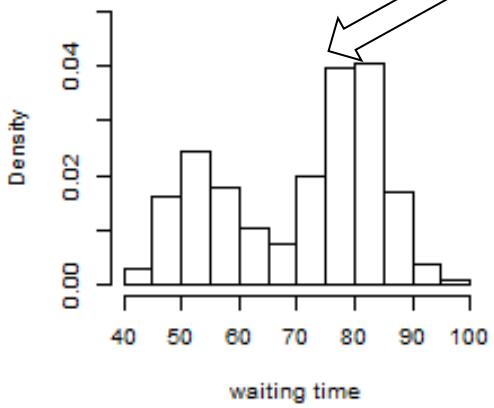
- **Zeitspanne** [Min] zwischen Ausbrüchen
- **Eruptionsdauer** [Min]

- Daten z.B. von hier
  http://stat.ethz.ch/Teaching/Datasets/geysir.dat

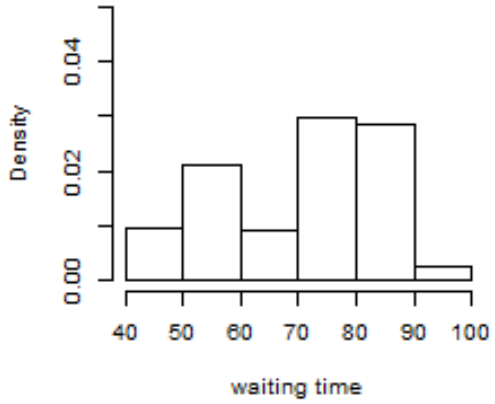
	A	B	C	D
1	Tag	Zeitspanne	Eruptionsdauer	
2	1	78	4.4	
3	1	74	3.9	
4	1	68	4	
5	1	76	4	
6	1	80	3.5	
7	1	84	4.1	
8	1	50	2.3	
9	1	93	4.7	
10	1	55	1.7	
11	1	76	4.9	
12	1	58	1.7	
13	1	74	4.6	
14	1	75	3.4	
15	2	80	4.3	
16	2	56	1.7	
17	2	80	3.9	
18	2	69	3.7	
19	2	57	3.1	
20	2	90	4	
21	2	42	1.8	
22	2	91	4.1	
23	2	51	1.8	

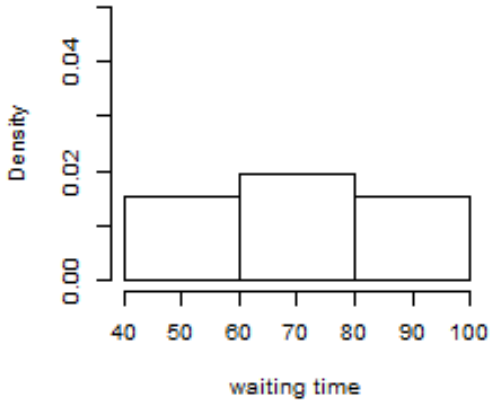
Histogramme der Zeitspanne





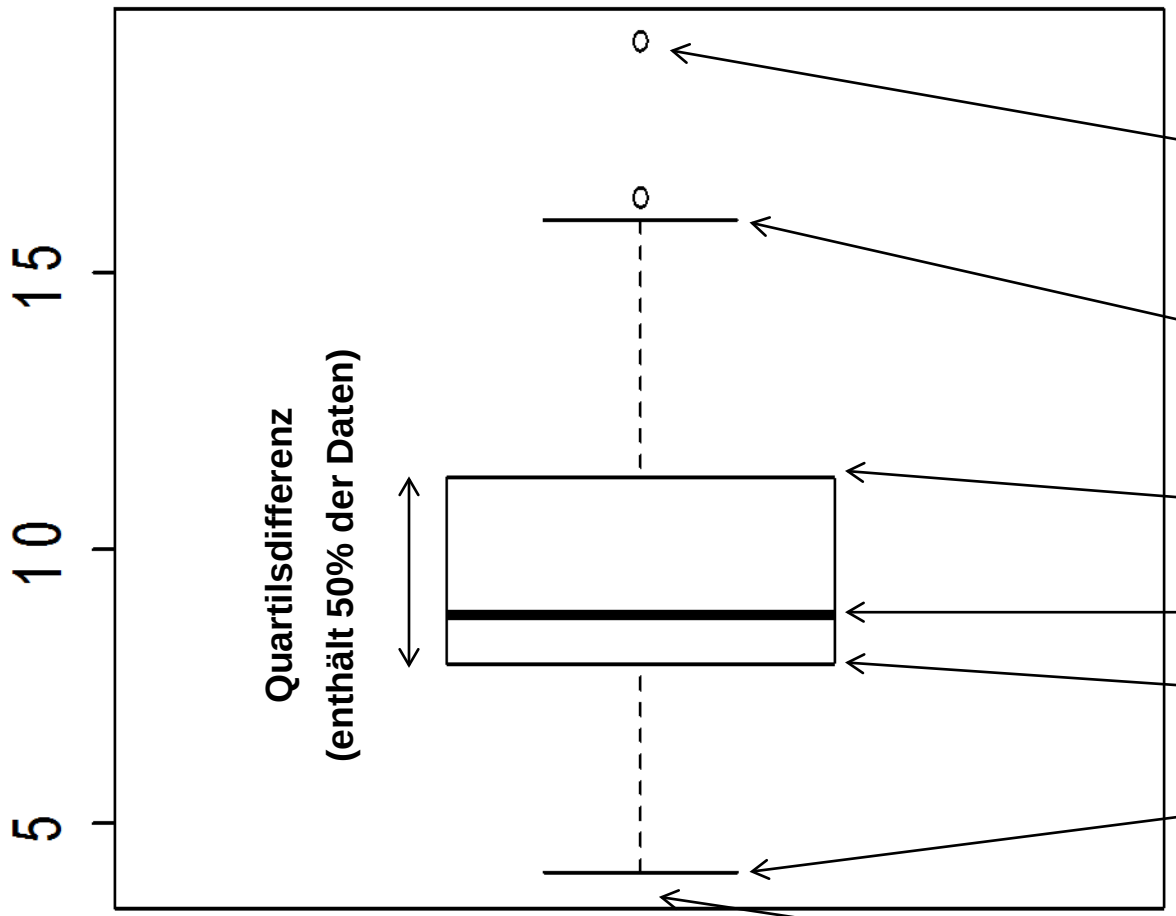
Relative Häufigkeit von Wartezeiten im Intervall [70,80] ist etwa
0.02*(75-70) + 0.04*(80-75) = 30%





