Deskriptive Statistik
Eindimensionale Daten

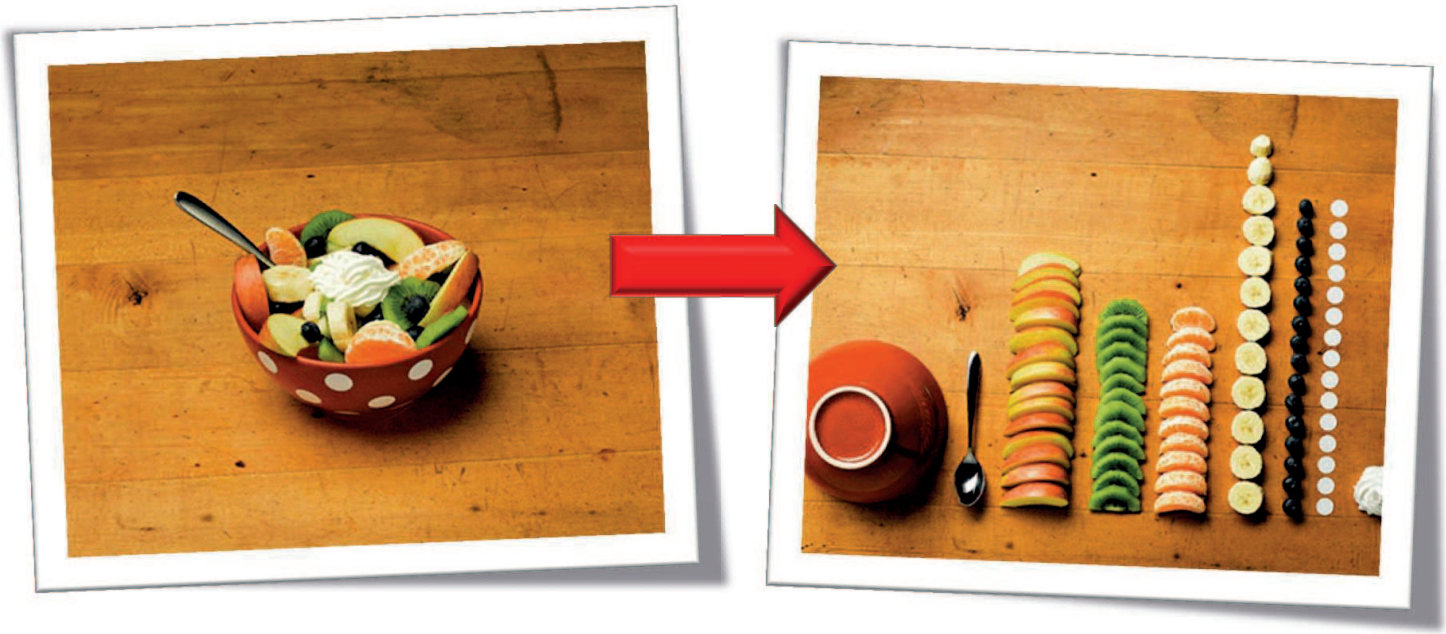
Peter Büchel

HSLU TA

Stat: SW02

Ziele der Deskriptiven Statistik

- Daten *zusammenfassen* durch *numerische Kennwerte*
- *Graphische Darstellung* der Daten



Daten

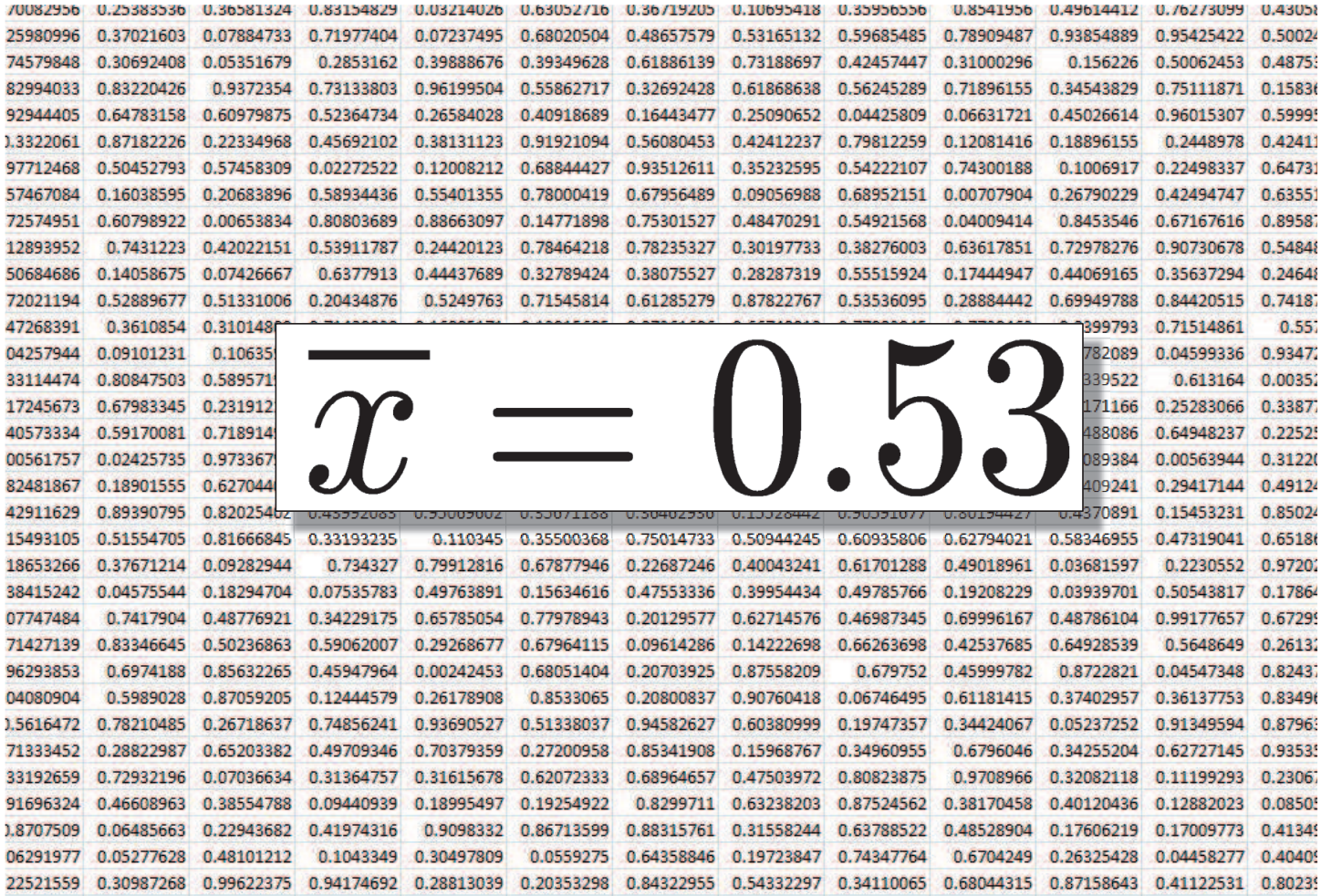
- In diesem Modul $\rightarrow$ meist *reale Daten*

