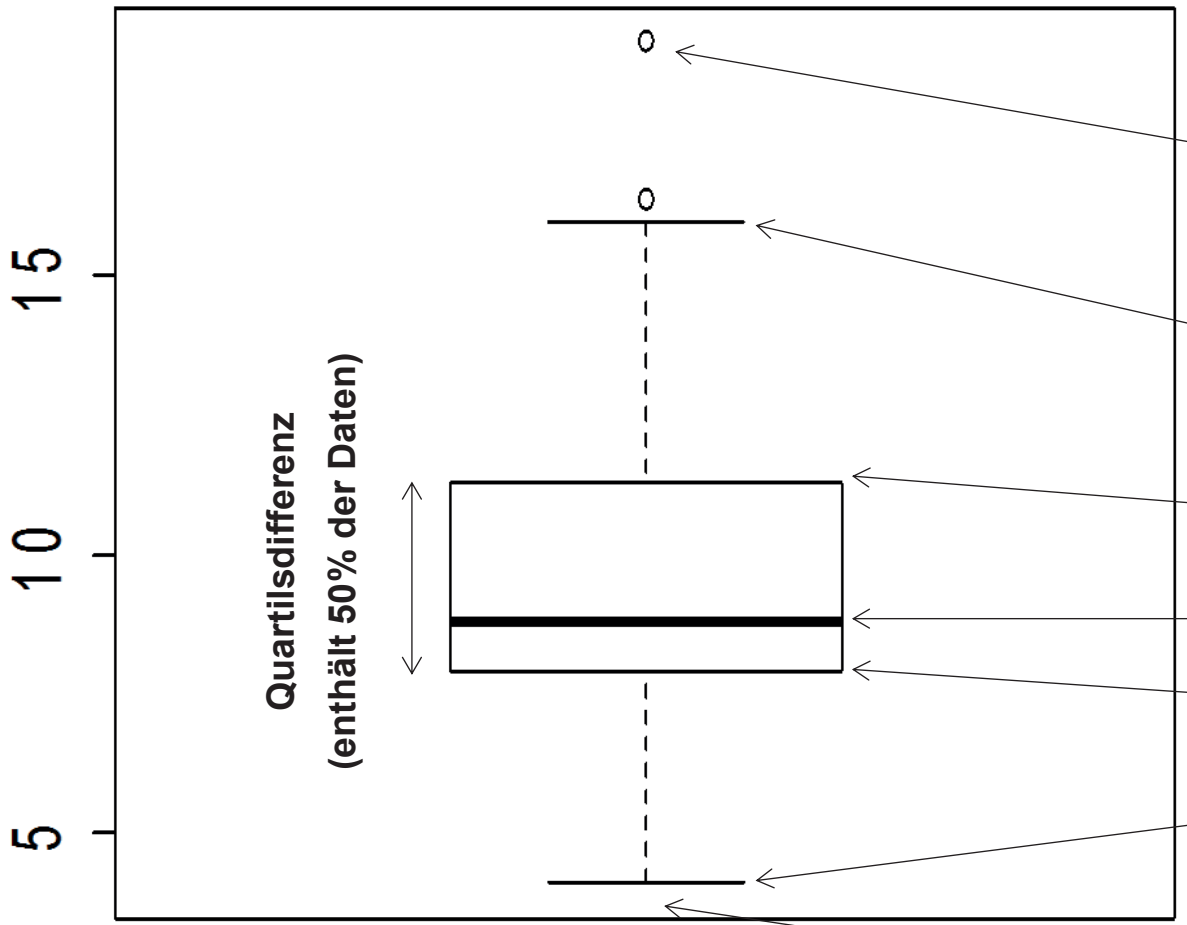
Boxplot
Lineare Regression

Peter Büchel

HSLU I

Stat: SW02

Boxplot: Schematischer Aufbau



Ausreisser
(falls vorhanden)

Grösste «normale» Beobachtung

Oberes Quartil

Median

Unteres Quartil

Kleinste «normale» Beobachtung

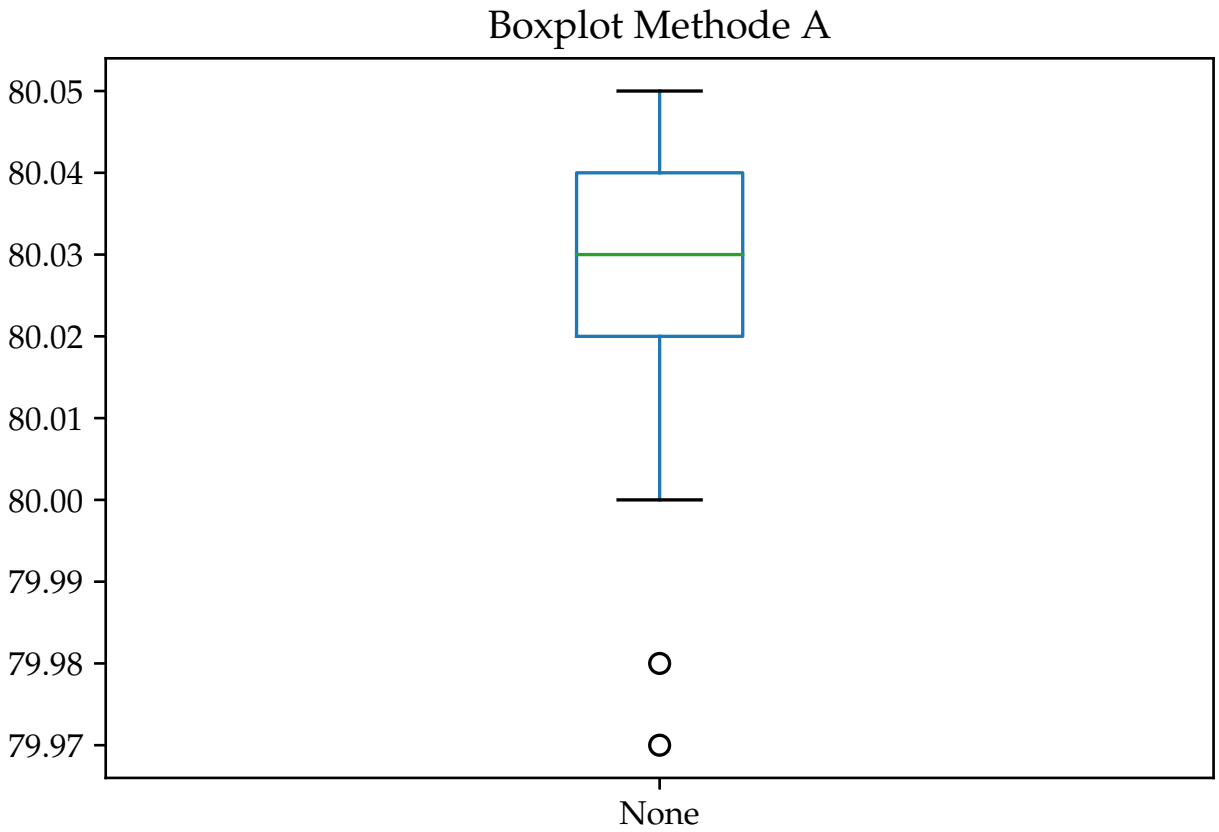
Ausreisser
(falls vorhanden)