- Histogramm ergibt oft einen guten Überblick: Symmetrie, Anzahl Gipfel, Lage, Streuung, …
- Je breiter die Klassen, je mehr werden die Daten zusammengefasst (“Erosion”)

Boxplot: Schematischer Aufbau



Ausreisser
(falls vorhanden)

Grösste «normale» Beobachtung

Oberes Quartil

Median

Unteres Quartil

Kleinste «normale» Beobachtung

Ausreisser
(falls vorhanden)

Boxplot: Schematischer Aufbau

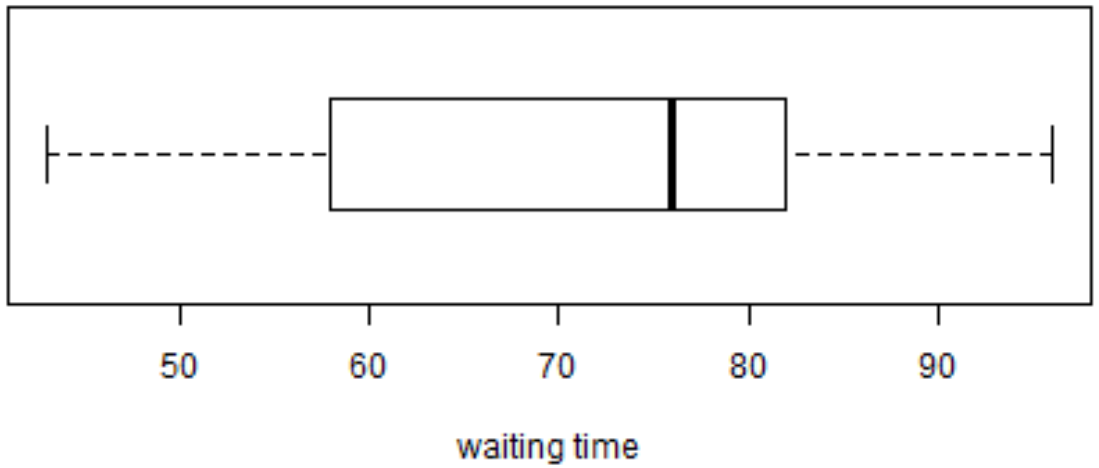
- Die **grösste normale Beobachtung** ist definiert als die grösste Beobachtung, die höchstens $1.5 \cdot IQR$ vom oberen Quartil entfernt ist, wobei IQR die Quartilsdifferenz ist:
  Also grösster Datenwert x_i mit $x_i - q_{0.75} < 1.5 \cdot IQR$

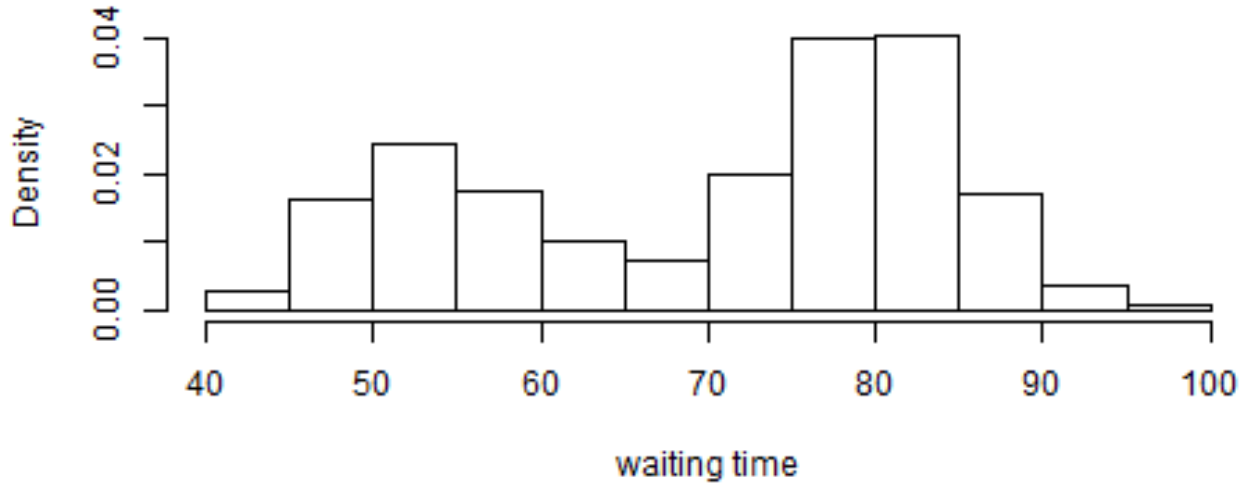
- Die **kleinste normale Beobachtung** ist entsprechend analog definiert mit dem unteren Quartil:
  Also kleinster Datenwert x_i mit $q_{0.25} - x_i < 1.5 \cdot IQR$

- **Ausreisser** sind Punkte, die ausserhalb dieser Bereiche liegen.