- *Datensatz*
  Wiederholte Messungen: Freigesetzte Wärme beim Übergang von Eis bei $-0.7\,^{\circ}\mathrm{C}$ zu Wasser bei $0\,^{\circ}\mathrm{C}$
  $\rightarrow$ 13 Werte (siehe Skript) (in cal/g) $\rightarrow$ Methode *A*

  $$79.98;\ 80.04;\ 80.02;\ldots\ 80.02;\ 80.00;\ 80.02$$

- Basierend auf den Daten: Diverse *Kennwerte* berechnen bzw. Daten *graphisch darstellen*

- **Warnung**: Wann immer wir einen Datensatz „reduzieren" (durch Kennzahlen oder Graphiken), geht *Information verloren*!


$\overline{x} = 0.53$

Überblick

- Bekannt: n beobachtete Datenpunkte (Messungen)

$$x_1,\ x_2,\ \ldots,\ x_n$$

  (z.B. Verkehrsaufkommen an n verschiedenen Tagen)
- Unterscheidung zwischen Lage- und Streuungsparametern
- *Lageparameter* („Wo liegen die Beobachtungen auf der Mess-Skala?")
  - Arithmetisches Mittel („Durchschnitt")
  - Median
  - Quantile
- *Streuungsparameter* („Wie streuen die Daten um ihre mittlere Lage?")
  - Empirische Varianz / Standardabweichung
  - Quartilsdifferenz

Arithmetisches Mittel

- Definition:

Arithmetisches Mittel

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Umgangsprachlich: *Durchschnitt*

- Beispiel Schmelzwärme: Arithmetische Mittel der $n = 13$ Messungen

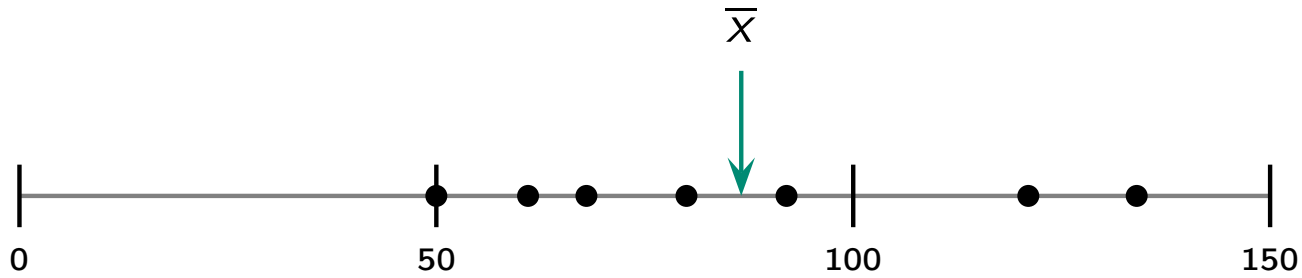
$$\overline{x} = \frac{79.98 + 80.04 + \ldots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.020\,77$$

Arithmetisches Mittel

- Python-Befehl

```
from pandas import Series
import pandas as pd

methodeA = Series([79.98, 80.04, 80.02, 80.04, 80.03,
80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])

methodeA.mean()
## 80.02076923076923
```

- Arithmetische Mittel: Anschaulich



Schwerpunkt der Daten

Streuung

- Arithmetisches Mittel : „Wo ist die „Mitte" der Daten?"
- Aber: Beispiel von (fiktiven) Schulnoten:

$$2;\ 6;\ 3;\ 5 \qquad \text{und} \qquad 4;\ 4;\ 4;\ 4$$

- Beide Mittelwert 4, aber Verteilung der Daten um Mittelwert ziemlich unterschiedlich
  - 1. Fall: Zwei gute und zwei schlechte Schüler
  - 2. Fall: Alle Schüler gleich gut
- Datensätze haben eine verschiedene *Streuung* um den Mittelwert

Streuung numerisch

- 1. Idee: Durchschnitt der *Unterschiede zum Mittelwert*

  1. Fall:

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

  Zweiter Fall auch 0 $\rightarrow$ Keine Aussage

- Problem: Unterschiede können *negativ* werden $\rightarrow$ Können sich aufheben