Boxplot: Schematischer Aufbau

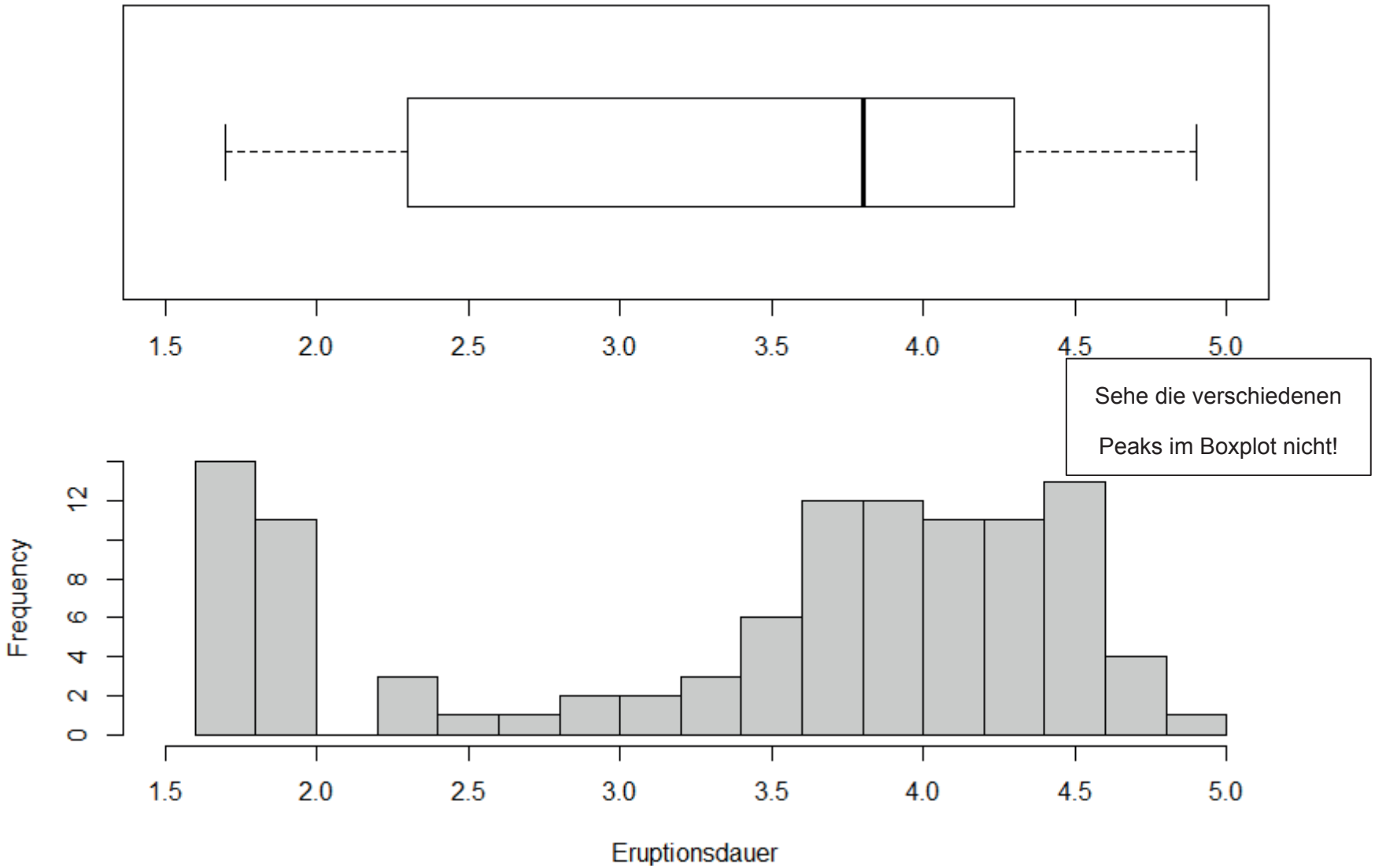
- *Grösste normale Beobachtung*: Grösste Beobachtung, die höchstens $1.5 \cdot r$ vom oberen Quartil entfernt ist (r: *Quartilsdifferenz*)
- *Kleinste normale Beobachtung*: Analog definiert mit dem unteren Quartil
- *Ausreisser* sind Punkte, die ausserhalb dieser Bereiche liegen