Boxplot und Histogramm der Wartezeiten zwischen Eruptionen

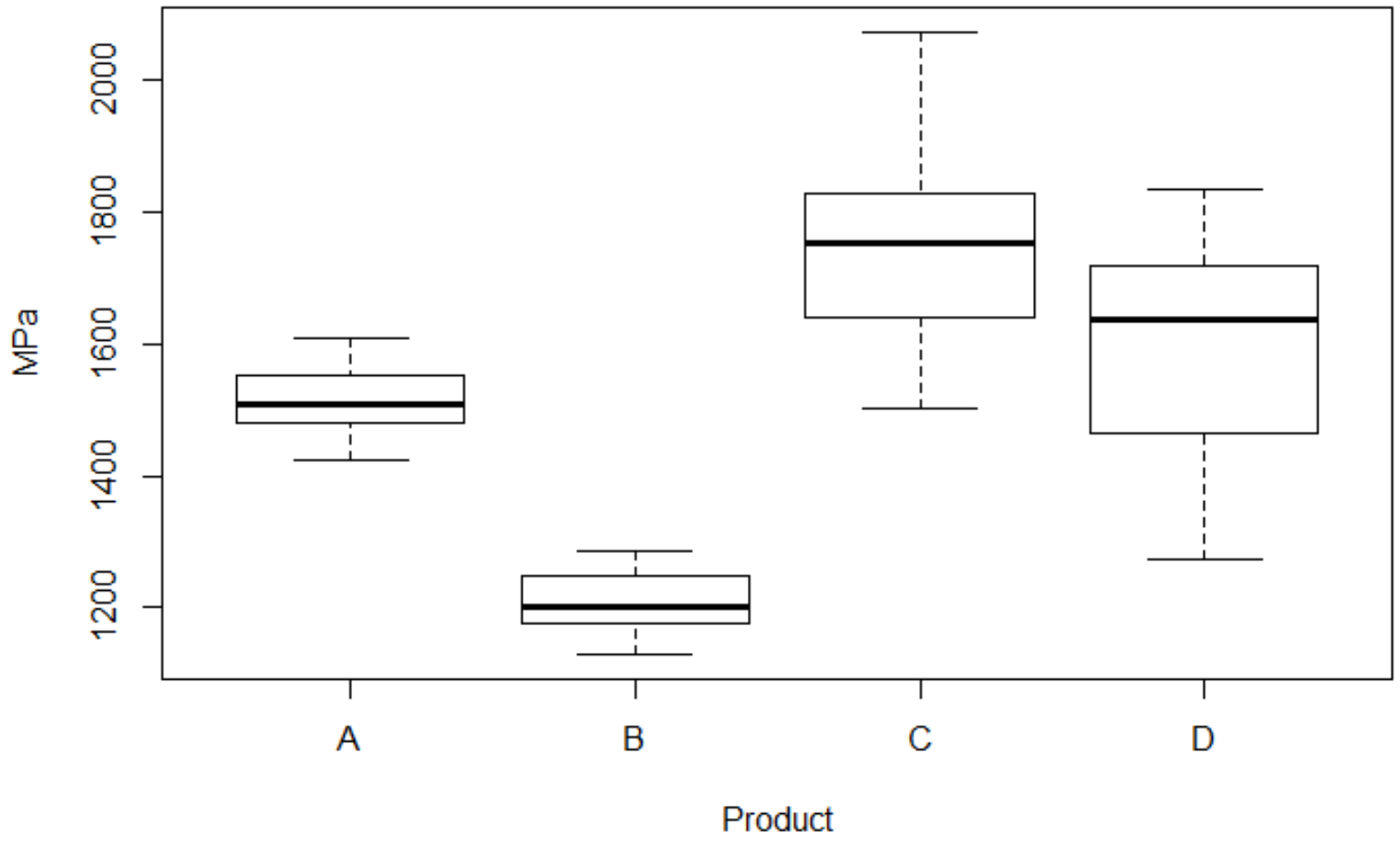
Wir sehen die verschiedenen Peaks im Boxplot nicht!





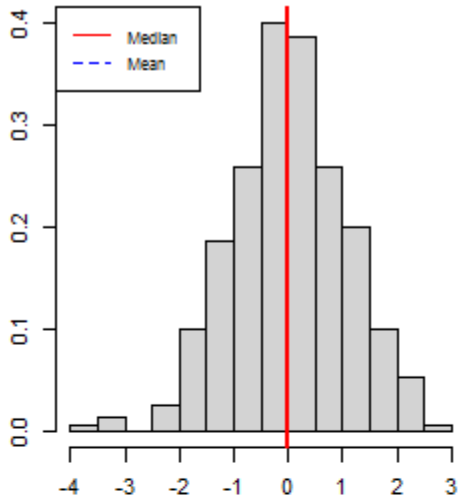
Mehrere Boxplots

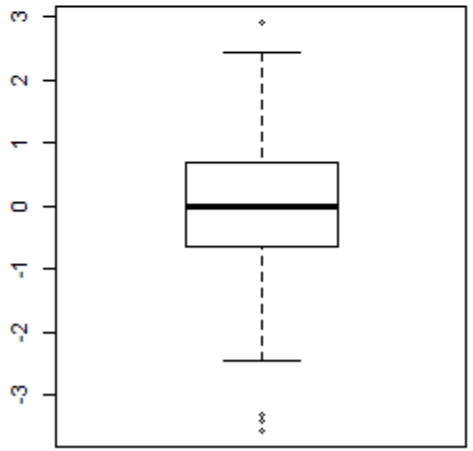
Mit mehreren Boxplots kann man einfach und schnell die Verteilung von verschiedenen Gruppen (Methoden, Produkte, …) vergleichen.