Streuung

- Nächste Idee: Unterschiede durch die Absolutwerte ersetzen
  1. Fall:

$$\frac{|(2-4)|+|(6-4)|+|(3-4)|+|(5-4)|}{4} = \frac{2+2+1+1}{4} = 1.5$$

- D.h.: Noten weichen im Schnitt 1.5 vom Mittelwert ab
- 2. Fall: Dieser Wert natürlich auch 0
- Je grösser dieser Wert (immer grösser gleich 0) , desto mehr unterscheiden sich die Daten bei gleichem Mittelwert untereinander
- Dieser Wert für die Streuung: *mittlere absolute Abweichung*
- Aber: Theoretische Nachteile

Empirische Varianz und Standardabweichung

- Besser: *Empirische Varianz* und *empirische Standardabweichung* für das Mass der Variabilität oder Streuung der Messwerte verwendet

- Definition:

Empirische Varianz *Var*(x) und Standardabweichung s_X

$$\text{Var}(x) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

und

$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Eigenschaften der Varianz

- Bei Varianz: Abweichungen $x_i - \overline{x}$ quadrieren , damit sich Abweichungen nicht gegenseitig aufheben können
- Nenner $n - 1$, anstelle von n $\rightarrow$ mathematisch begründet
- Standardabweichung ist die Wurzel der Varianz
- Durch Wurzelziehen wieder dieselbe Einheit wie bei Daten selbst
- Ist empirische Varianz (und damit die Standardabweichung) gross, so ist die Streuung der Messwerte um das arithmetische Mittel gross
- *Wert* der empirische Varianz hat keine physikalische Bedeutung
  $\rightarrow$ Man weiss nur, je grösser der Wert umso grösser die Streuung

Beispiele: Schmelzwärme

- Arith. Mittel der $n = 13$ Messungen ist $\overline{x} = 80.02$ (siehe oben)
- Empirische Varianz:

$$\text{Var}(x) = \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \cdots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1}$$
$$= 0.000\,574$$

- Empirische Standardabweichung:

$$s_x = \sqrt{0.000574} = 0.024$$

- D.h.: „mittlere" Abweichung vom Mittelwert 80.02 cal/g ist 0.024 cal/g

Beispiele: Schmelzwärme

- Von Hand sehr mühsam. Mit `pandas`-Methoden:
  Varianz:

```
methodeA.var()
## 0.0005743589743590099
```

  Standardabweichung:

```
methodeA.std()
## 0.023965787580611863
```

Median

- Ein weiteres Lagemass für die „Mitte" $\rightarrow$ *Median*
- Sehr vereinfacht gesagt: Wert, bei dem die Hälfte der Messwerte unter diesem Wert liegen
- Beispiel: Prüfung in der Schule ist Median 4.6
- D.h.: Hälfte der Klasse liegt unter dieser Note
- Umgekehrt liegen die Noten der anderen Hälfte *über* dieser Note
- Obige Interpretation des Medians ist sehr vereinfacht dargestellt. Die exakte Definition folgt nun.

Geordnete Strichprobe

- Datensatz in aufsteigender Reihenfolge *ordnen*
- Bezeichnung der *geordneten Daten* mit $x_{(i)}$:

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

- Beispiel: $x_1 = 3, x_2 = 7, x_3 = 2$:

$$x_{(1)} = x_3 = 2, \qquad x_{(2)} = x_1 = 3, \qquad x_{(3)} = x_2 = 7$$

Median

- Bestimmung *Median*: Daten zuerst der Grösse nach ordnen:

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

- Für die Daten der Methode A heisst dies

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

- Median ist nun sehr einfach zu bestimmen
- Unter diesen 13 Messungen: Wert der mittleren Beobachtung
- Dies ist in diesem Fall der Wert der 7. Beobachtung:

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

Median

- Median des Datensatzes der Methode A ist 80.03
- D.h.: Knapp die Hälfte der Messwerte, nämlich 6 Beobachtungen sind kleiner oder gleich 80.03
- Ebenso sind 6 Messwerte grösser oder gleich dem Median
- Ungerade Anzahl Messungen $\rightarrow$ *Genau* eine mittlere Messung

Median

- Vorher: Anzahl der Daten ungerade und damit ist die mittlere Beobachtung eindeutig bestimmt
- Anzahl Daten gerade $\rightarrow$ *Keine* mittlere Beobachtung
- *Definition* Median: Mittelwert der beiden mittleren Beobachtungen
- Beispiel: Datensatz der Methode B hat 8 Beobachtungen
- Ordnen den Datensatz: Median Durchschnitt von der 4. und 5. Beobachtung

$$79.94;\ 79.95;\ 79.97;\ \underbrace{79.97;\ 79.97}\ ;\ 79.94;\ 80.02;\ 80.03$$

$$\frac{79.97 + 79.97}{2} = 79.97$$

Median

- Mit `pandas` erhalten wir für die Methode A

```
methodeA.median()
## 80.03
```

und für die Methode B

```
methodeB = Series([80.02, 79.94, 79.98, 79.97, 79.97,
80.03, 79.95, 79.97])

methodeB.median()
## 79.97
```

- Als Median kann Wert auftreten, der in Messreihe nicht vorkommt
- Annahme: Mittlere Beobachtungen der Methode B sind Werte 79.97 und 79.98:

$$\frac{79.97 + 79.98}{2} = 79.975$$

Median vs. arithmetisches Mittel

- Zwei Lagemasse für die Mitte eines Datensatzes
- Welches ist nun „besser”?
- Dies kann man so nicht sagen, das kommt auf die jeweilige Problemstellung an. Am besten werden beide Masse gleichzeitig verwendet.
- Eigenschaft des Medians: *Robustheit*
- Das heisst: Wird viel weniger stark durch extreme Beobachtungen beeinflusst als das arithmetisches Mittel