Python

```
methodeA.plot(kind="box", title="Boxplot Methode A")
```

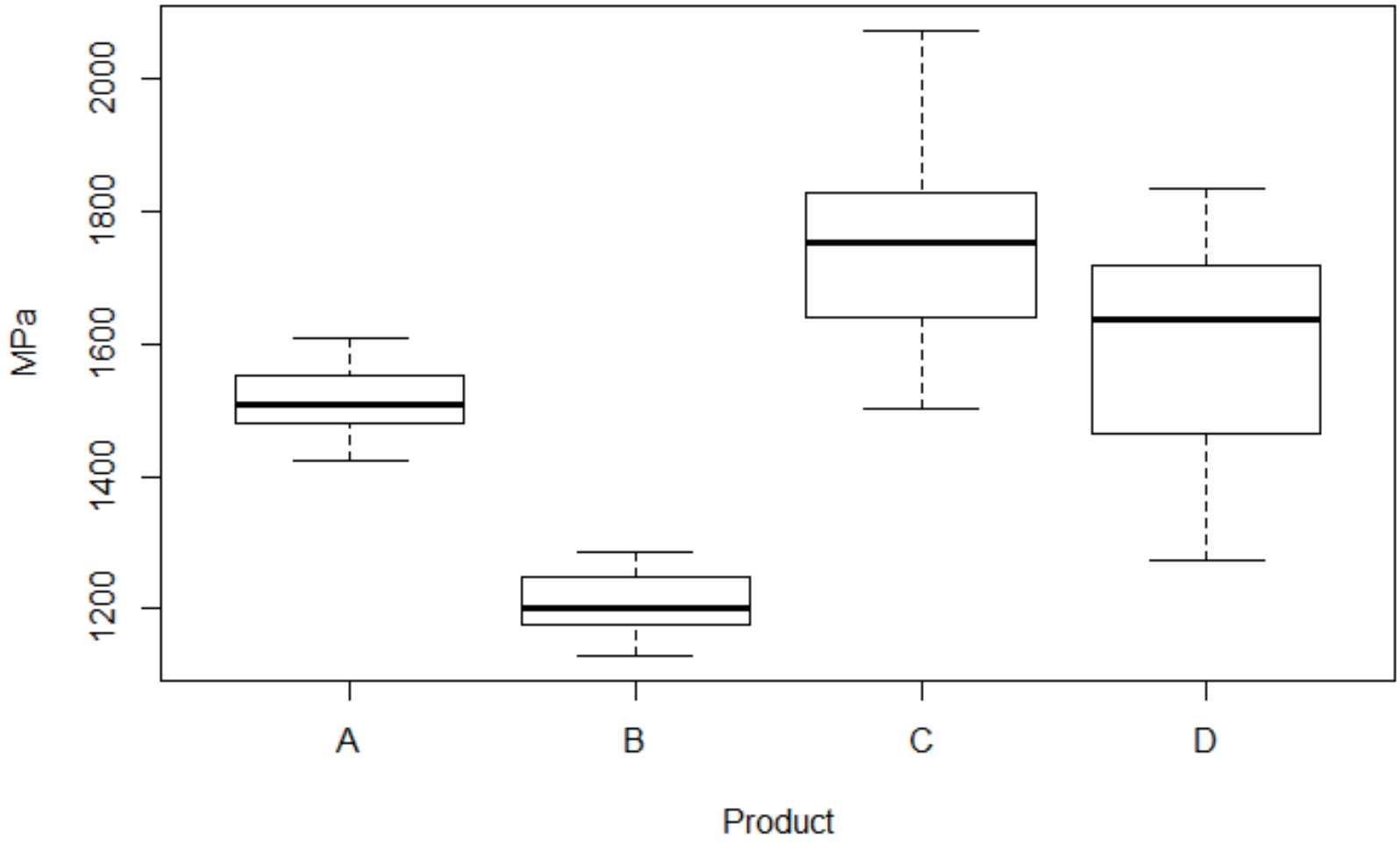


Boxplot und Histogramm der Eruptionsdauer

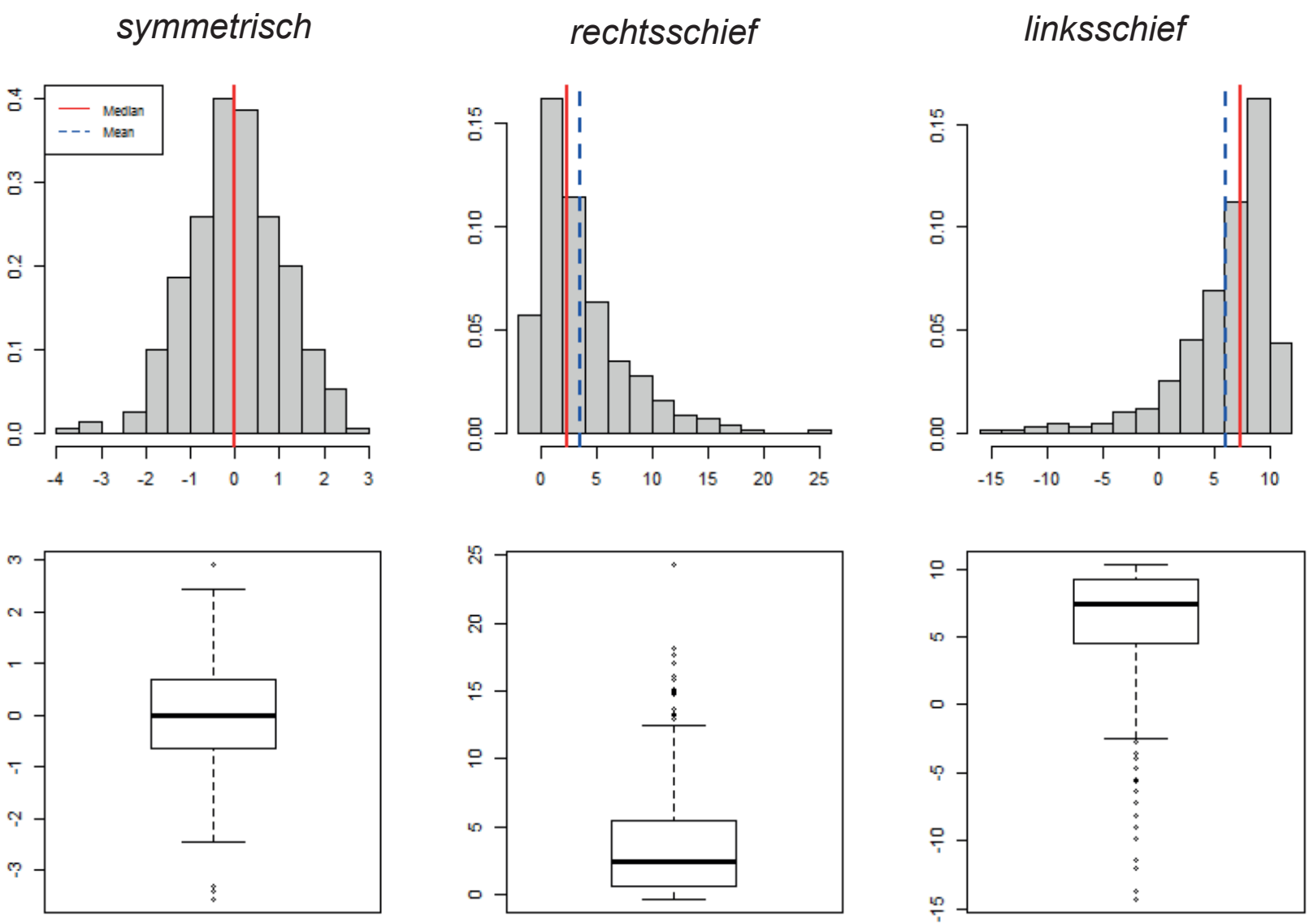


Mehrere Boxplots

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



Schiefe



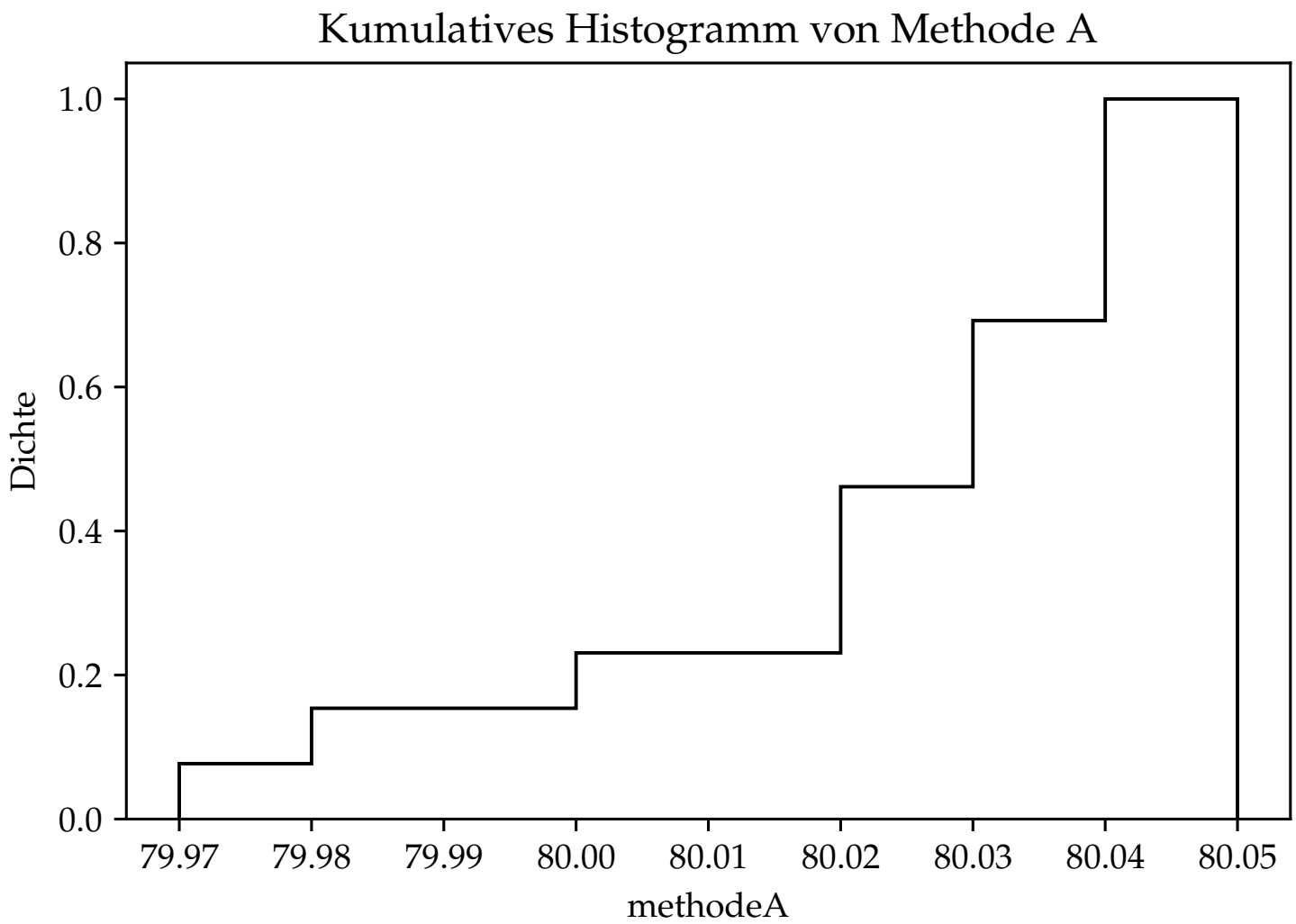
Boxplot: Bemerkungen

- Im *Boxplot* sind ersichtlich:
  - Lage
  - Streuung
  - Schiefe
- Man sieht aber z.B. *nicht*, ob eine Verteilung mehrere „Peaks" hat

Empirische kumulative Verteilungsfunktion

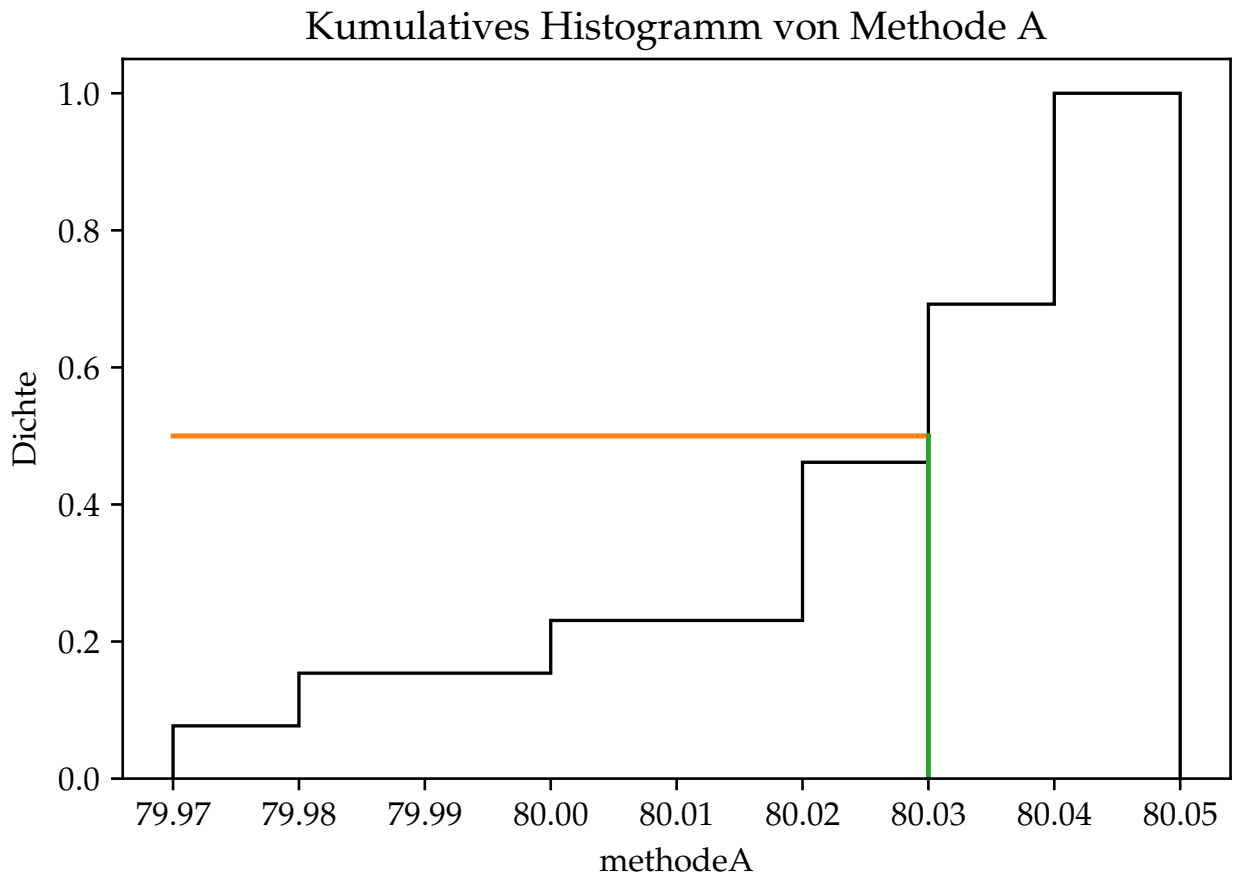
- *Empirische kumulative Verteilungsfunktion* $F_n(\cdot)$ ist eine Treppenfunktion, mit:
  - Links von $x_{(1)}$ ist die Funktion gleich null
  - Bei jedem $x_{(i)}$ wird ein Sprung der Höhe $\frac{1}{n}$ gemacht
  - Wert kommt mehrmals vor $\rightarrow$ Sprung entsprechendes Vielfache von $\frac{1}{n}$