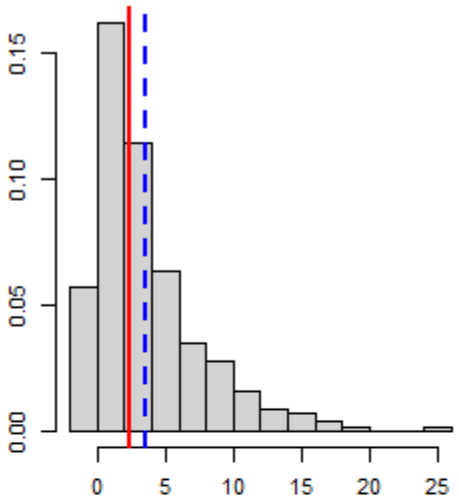
Schiefe

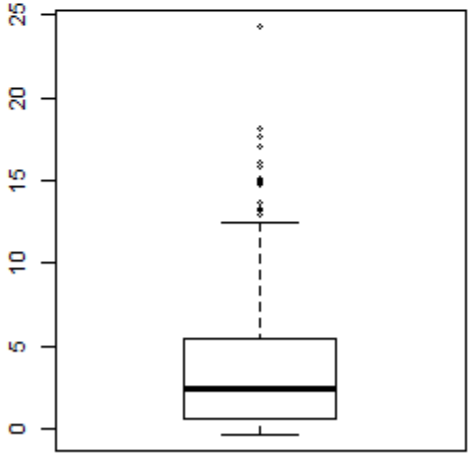
symmetrisch



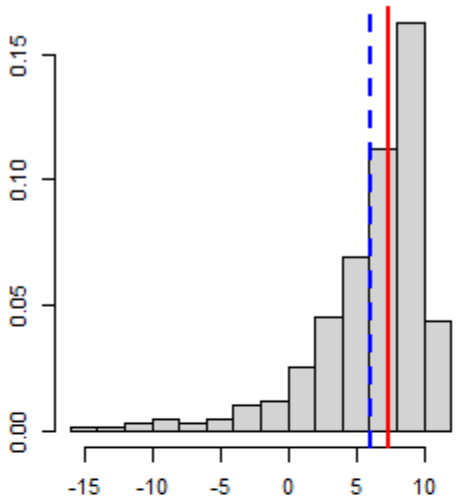


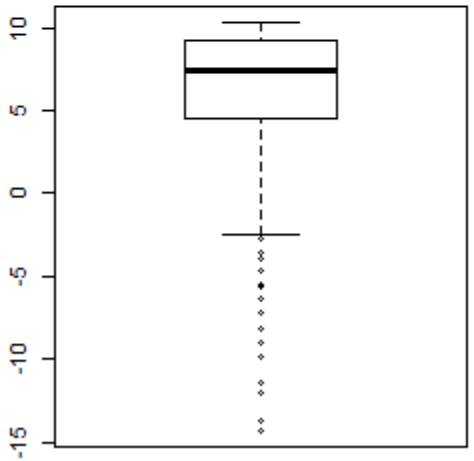
rechtsschief





linksschief





Boxplot: Bemerkungen

- Ein Boxplot ist eine grobere Zusammenfassung als ein Histogramm.
  Es eignet sich gut um mehrere Datensätze zu vergleichen.

- Im Boxplot sind ersichtlich:
  - Lage
  - Streuung
  - Schiefe

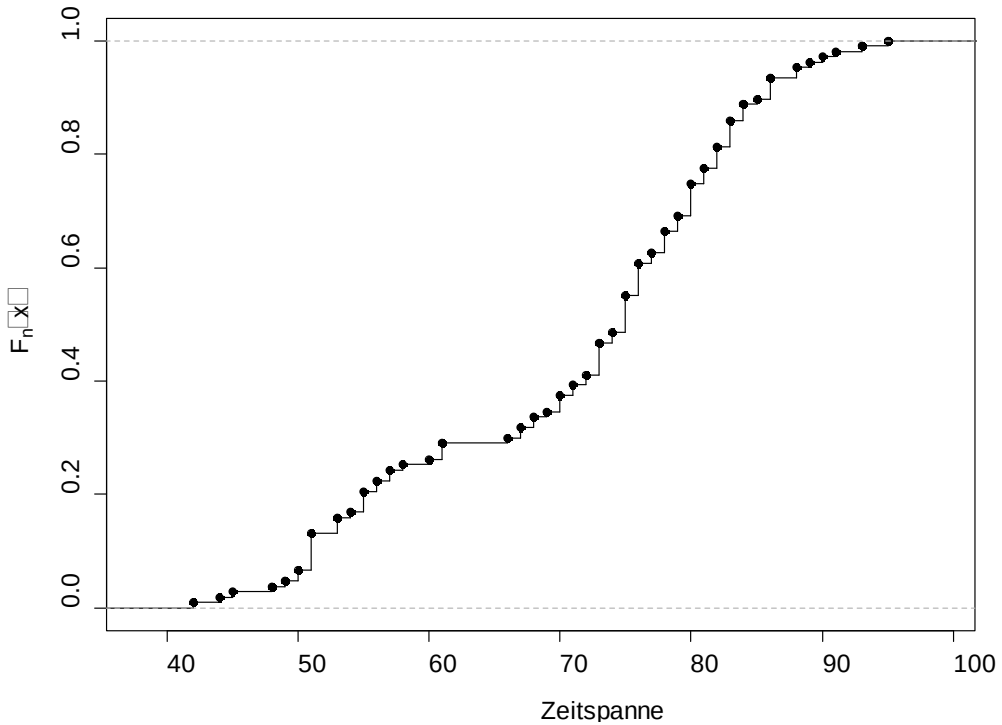
- Man sieht aber z.B. **nicht**, ob eine Verteilung mehrere «Peaks» (Gipfel) hat.

Empirische kumulative Verteilungsfunktion

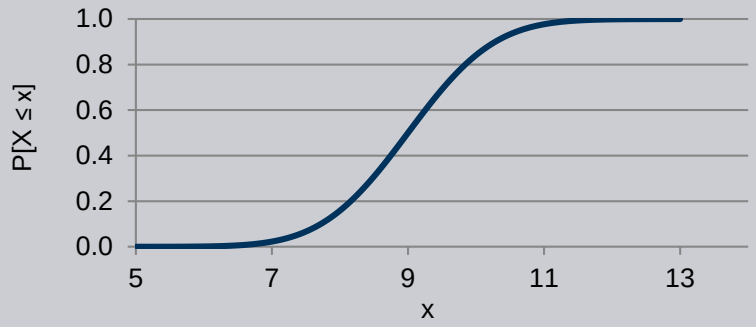
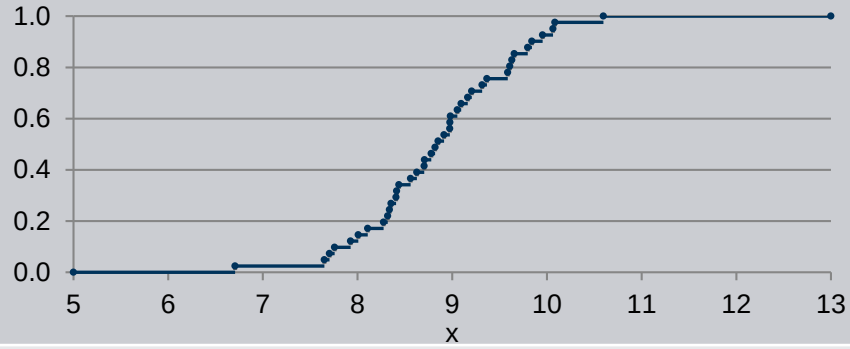
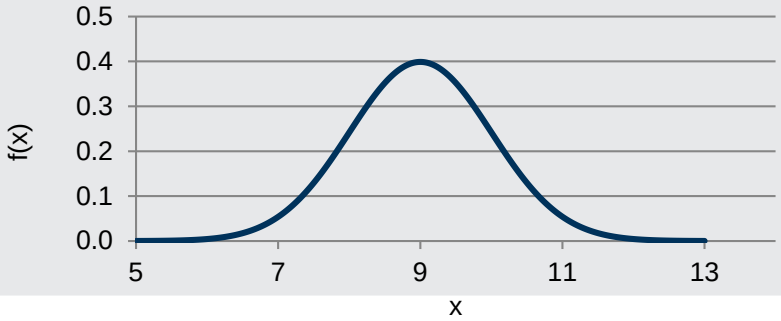
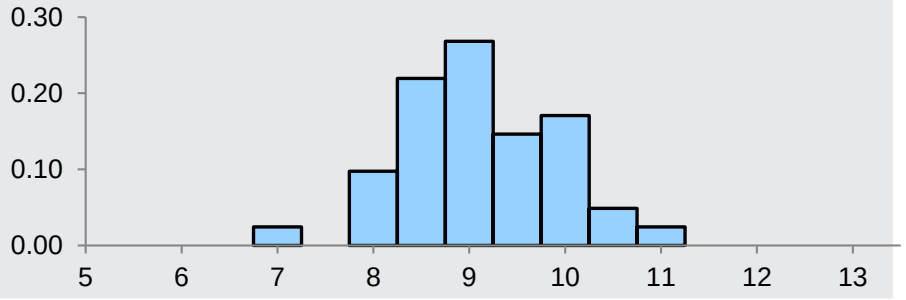
- Empirische kumulative Verteilungsfunktion ist definiert als der **Anteil der Punkte, die kleiner als ein bestimmter Wert** x sind, d.h.