Median vs. arithmetisches Mittel

- Beispiel: Bei der grössten Beobachtung ($x_9 = 80.05$) ist ein Tippfehler passiert und $x_9 = 800.5$ eingegeben worden
- Das arithmetische Mittel ist dann

$$\overline{x} = 135.44$$

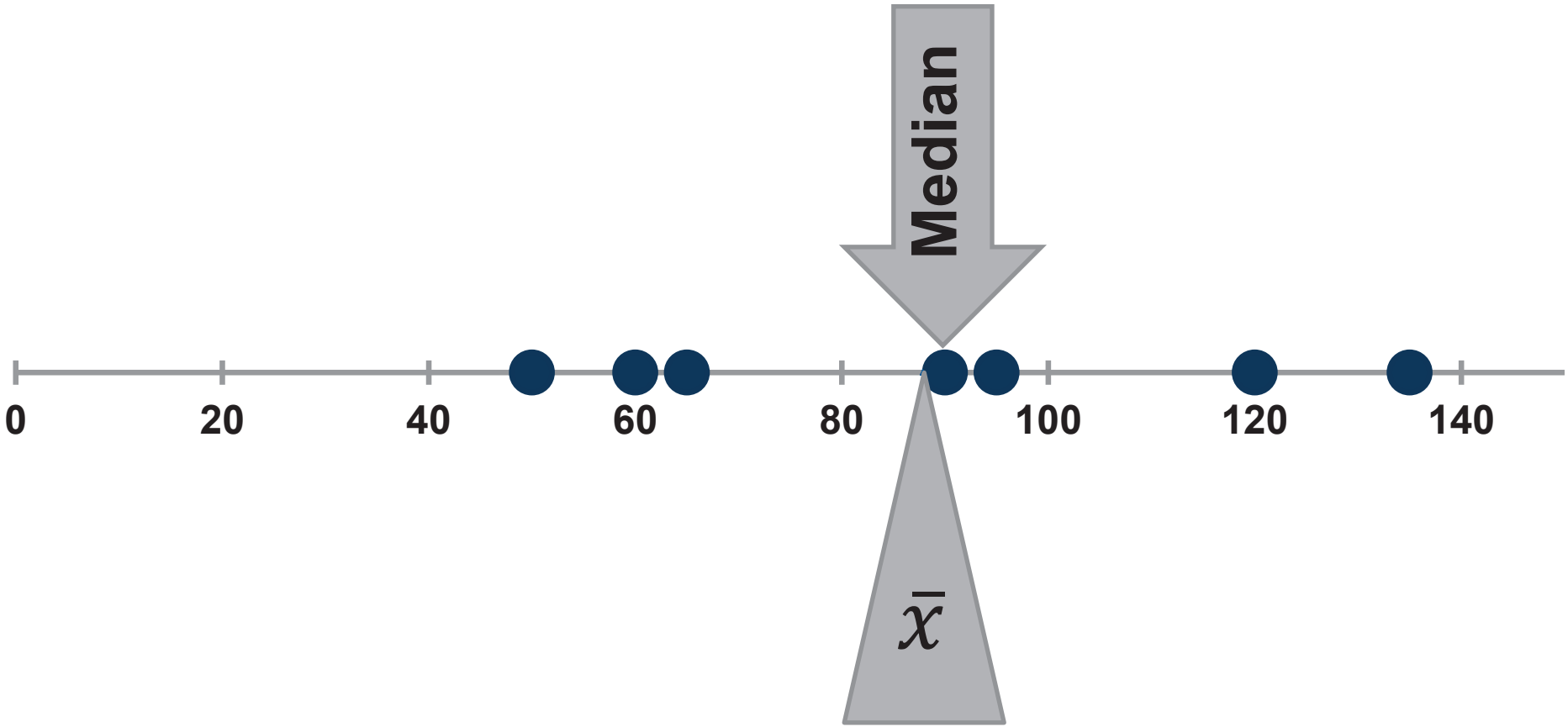
- Der Median ist aber nach wie vor

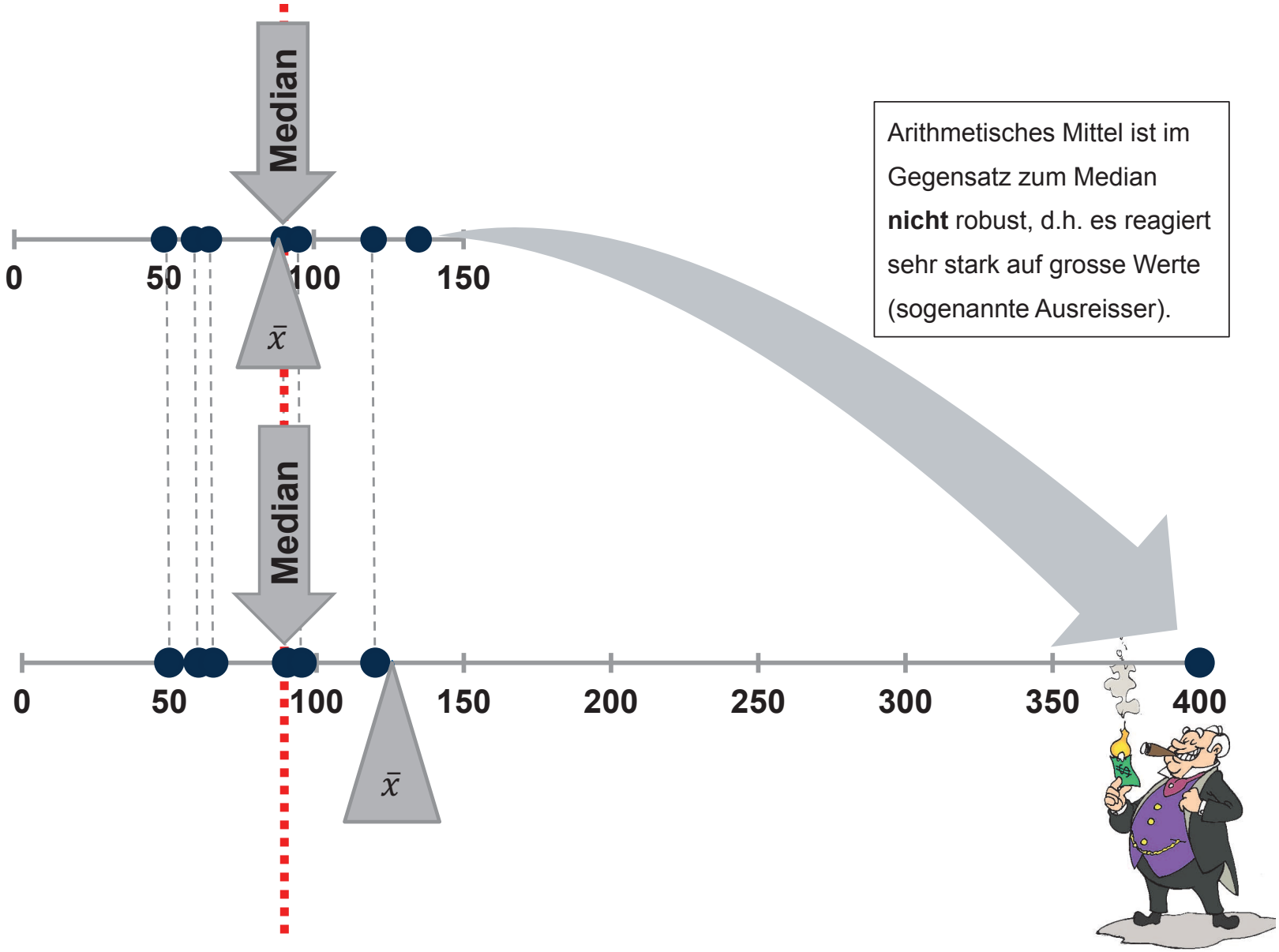
$$x_{(7)} = 80.03$$

- Arithmetisches Mittel: Durch Veränderung einer Beobachtung sehr stark beeinflusst
- Median hier gleich bleibt $\rightarrow$ robust

Arith. Mittel vs. Median: Einkommen [k CHF]

7 Beobachtungen





Arithmetisches Mittel ist im Gegensatz zum Median **nicht** robust, d.h. es reagiert sehr stark auf grosse Werte (sogenannte Ausreisser).

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

Quartile

- Median: Wert, wo die Hälfte der Beobachtungen kleiner (oder gleich) wie dieser Wert sind
- Analoge Überlegung: Unteres und oberes Quartil
- Unteres Quartil: Wert, wo 25 % aller Beobachtungen kleiner oder gleich und 75 % grösser oder gleich sind wie dieser Wert
- Oberes Quartil: Wert, wo 75 % aller Beobachtungen kleiner oder gleich und 25 % grösser oder gleich wie dieser Wert sind
- Achtung: Meist gibt es nicht *exakt* 25 % der Beobachtungen
- Man *definiert* Wert für das untere Quartil bzw. obere Quartil

Beispiel: Schmelzwärme

- Methode A hat $n = 13$ Messpunkte $\quad \rightarrow \quad$ 25 % davon ist 3.25
- Man *wählt* nächstgrösseren Wert $x_{(4)}$ als unteres Quartil:

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

- Unteres Quartil ist 80.02