- Beispiel: Kumulative Verteilungsfunktion der Methode *A*



- Abbildung entsteht wie folgt:
  - Jeder Beobachtung wird ein Dichtewerte von $\frac{1}{13}$ zugeordnet
  - Links von 79.97 ist Funktion 0 (es hat keinen kleineren Beobachtungswert)
  - Bei 79.97 macht die Funktion einen Sprung auf $n = \frac{1}{13} \approx 0.077$
  - Funktion bleibt dann gleich bis 80.00, da es vorher keinen zusätzlichen Beobachtungswert gibt
  - Bei 80.00 macht die Funktion wieder einen Sprung um 0.077 nach oben, weil es dort einen Messwert hat
  - Bei 80.02 macht die Funktion einen Sprung um $3 \cdot 0.077$ nach oben, da es dort 3 Beobachtungswerte gibt
  - usw.
  - Bei 80.05 letzten Sprung $\rightarrow$ Funktionswert wird 1

- Was kann man aus der kumulativen Verteilungfunktion herauslesen?
  - Bei 0.5 auf vertikaler Achse werden gerade die Hälfte aller Werte aufsummiert



- Zeichnen von 0.5 horizontale Linie $\rightarrow$ grüne Linie in Abbildung schneidet kumulative Verteilungsfunktion bei 80.03
- Das entspricht gerade dem Median
- Dort, wo die kumulative Funktion steil, viele Beobachtungswerte
- D.h.: die meisten Beobachtungswerte liegen hier zwischen 80.02 und 80.04
- Die Werte entsprechen aber gerade dem unteren und oberen Quartil

Python

```
methodeA.plot(kind="hist", cumulative=True, histtype="step",
normed=True, bins=8, edgecolor="black")
```

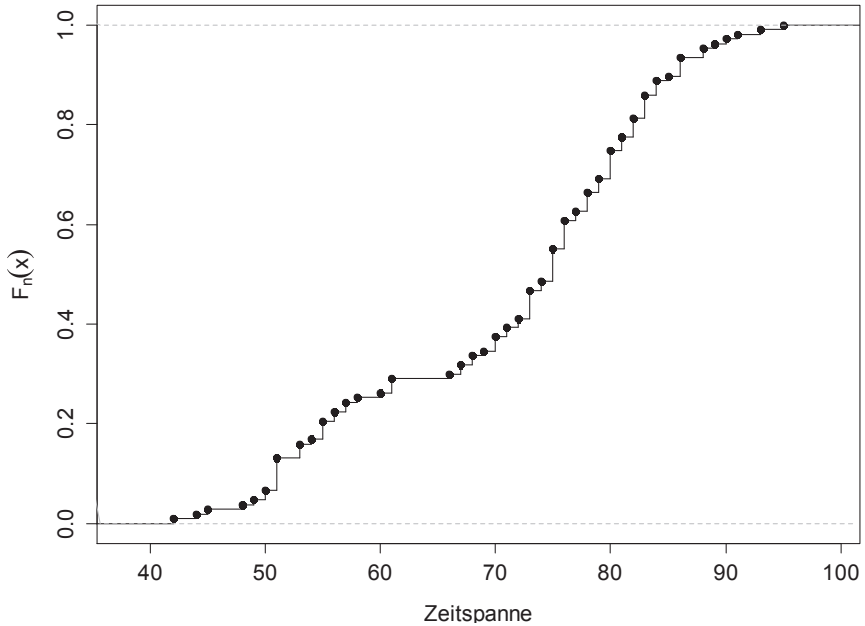
- Kumulative Verteilungsfunktion: `cumulative=True`

Empirische kumulative Verteilungsfunktion

- Empirische kumulative Verteilungsfunktion ist definiert als der *Anteil der Punkte kleiner als ein bestimmter Wert*

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

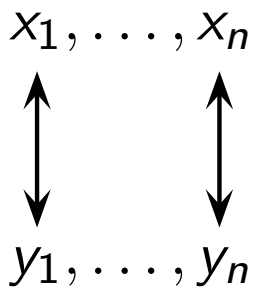
- Kumulative Verteilungsfkt. für Zeitspanne im Geysir-Datensatz



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

Deskriptive Statistik: 2 Dimensionen

- Betrachten nun *paarweise* beobachtete Daten: *Zwei* Messgrössen pro Messeinheit

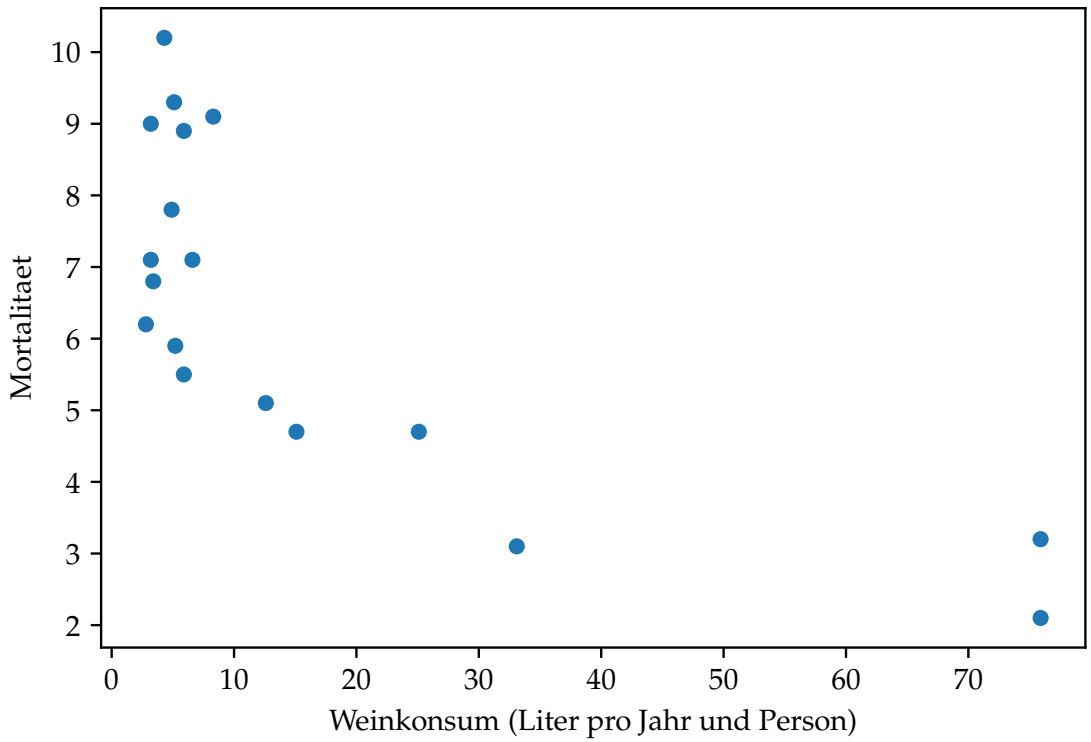


- Weinkonsumation (Liter pro Person pro Jahr) und Mortalität aufgrund von Herzkreislauferkrankung (Todesfälle pro 1000) in 18 Ländern
- Eruptionsdauer (y_i) und die Zeitspanne (x_i) zum vorangehenden Ausbruch des Old Faithful Geysir

Daten: Weinkonsum - Mortalität

Land	Weinkonsum	Mortalität Herzerkrankung
Norwegen	2.8	6.2
Schottland	3.2	9.0
Grossbritannien	3.2	7.1
Irland	3.4	6.8
Finnland	4.3	10.2
Kanada	4.9	7.8
Vereinigte Staaten	5.1	9.3
Niederlande	5.2	5.9
New Zealand	5.9	8.9
Dänemark	5.9	5.5
Schweden	6.6	7.1
Australien	8.3	9.1
Belgien	12.6	5.1
Deutschland	15.1	4.7
Österreich	25.1	4.7
Schweiz	33.1	3.1
Italien	75.9	3.2
Frankreich	75.9	2.1