$$F_n(x) = \frac{1}{n}\text{Anzahl}\{i \mid x_i \leq x\}$$

- Bild



Treppenfunktion: Sprunghöhe $1/n$ bei Beobachtungen x_i (bzw. ein Vielfaches davon, wenn es mehrere Beobachtungen mit dem gleichen Wert x_i gibt).

Modell (“Theorie”) ← $n \to \infty$	Daten (beobachtete Stichprobe)
Erwartungswert $E[X]$	Arithm. Mittel $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Varianz $\mathrm{Var}(X)$	Empirische Varianz $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
Kumulative Verteilungsfunktion 	Empirische kumulative Verteilungsfunktion 
Dichte 	Histogramm (normiert auf Fläche 1) 

Deskriptive Statistik: 2 Dimensionen

- Wir haben nun **paarweise** beobachtete Daten

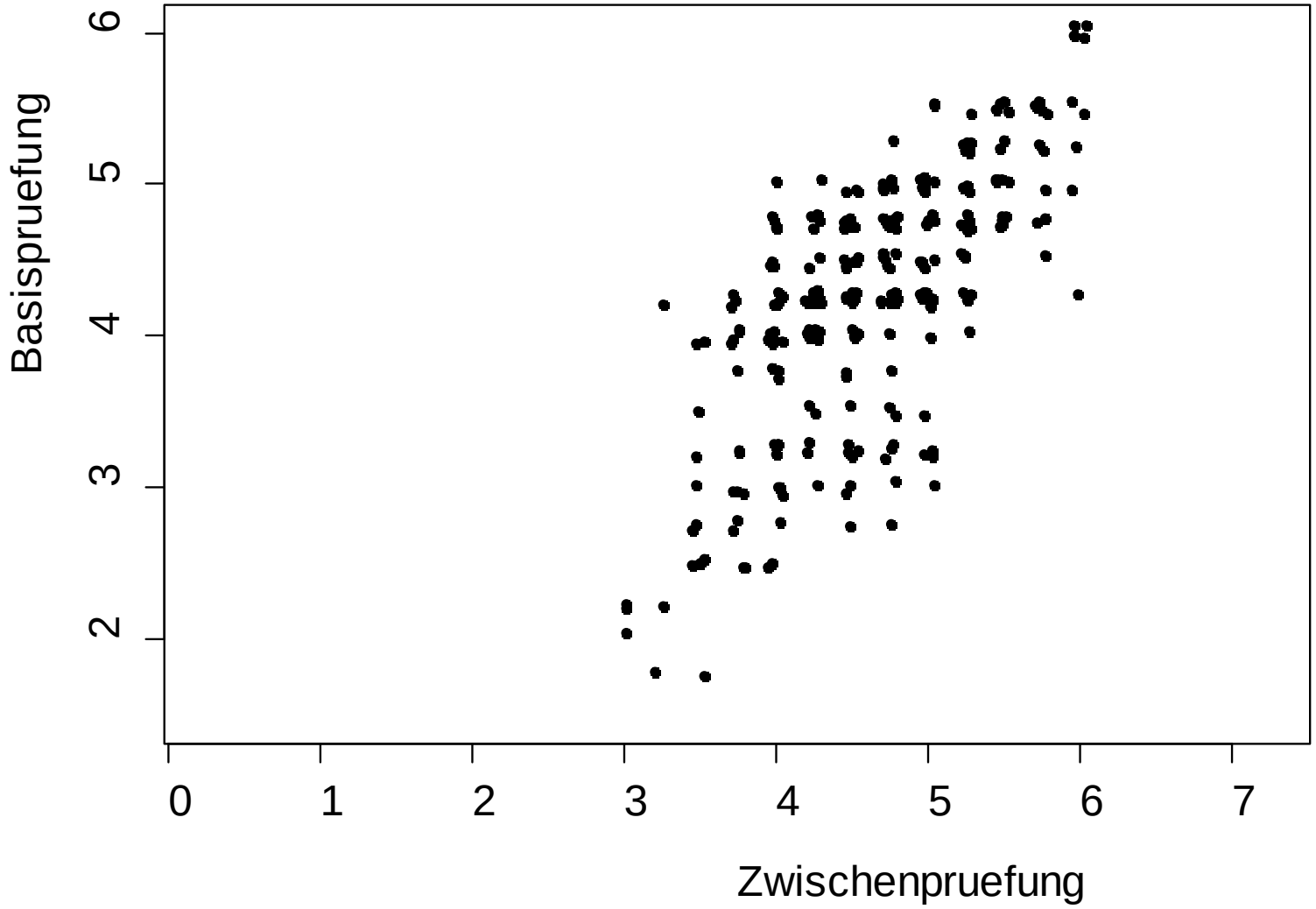
$$x_1, \dots, x_n$$
$$\updownarrow \quad \updownarrow$$
$$y_1, \dots, y_n$$

- Zum Beispiel die Note der Basisprüfung (y_i) und die Note der Zwischenprüfung (x_i) der Studenten. Oder die Eruptionsdauer (y_i) und die Zeitspanne (x_i) zum vorangehenden Ausbruch des Old Faithful Geysir.