- Knapp ein Viertel der Messwerte ist gleich oder kleiner 80.02
- Oberes Quartil: *Wählen* $x_{(10)}$, da für $0.75 \cdot 13 = 9.75$ die Zahl 10 der nächsthöhere Wert ist

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

- Knapp drei Viertel der Messwerte sind kleiner oder gleich 80.04

- Methode B: 25 % der Werte 2 $\quad \rightarrow \quad$ ganze Zahl
- *Wählen* dann *nächste* Beobachtung $x_{(3)}$ als unteres Quartil

$$79.94; 79.95; \underbrace{79.97}; 79.97; 79.97; 79.94; 80.02; 80.03$$

- Unteres Quartil der Methode B ist also 79.97

Bemerkungen

- Hier jeweils aufgerundet, falls 25 % bzw. 75 % der Anzahl Beobachtungen nicht ganz ist
- Hätten auch *abrunden* können $\rightarrow$ Andere Werte für die Quartile
- Für grosse Datensätze: Spielt praktisch keine Rolle, ob auf- oder abgerundet oder gerundet wird
- Aber: Es gibt keine einheitliche Definition für die Quartile

pandas

- `pandas` kennt keine eigenen Befehle für die Quartile
- Allgemeinerer Befehl `quantile` (Quantile kommen gleich)
- `Python`: Quartile nach unserer Definition Option `interpolation="lower"`
- Für das untere Quartil der Methode *A* lautet der Befehl

```
methodeA.quantile(q=.25, interpolation="lower")
## 80.02
```

und für das obere

```
methodeA.quantile(q=.75, interpolation="lower")
## 80.04
```

pandas

- Methode *B*:

```
methodeB.quantile(q=.25, interpolation="lower")
## 79.95
```

- Dies entspricht *nicht* unserem oben bestimmten Wert 79.97
- Python führt noch Korrekturfaktor zur Berechnung der Quantile ein
- Dieser ist aber für grosse Datensätze irrelevant