Zweidimensionales Streudiagramm

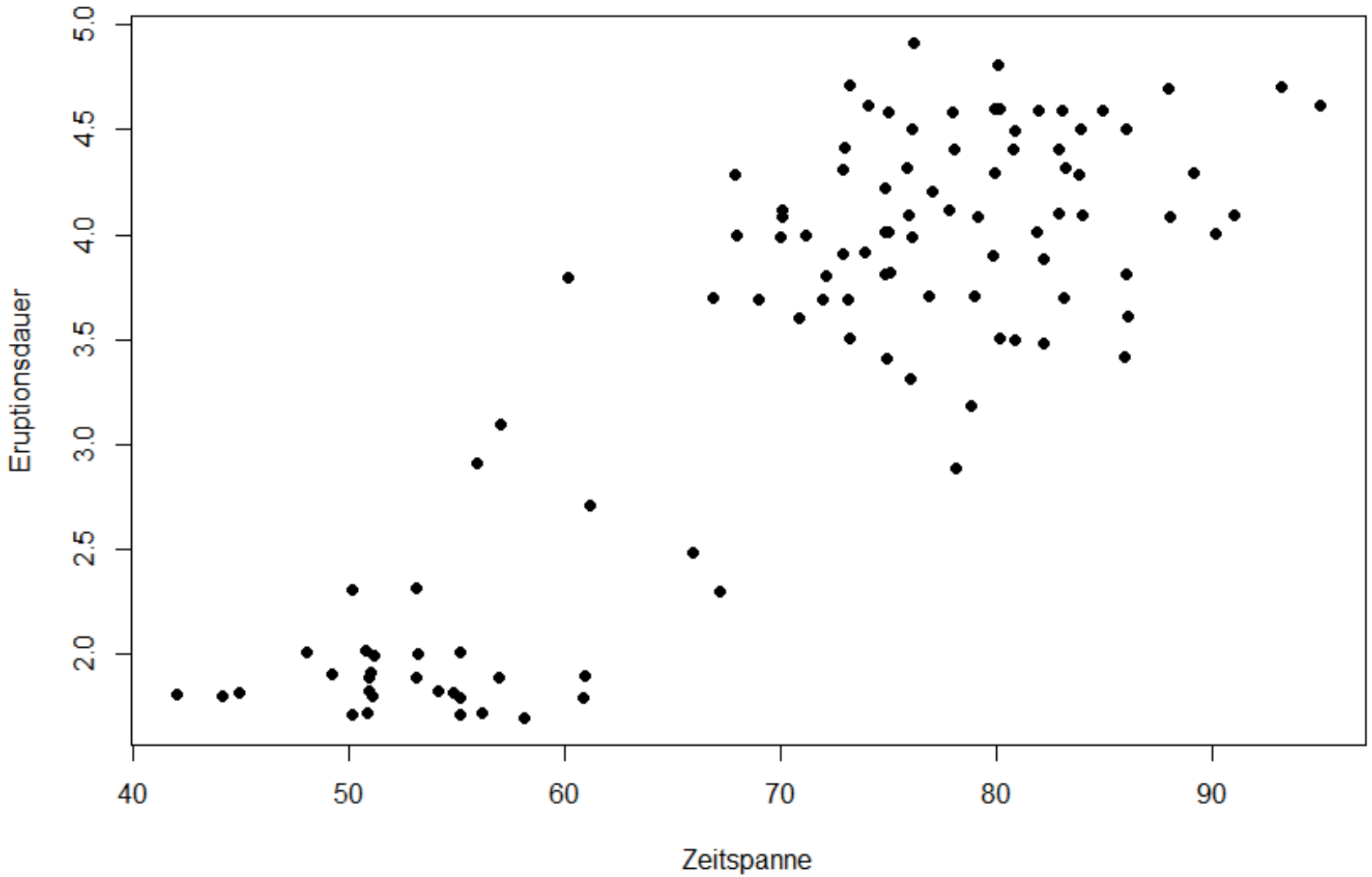


- Plot deutet an, dass hoher Weinkonsum weniger Sterblichkeit wegen Herz-Kreislauferkrankungen zur Folge hat
- Kann Zufall sein (keine Kausalität)
- Heisst *nicht*, dass Weinkonsum gesund ist (Leber!)

Streudiagramm mit Python

```
import pandas as pd
from pandas import DataFrame, Series
import numpy as np

mort = DataFrame({
    "wine": ([2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9,
       5.9, 6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9]),
    "mor": ([6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9,
        5.5, 7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1])
})

mort.plot(kind="scatter", x="wine", y="mor")

plt.xlabel("Weinkonsum (Liter pro Jahr und Person)")
plt.ylabel("Mortalitaet")

plt.show()
```

Beispiel Old Faithful



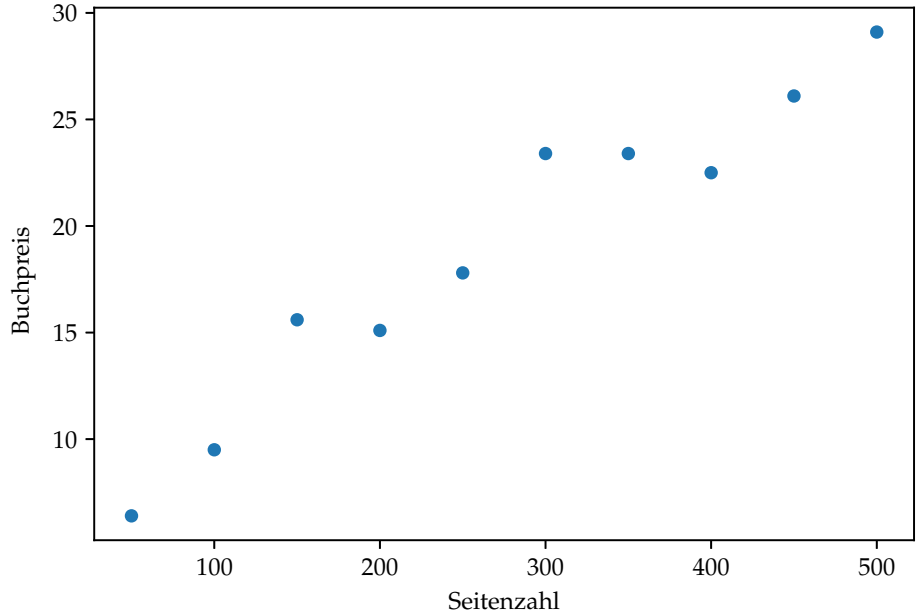
(Fiktives) Beispiel für Lineare Regression

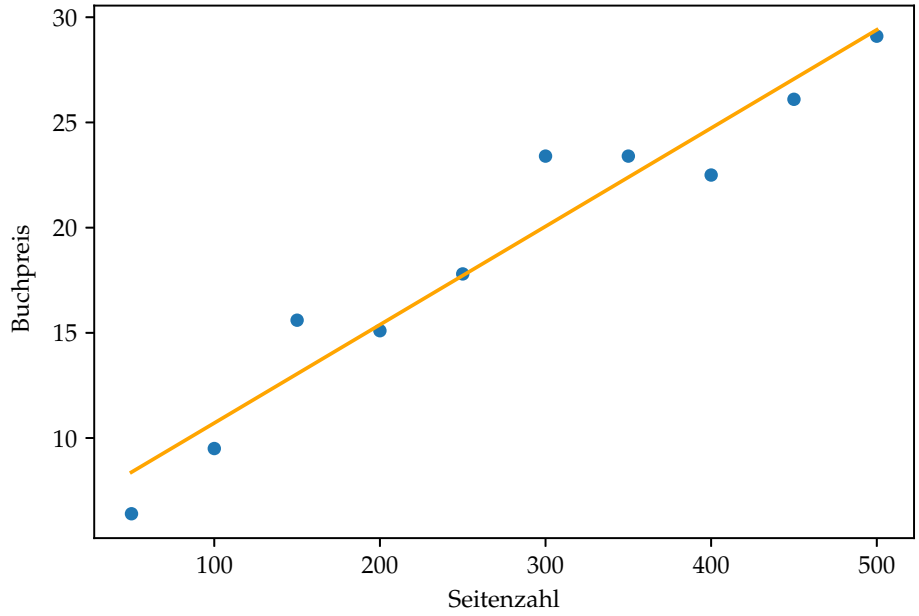
- Kunde kauft in Buchhandlung 10 Bücher

	Seitenzahl	Buchpreis (SFr)
Buch 1	50	6.4
Buch 2	100	9.5
Buch 3	150	15.6
Buch 4	200	15.1
Buch 5	250	17.8
Buch 6	300	23.4
Buch 7	350	23.4
Buch 8	400	22.5
Buch 9	450	26.1
Buch 10	500	29.1

- *Beobachtung*:
  - Je dicker ein Roman ist, desto teurer ist er in der Regel
  - Es gibt Zusammenhang zwischen Seitenzahl x und Buchpreis y
- *Ziel:* Formelmässiger Zusammenhang zwischen Buchpreis und Seitenzahl
- Vorhersagen über Buchpreis für Bücher mit Seitenzahlen, die in Liste nicht auftauchen

Streudiagramm und Regressionsgerade





Regressionsgerade und Residuum

- Vermutung: Gerade scheint recht gut zu den Daten zu passen
- Diese Gerade hätte die Form:

$$y = a + bx$$

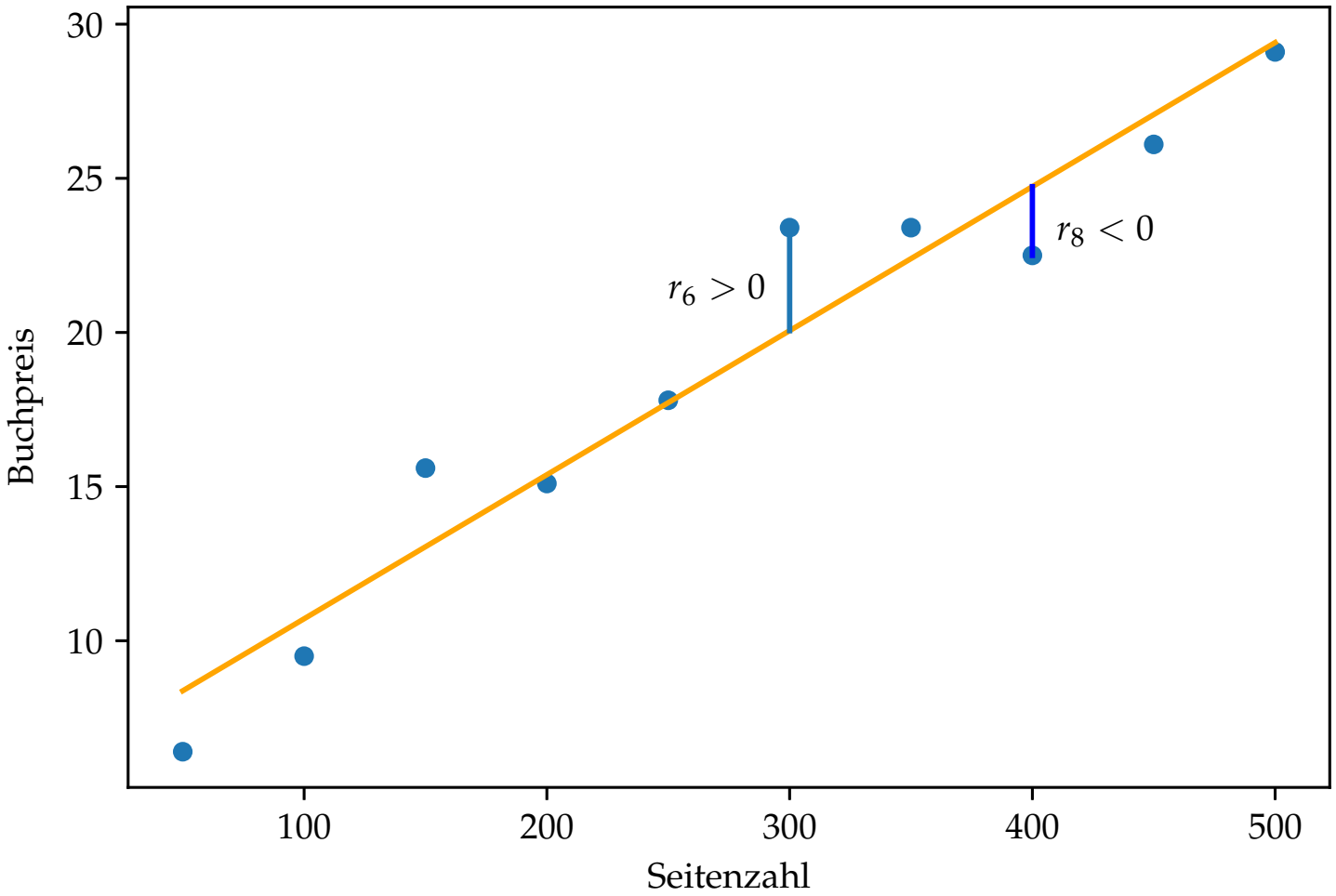
  mit
  - y: Buchpreis; x: Seitenzahl
  - a: Grundkosten des Verlags, b : Kosten pro Seite
- Problem: Gerade finden, die möglichst gut zu allen Punkten passt?

- Möglichkeit: Vertikale Abstände zwischen Beobachtung und Gerade zusammenzählen
- Dabei sollte eine kleine Summe der Abstände eine gute Anpassung bedeuten
- Abstände von Messpunkten zu Geraden $\rightarrow$ neuer Begriff:

Residuum

Der vertikalen Abstand zwischen einem Beobachtungspunkt (x_i, y_i) und der Geraden (der Punkt auf der Geraden ist $(x_i, a + bx_i)$ heisst *Residuum*:

$$r_i = y_i - a - bx_i$$


30
25
20
15
10
Buchpreis
100
200
300
400
500
Seitenzahl
$r_6 > 0$
$r_8 < 0$

- Beispiel: Residuen r_6 und r_8 für *diese* Gerade in Abbildung
- Residuum r_6 positiv, da Punkt überhalb der Gerade
- Entsprechend ist $r_8 < 0$
- Gerade $y = a + bx$ so bestimmen, dass die Summe

$$r_1 + r_2 + \ldots + r_n = \sum_i r_i$$

  minimal wird
- Minimierung von $\sum_i r_i$ hat aber eine **gravierende Schwäche**: Falls Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegen: Summe der Abstände etwa null
- Dabei passt die Gerade gar nicht gut zu den Datenpunkten!

Methode der kleinsten Quadrate

- Eine andere Möglichkeit besteht darin, die Quadrate der Abweichungen aufzusummieren, also

$$r_1^2 + r_2^2 + \cdots + r_n^2 = \sum_i r_i^2$$

- Parameter a und b so wählen, dass diese Summe minimal wird

- Python berechnet für Beispiel die Werte $a = 6.04$ und $b = 0.047$
  - Grundkosten des Verlags sind also rund 6 SFr. (Preis des Buches für 0 Seiten)
  - Pro Seite verlangt der Verlag rund 5 Rappen
  - Geradengleichung:

$$y = 6.04 + 0.04673x$$

Bestimmung der Parameter a und b

- *Frage:* Wie berechnet der Computer die Parameter a und b?

- Die Parameter a, b minimieren (Methode der Kleinsten-Quadrate)

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Die Lösung dieses Optimierungsproblem ergibt:

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

$$a = \overline{y} - b\overline{x}$$

wobei $\overline{x}$ und $\overline{y}$ die Mittelwerte der jeweiligen Daten

- Diese Gerade $y = a + bx$ wird auch *Regressionsgerade* genannt

Lineare Regression mit Python

- Code:

```
b, a = np.polyfit(book["pages"], book["price"], deg=1)
print(a, b)
```

```
## 6.039999999999996 0.04672727272727273
```

- Befehl

```
np.polyfit(book["pages"], book["price"], deg=1)
```

  passt ein Polynom vom Grad 1 (lineare Funktion) an Daten an

- Ausgabe von 2 Werten: der erste ist die Steigung der Geraden, der zweite der y–Achsenabschnitt

- Python findet also $a = 6.04$ und $b = 0.0467$

Plotten der Regressionsgerade

- Diese Gerade wird in Python wie folgt gezeichnet:

```
book.plot(kind="scatter", x="pages", y="price")
b, a = np.polyfit(book["pages"], book["price"], deg=1)

x = np.linspace(book["pages"].min() ,book["pages"].max())

plt.plot(x, a+b*x, c="orange")

plt.xlabel("Seitenzahl")
plt.ylabel("Buchpreis")

plt.show()
```

- Der Befehl

```
x = np.linspace(book["pages"].min(), book["pages"].max())
```

erzeugt einen Vektor `x` der Länge 50, der als 1. Wert den Minimalwert von pages im Dataframe `book` hat und als letzten Wert dessen Maximalwert.

Beispiel: Buchpreis

- Mit diesem Modell: Preis für Bücher mit Seitenzahlen berechnen, die in der Tabelle nicht vorkommen
- Wieviel würde nach diesem Modell ein Buch von 375 Seiten kosten?
- $x = 375$ in die Geradengleichung oben einsetzen:

$$y = 6.04 + 0.04673 \cdot 375 \approx 23.60$$

- Das Buch dürfte also etwa CHF 23.60 kosten
- Dieses Modell ist allerdings nur begrenzt gültig
- Vor allem bei *Extrapolationen* muss man vorsichtig sein
- Möglich: Was kostet ein Buch mit einer Million Seiten?
- Oder ein Buch mit -100 Seiten? $\rightarrow$ Nicht realistisch!

Beispiel: Körpergrösse Vater-Sohn

- Vermutung: Zusammenhang zwischen der Körpergrösse der Väter und der Grösse der Söhne
- Der britische Statistiker Karl Pearson trug dazu um 1900 die Körpergrösse von 10 (in Wahrheit waren 1078) zufällig ausgewählten Männern gegen die Grösse ihrer Väter auf

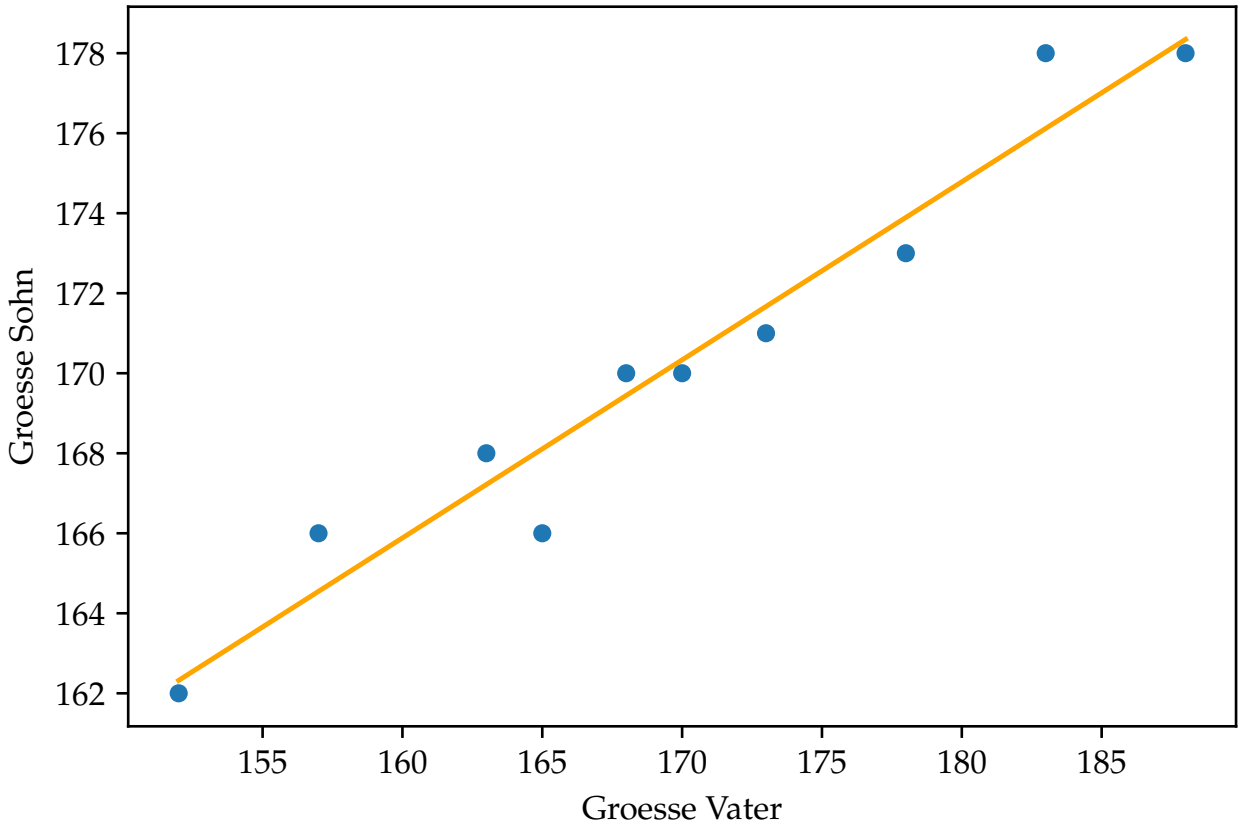
Grösse des Vaters	152	157	163	165	168	170	173	178	183	188
Grösse des Sohnes	162	166	168	166	170	170	171	173	178	178