- Neue Grafiken/Kennzahlen:
  - zweidimensionales Streudiagramm
  - empirische Kovarianz und Korrelation

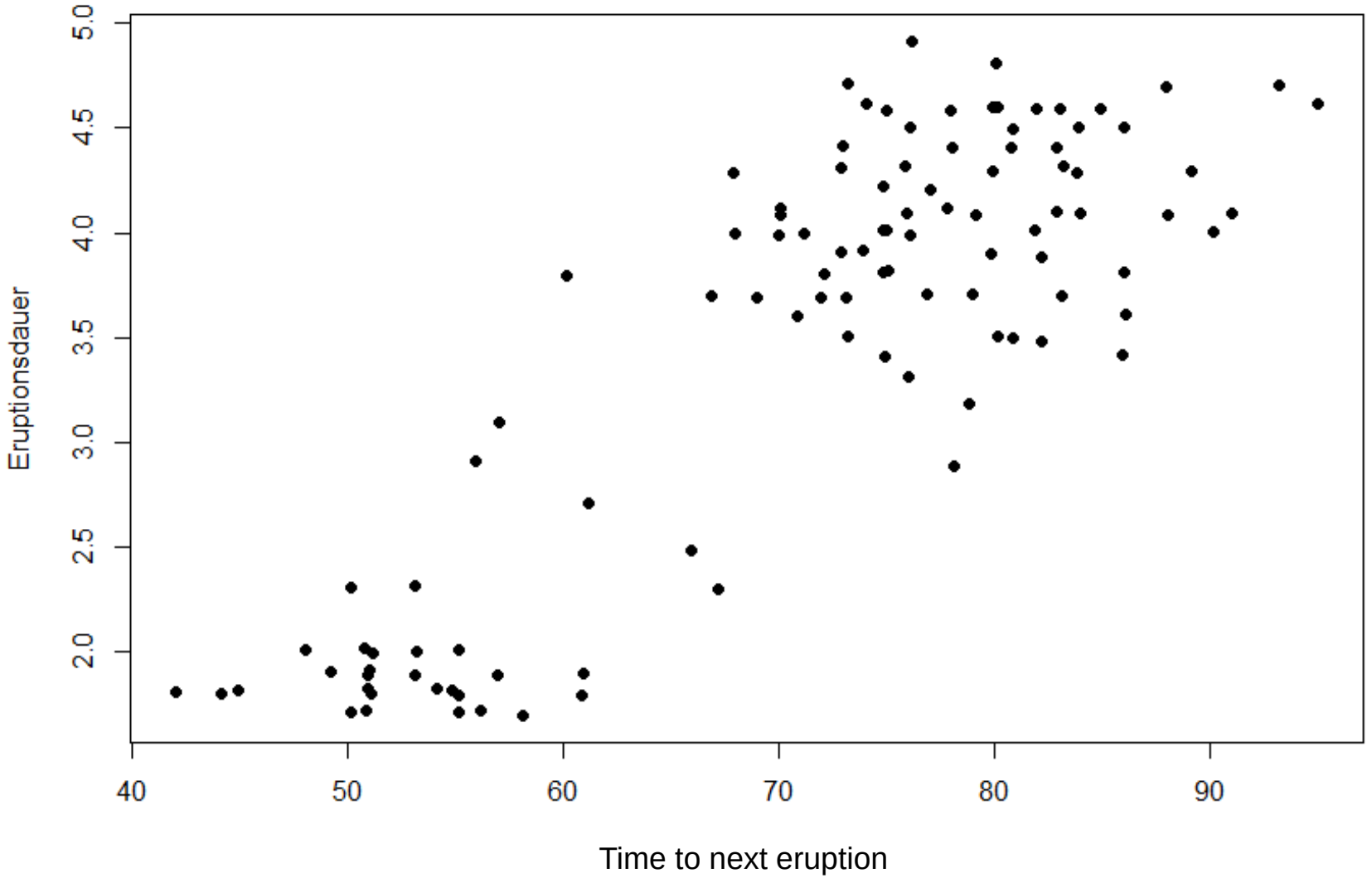
Zweidimensionales Streudiagramm

Beispiel der Zwischen- und Basisprüfung (mit «jittering»):

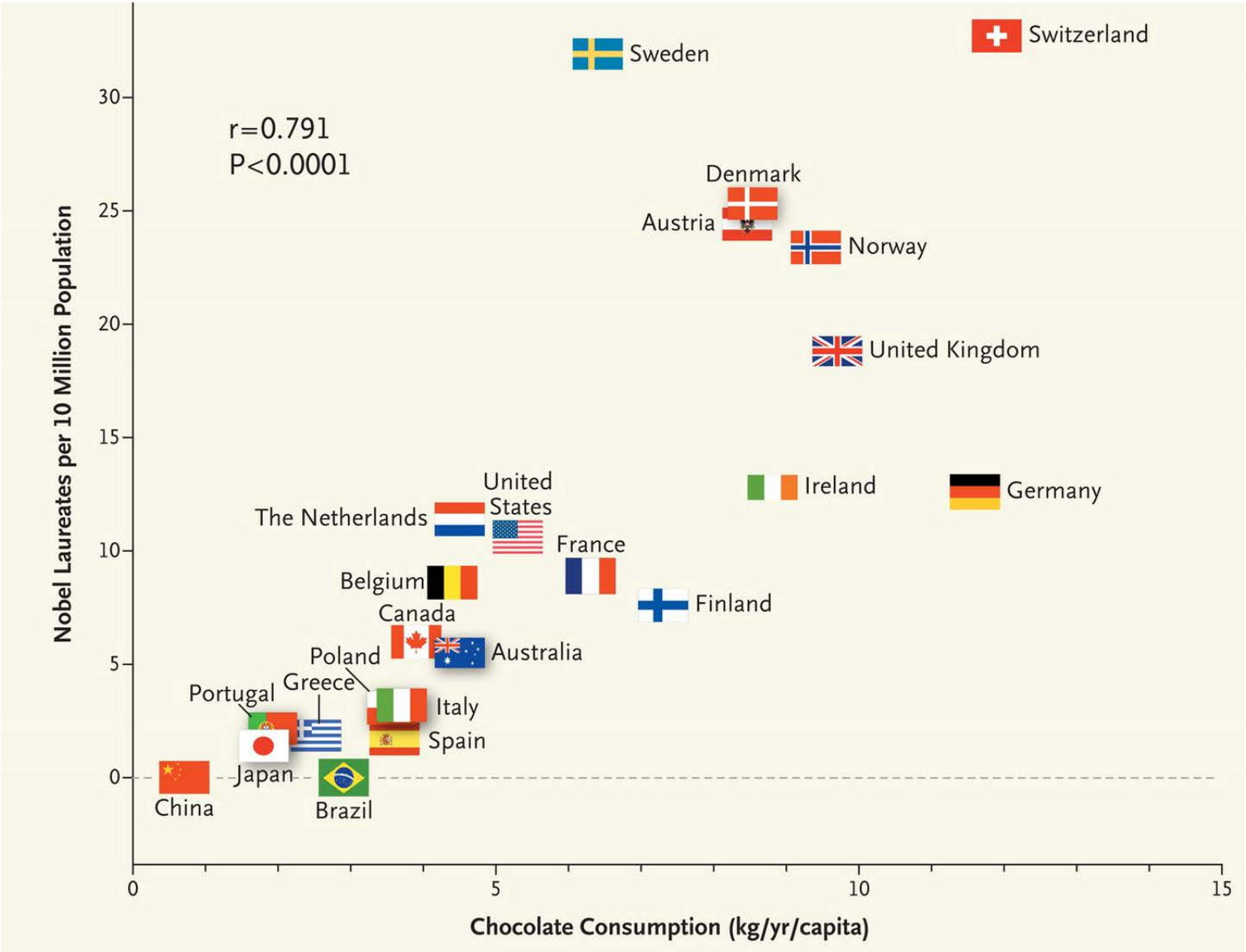


Zweidimensionales Streudiagramm

Beispiel Old Faithful:



Zusammenhänge gibt es viele…



Quelle: The New England Journal of Medicine

Empirische Kovarianz und Korrelation

- **Empirische Kovarianz**

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

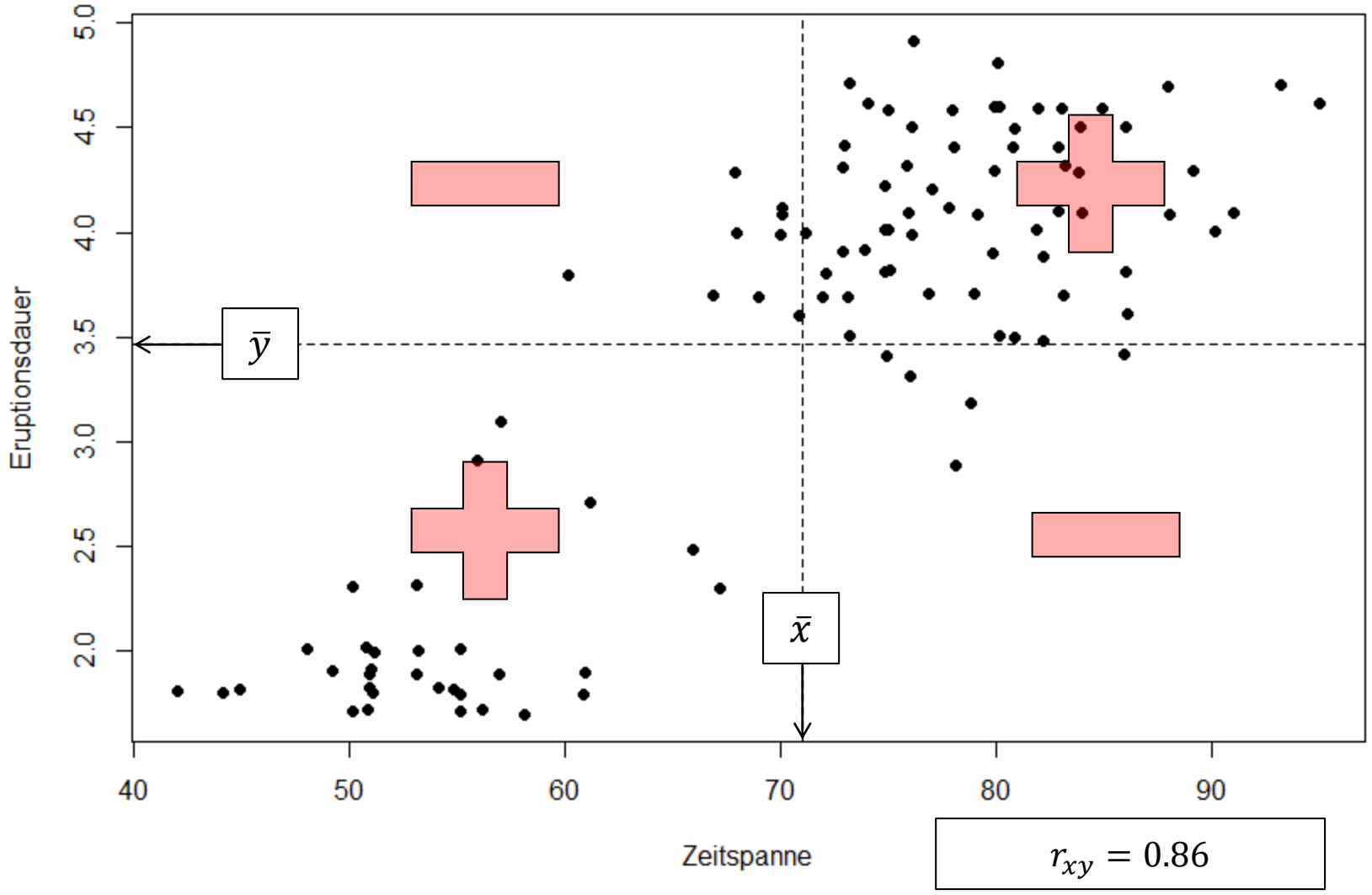
- **Empirische Korrelation**

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \in [-1, 1]$$

wobei s_x, s_y die empirischen Standardabweichungen sind.