- Es *scheint* einen Zusammenhang zu geben: Je grösser der Vater, desto grösser der Sohn
- Streudiagramm: Möglicher linearer Zusammenhang besteht

- Die Punktwolke „folgt" der Geraden

$$y = 0.445x + 94.7$$

(mit der Methode der Kleinsten Quadrate aus den Daten)



- Möglich: In Tabelle nicht vorkommende Grösse von 180 cm des Vater, den zu erwartenden Wert für die Grösse seines Sohnes berechnen:

$$y = 0.445 \cdot 180 + 94.7 \approx 175\,\text{cm}$$

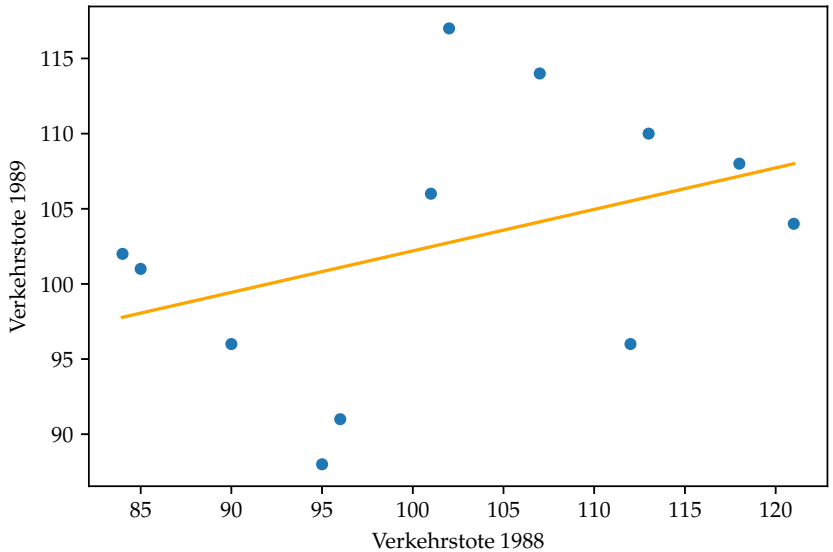
- Achtung: Formel nicht dort anwenden, wo man es *nicht* darf
- Für $x = 0$ erhält man einen Wert von 94.7
- Was heisst dies aber? Wenn der Vater 0 cm gross ist, so ist der Sohn ungefähr 95 cm gross $\rightarrow$ Macht keine Sinn!

Beispiel: Autounfälle

- Tabelle stellt einen Zusammenhang zwischen den Zahlen der Verkehrstoten her, die es 1988 und 1989 in zwölf Bezirken in den USA geben hat

Bezirk	1	2	3	4	5	6	7	8	9	10	11	12
Verkehrstote 1988	121	96	85	113	102	118	90	84	107	112	95	101
Verkehrstote 1989	104	91	101	110	117	108	96	102	114	96	88	106

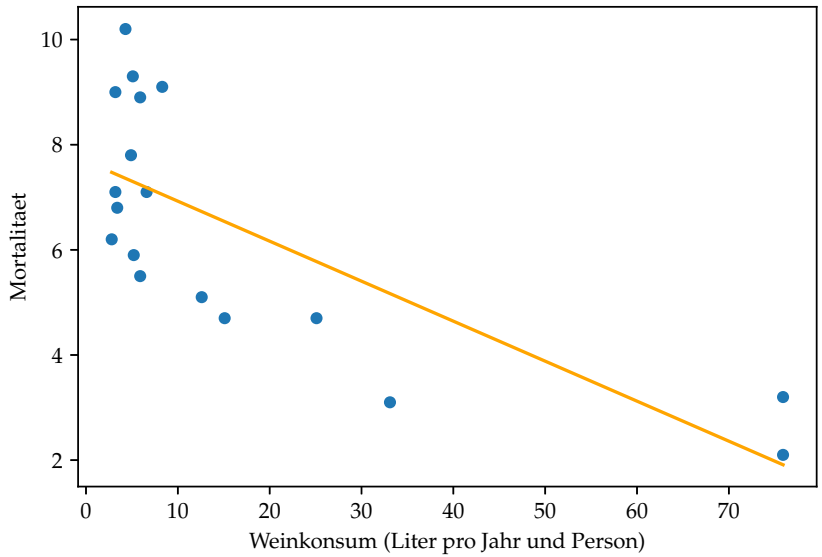
- Es besteht kein offensichtlicher Zusammenhang
- Streudiagramm: kein offensichtlicher Zusammenhang



- Zu erwarten, da es zwischen den Verkehrstoten der einzelnen Bezirke keinen Zusammenhang gibt
- In Abbildung ist noch die Regressionsgerade eingezeichnet
- Können sie zwar berechnen/einzeichnen, *aber diese macht hier gar keinen Sinn*
- *Immer* Berechnung und Plot vergleichen

Beispiel: Weinkonsum

- Schon gesehen: Sterblichkeit vs. Weinkonsum



- Regressionsgerade

$$y = 7.68655 - 0.07608x$$

- Zusammenhang der Daten nicht linear ist (folgt eher einer Hyperbel)
- Die Regressionsgerade sagt hier wenig über den wahren Zusammenhang aus

Wie gut passt die Regressionsgerade?

- Regressionsgerade kann (fast) immer bestimmt werden
- Letzten beiden Beispiele: Regressionsgerade sagt sehr wenig über die wirkliche Verteilung der Punkte im Streudiagramm aus
- Dafür gibt es zwei Gründe
  - Punkte folgen scheinbar gar keiner Gesetzmässigkeit
  - Punkte folgen einer nichtlinearen Gesetzmässigkeit
- Wie kann man feststellen, ob ein linearer Zusammenhang der Daten besteht oder nicht?
- Möglichkeit: Situation graphisch betrachten
- Wert angeben, der den Zusammenhang numerisch beschreibt

Empirische Korrelation

Numerischer Wert der linearen Abhängigkeit von zwei Grössen:

Empirische Korrelation

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \overline{y})^2)}}$$

- Empirische Korrelation ist dimensionslose Zahl zwischen -1 und $+1$
- Misst Stärke und Richtung der *linearen Abhängigkeit* zwischen den Daten x und y
- $r = +1$: Punkte liegen auf steigender Geraden : $y = a + bx$ mit $a \in \mathbb{R}$ und ein $b > 0$
- $r = -1$: Punkte liegen auf fallender Geraden : $y = a + bx$ mit $a \in \mathbb{R}$ und ein $b < 0$
- Sind x und y unabhängig (d.h. kein Zusammenhang), so ist $r = 0$

Berechnung von Korrelation mit Python

- Seitenzahl-Preis-Beispiel mit Python

```
book.corr().iloc[0,1]

## 0.9681121878410434
```

- Wert sehr nahe bei 1 $\rightarrow$ starker linearer Zusammenhang
- Wert positiv $\rightarrow$ „je mehr, desto mehr" Zusammenhang
- Der Befehl

```
book.corr()
```

allgemeiner $\rightarrow$ Korrelationsmatrix

Empirische Korrelation: Beispiele

- Beispiel der Körpergrösse von Vater und Sohn: Erwarten hohen Korrelationskoeffizienten, da Daten nahe der Regressionsgerade

  $\rightarrow$ 0.973

- Verkehrsunfällen: Kein Zusammenhang $\rightarrow$ Tiefer Korrelationskoeffizienten

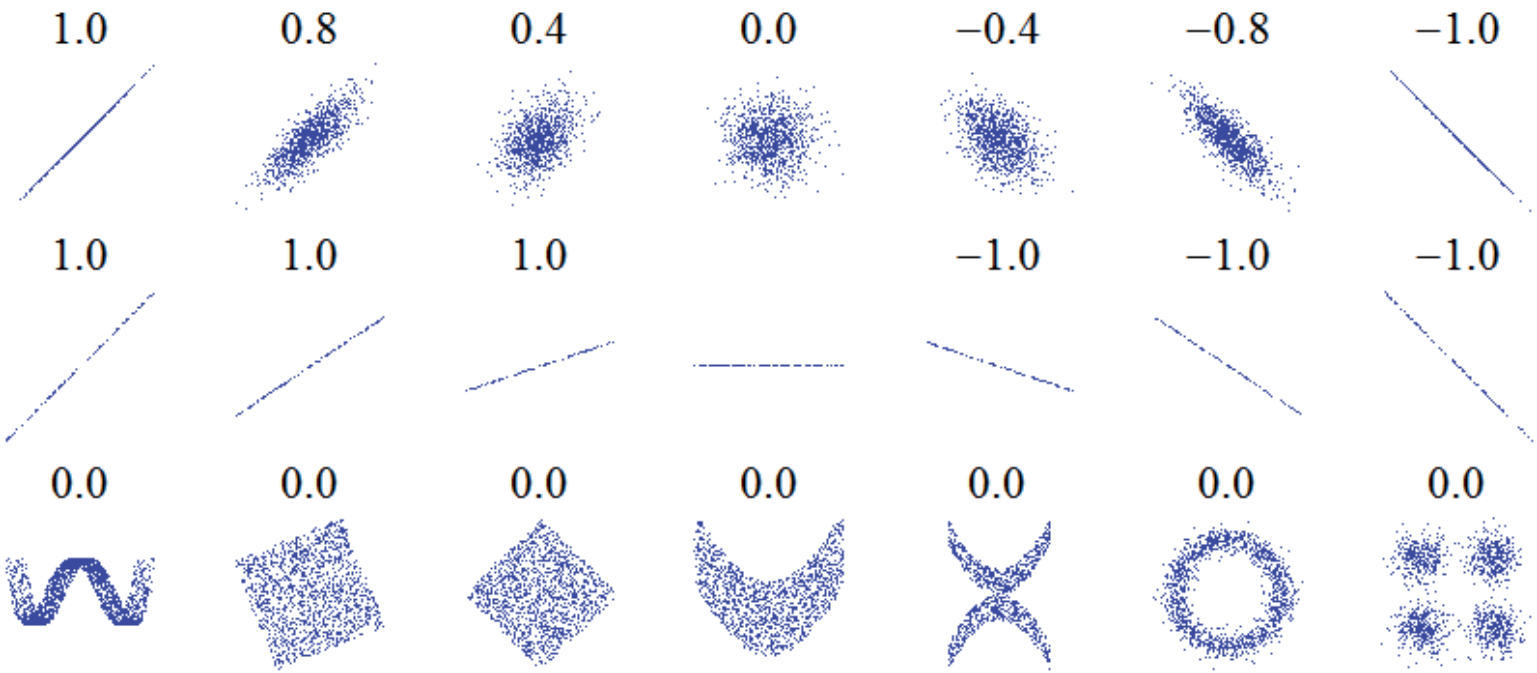
  $\rightarrow$ 0.386

- Weinkonsum: Keinen allzu grossen Korrelationskoeffizient (keine Gerade), aber er sollte negativ sein, da mit steigendem Weinkonsum die Mortalität sinkt

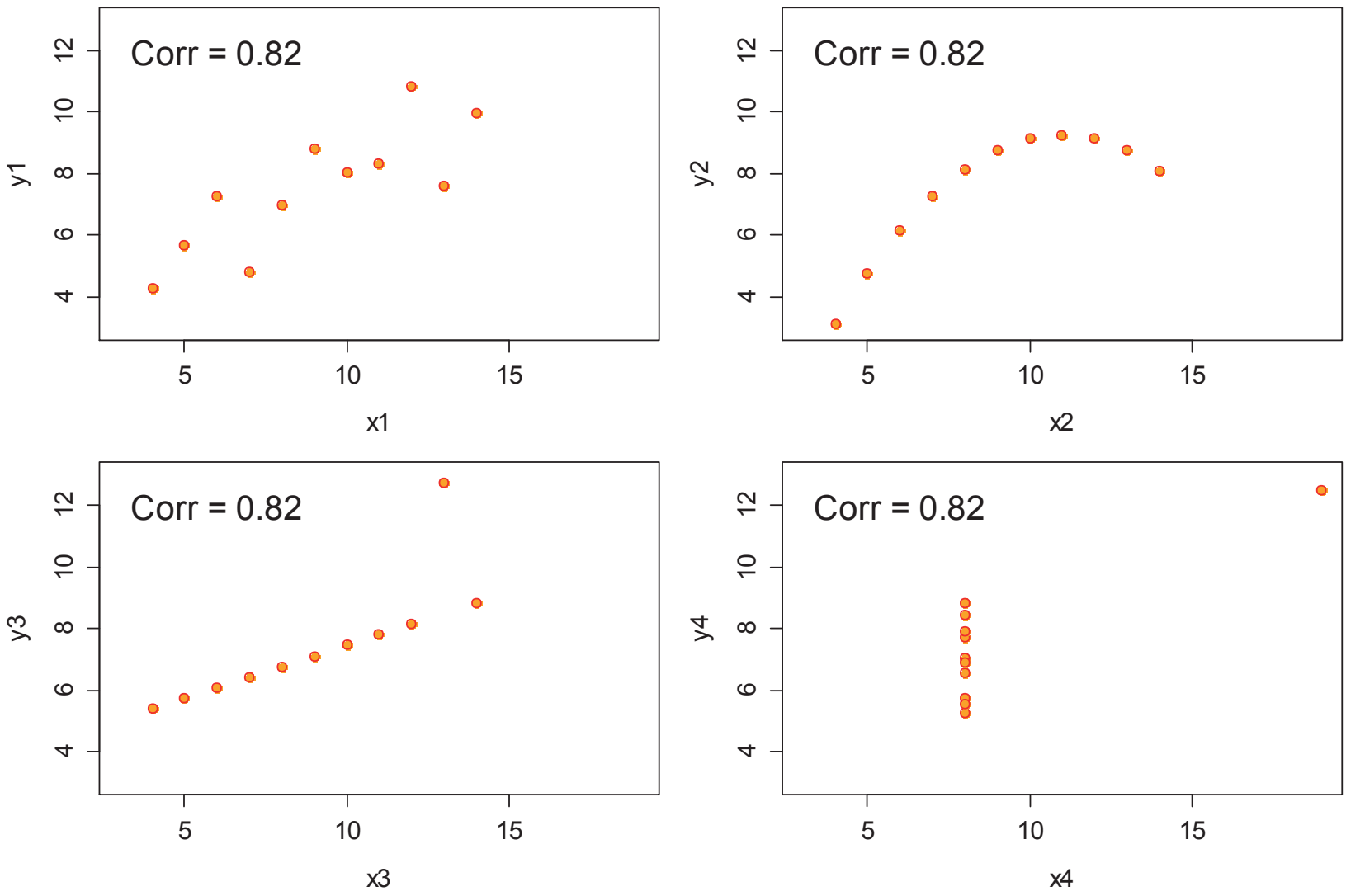
  $\rightarrow$ -0.746.

Empirische Korrelation: Bemerkungen

- Korrelation misst „nur" den *linearen Zusammenhang*
- Man sollte daher die Daten immer auch anschauen, statt sich „blind" auf Kennzahlen zu verlassen



Empirische Korrelation: Bemerkungen



Datenbereinigung

- Problem: Fehlende Daten im Datensatz
- Bevor man aber fehlende Datenpunkte entfernt oder für fehlende Datenpunkte einfach neue (interpolierte) Datenwerte einsetzt (engl. *data imputation*), sollte man verstehen, warum diese Datenwerte fehlen. Wir unterscheiden folgende Fälle:
  - *Missing Completely at Random (MCAR)*
  - *Missing at Random (MAR)*
  - *Missing Not at Random (MNAR)*

Missing Completely at Random (MCAR)

- In diesem Fall ist die Ursache für das Fehlen der Variable völlig unsystematisch.
- Beispiel: Betrachten wir als Beispiel eine Studie, bei welcher der Grund für die Fettleibigkeit bei $K12$-Kindern ermittelt wird
- MCAR bedeutet in diesem Fall, dass die Eltern zum Beispiel vergessen haben, ihre Kinder in die Klinik zur Studie zu bringen.

Missing at Random (MAR)

- In diesem Fall liegt dem Fehlen von Daten eine gewisse Systematik zugrunde
- Beispiel: Bei Verwendung der obengenannten *K*12-Studie sind die fehlenden Daten in diesem Fall zum Beispiel auf den Umzug der Eltern in eine andere Stadt zurückzuführen, weshalb die Kinder die Studie aufgeben mussten - das Fehlen hat nichts mit der Studie zu tun.

Missing Not at Random (MNAR)

- Ein möglicher nichtzufälliger Grund für das Fehlen von Datenwerten ist, dass der fehlende Wert vom hypothetischen Wert abhängt.

- Beispiel: falls Eltern durch die Art der Studie beunruhigt sind und nicht wollen, dass ihre Kinder beispielsweise gemobbt werden, deswegen zogen sie ihre Kinder aus der Studie zurück. Die Schwierigkeit mit MNAR-Daten ist intrinsisch und hat hier mit dem Problem der Identifizierbarkeit der Studienteilnehmer zu tun.