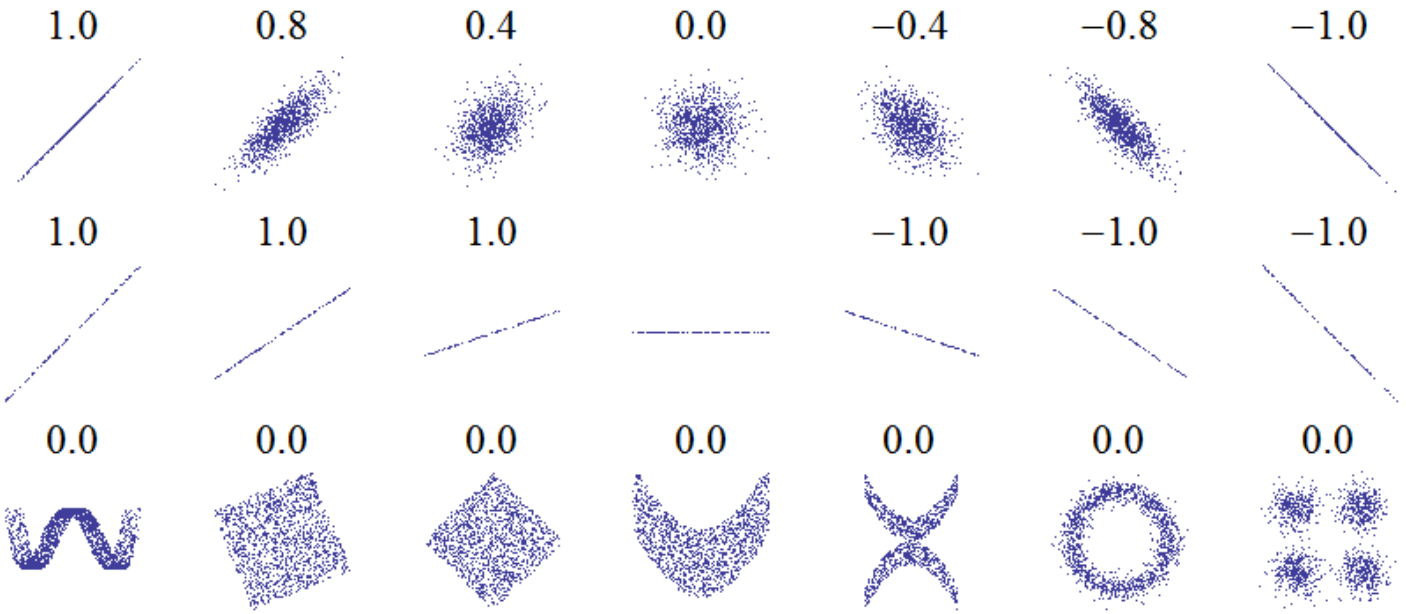
Empirische Kovarianz und Korrelation

Beitrag eines Datenpaares zur empirischen Kovarianz/Korrelation



Empirische Korrelation: Bemerkungen

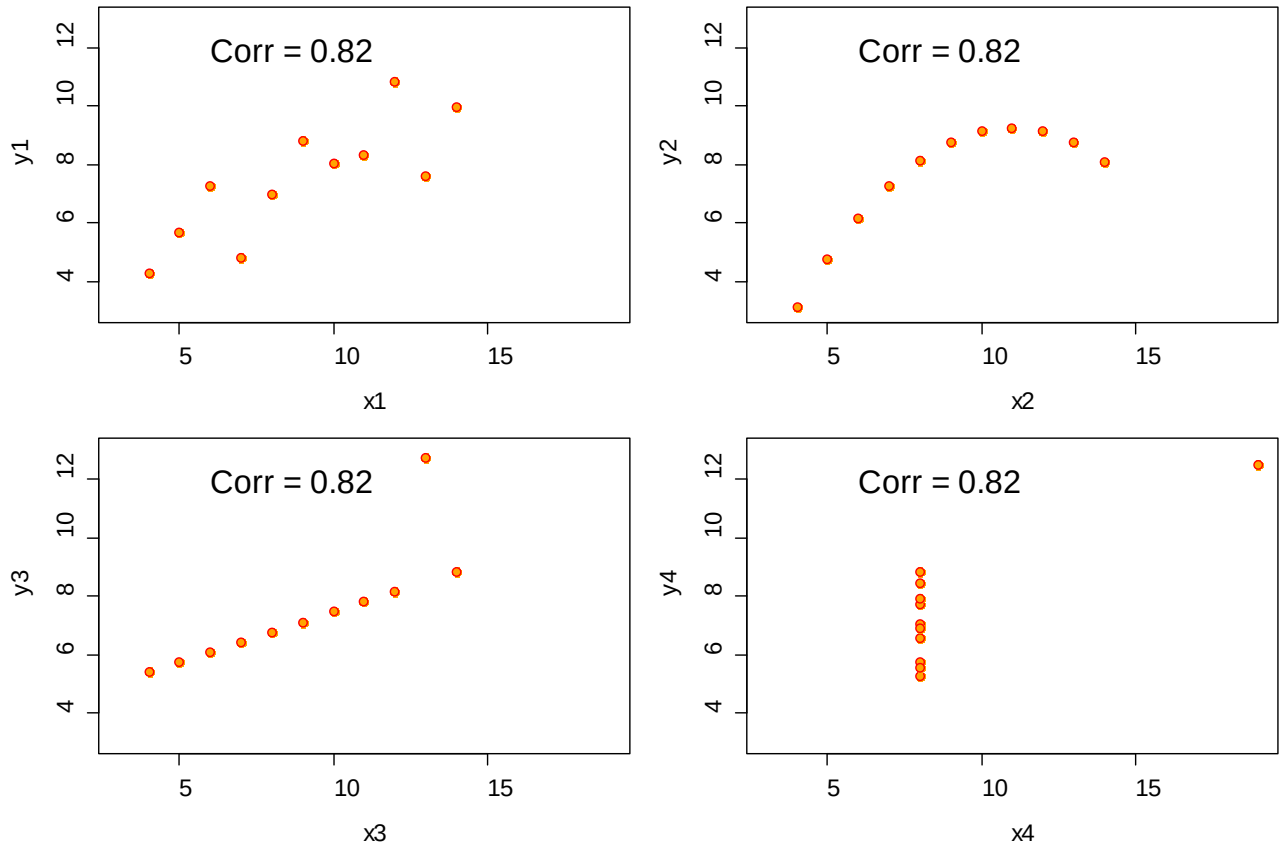
- Korrelation misst «nur» den **linearen Zusammenhang**.
- Das Zeichen von r_{xy} misst die «Richtung» der linearen Zusammenhang. Der Betrag $|r_{xy}|$ misst die «Stärke» der linearen Zusammenhang.



Quelle: Wikipedia

Pass auf:
Hier gibt es einen nicht-linearen Zusammenhang zwischen *X* und *Y*, der nicht von r_{xy} detektiert wird.

Empirische Korrelation: ein anderes klassisches Beispiel



- Man sollte die Daten immer auch anschauen, statt sich «blind» auf Kennzahlen zu verlassen.