Quartilsdifferenz

- *Quartilsdifferenz* ist ein Streuungsmass für die Daten

$$\text{oberes Quartil} - \text{unteres Quartil}$$

- Es misst die Länge des Intervalls, das etwa die Hälfte der „mittleren" Beobachtungen enthält
- Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte um den Median und umso kleiner ist die Streuung
- Dieses Streuungsmass ist robust
- Quartilsdifferenz der Methode A

$$80.04 - 80.02 = 0.02$$

pandas

- Mit pandas

```
q75, q25 = methodeA.quantile(q = [.75, .25],
interpolation="lower")

iqr = q75 - q25
iqr
## 0.020000000000010232
```

- Die (oder ungefähr die) Hälfte der Messwerte liegt also in einem Bereich der Länge 0.02

Quantile

- Quartile auf jede andere Prozentzahl verallgemeinern: Quantile
- 10 %-Quantil: Wert, wo 10 % der Werte kleiner oder gleich und 90 % der Werte grösser oder gleich diesem Wert sind
- *Empirische α-Quantil*: Wert, wo $\alpha \times 100\,\%$ der Datenpunkte kleiner oder gleich und $(1-\alpha) \times 100\,\%$ der Punkte grösser oder gleich sind
- Definition

Empirische α-Quantile ($0 < \alpha < 1$)

$x_{(\alpha n+1)}$, falls $\alpha \cdot n$ eine natürliche Zahl ist

$x_{(k)}$, wobei k die Zahl $\alpha \cdot n$ aufgerundet ist, falls $\alpha \cdot n \notin \mathbb{N}$

- Wie bei den Quartilen:
  - Aufgerunden, falls die entsprechende Prozentzahl der Beobachtungen nicht ganz;
  - sonst nächsthöheren Wert
- Empirischer Median ist empirisches 50 %–Quantil
- Empirisches 25 %–Quantil ist unteres Quartil
- Empirisches 75 %–Quantil ist oberes Quartil

- `pandas`: 10 %– und 70 %–Quantil der Methode A:

```
methodeA.quantile(q=.1, interpolation="lower")
methodeA.quantile(q=.7, interpolation="lower")
```

```
## 79.98
## 80.03
```

- Knapp 10 % der Messwerte sind kleiner oder gleich 79.98
- Entsprechend: Knapp 70 % der Messwerte kleiner oder gleich 80.03

Beispiel

- Noten an Prüfung in Schulklasse mit 24 SchülerInnen:

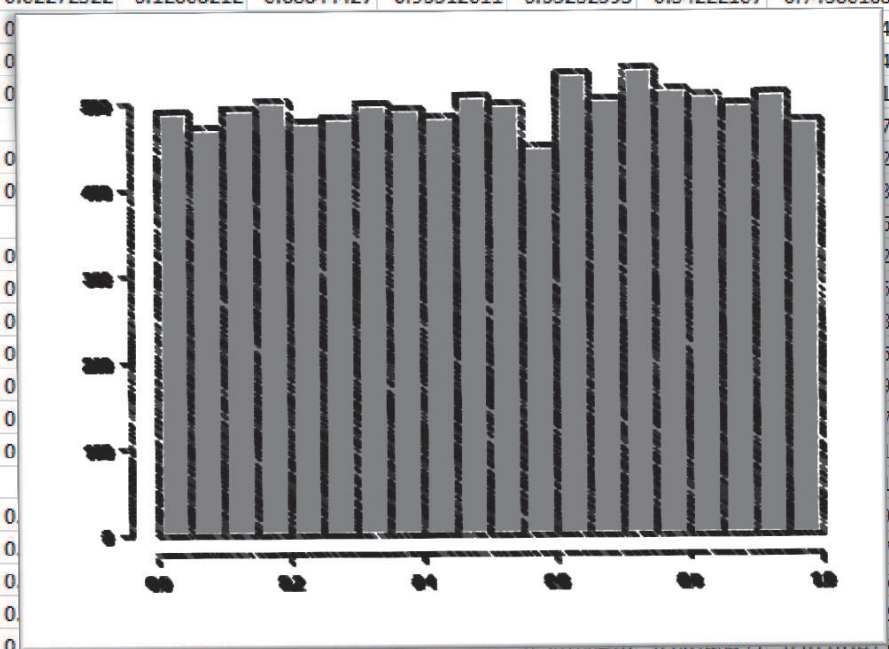
  4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

- Verschiedene Quantile mit `pandas`:

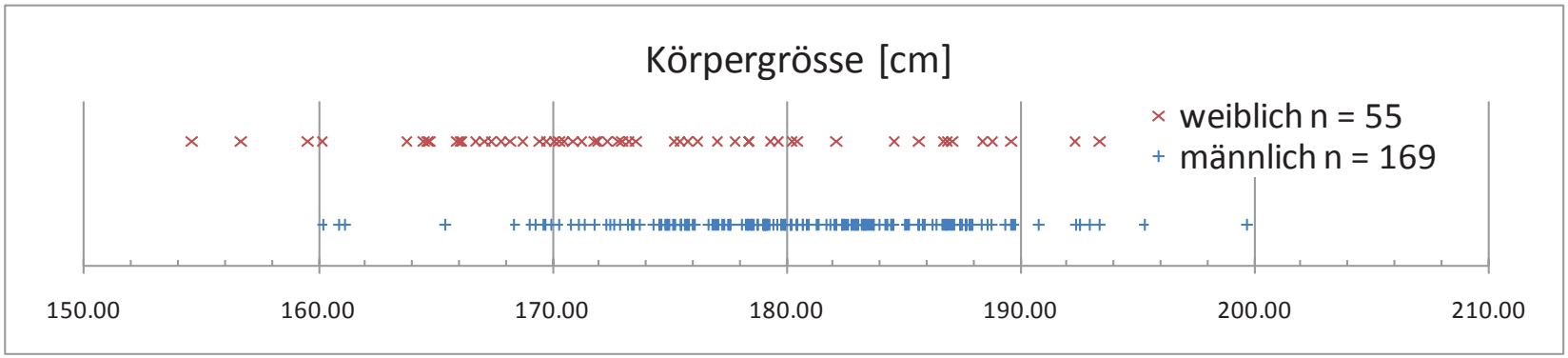
```
noten = Series([4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4,
3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1])

noten.quantile(q = np.linspace(.2,1,5), interpolation="lower")
## 0.2    3.6
## 0.4    4.2
## 0.6    4.9
## 0.8    5.5
## 1.0    6.0
## dtype: float64
```

- D.h.: Knapp 20 % der SuS sind schlechter als 3.6
- 20 % SchülerInnen nicht möglich, da dies 4.8 SuS wären
- 60 %–Quantil: (Knapp) diese Anzahl Prozent der SuS waren schlechter oder gleich einer 4.9

Graphische Darstellungen



Eindimensionales Streudiagramm

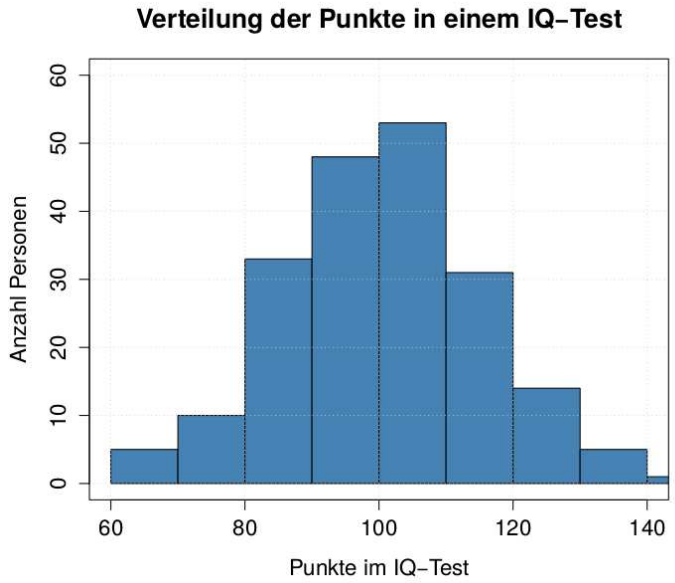


- Guter Überblick, falls nicht zu viele Daten vorhanden sind
- Achtung bei diskret verteilten Daten (Punkte liegen aufeinander!)

Histogramm

- *Histogramm*: Graphischer Überblick über die auftretenden Werte
- Aufteilung des Wertebereichs in k *Klassen* (Intervalle)
- Faustregel:
  - bei weniger als 50 Messungen ist die Klassenzahl 5 bis 7
  - bei mehr als 250 Messungen wählt man 10 bis 20 Klassen
- Zeichne für jede Klasse einen *Balken*, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist

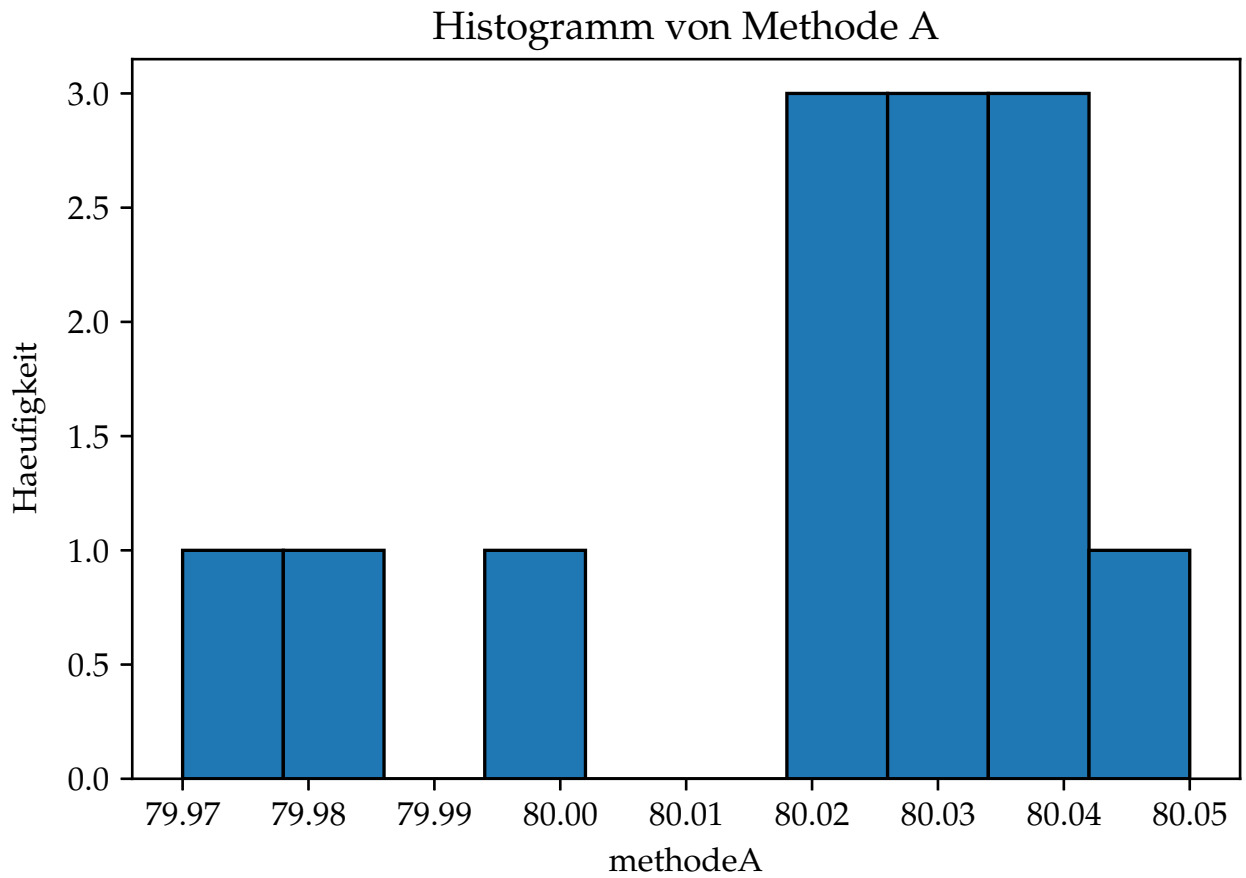
Beispiel: IQ-Test



- Histogramm von IQ-Test Ergebnis von 200 Personen
- Breite der Klassen: 10 IQ-Punkte; für jede Klasse gleich
- Höhe der Balken: Anzahl Personen, die in diese Klasse fallen
- Beispiel: ca. 14 Personen fallen in die Klasse zwischen 120 - 130 IQ-Punkten

Python

Für die Methode *A* sieht das Histogramm wie folgt aus:



Python

Es wurde mit folgendem Code erzeugt:

```
import pandas as pd
from pandas import DataFrame, Series
import matplotlib.pyplot as plt

methodeA = Series([79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])

methodeB = Series([80.02, 79.94, 79.98, 79.97, 79.97, 80.03,
79.95, 79.97])

methodeA.plot(kind="hist", edgecolor="black")

plt.title("Histogramm von Methode A")
plt.xlabel("methodeA")
plt.ylabel("Haeufigkeit")

plt.show()
```

Bemerkungen

- Methode A 13 Messungen $\rightarrow$ 10 Balken (`pandas`-default)
- Bedeutung der Anzahlen (Frequency):
  - 10 Klassen mit Werten im Bereich $[79.97, 80.05]$ $\rightarrow$ Balkenbreite

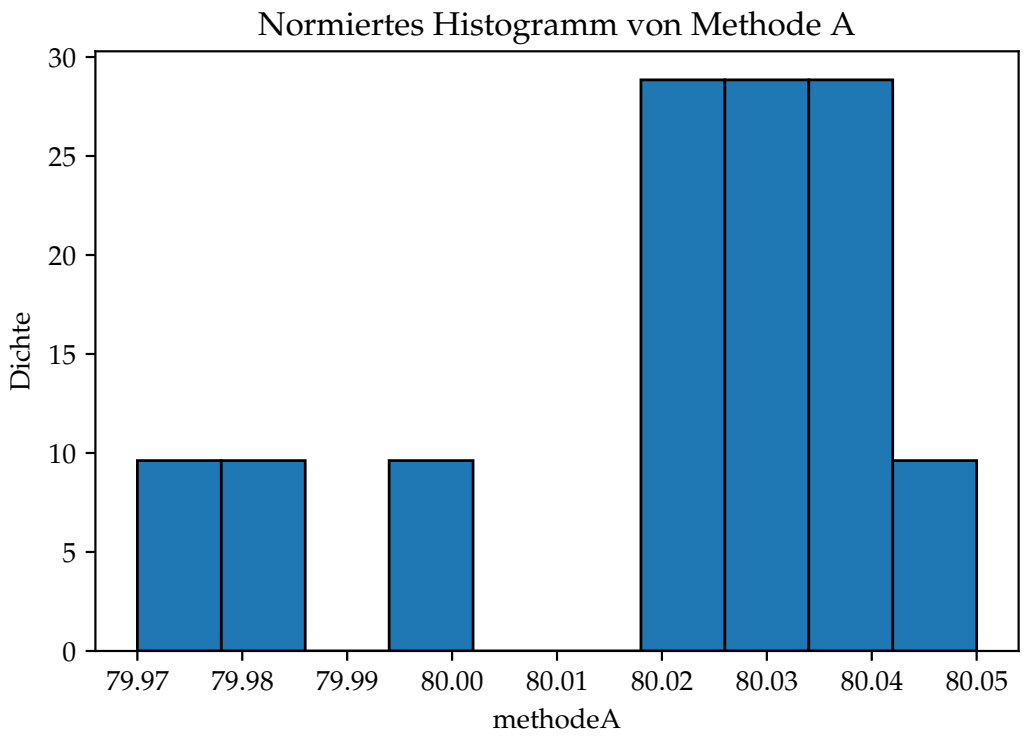
$$\frac{80.05 - 79.97}{10} = 0.008$$

  - 1. Klasse 79.97-79.978: Anzahl Beobachtungen 79.97 berücksichtigt
  - 2. Klasse Werte 79.98; usw.
- `pandas` selbst keine Graphiken $\rightarrow$ Bibliothek `matplotlib`
- Pandas-Attribut `plot` für Plots $\rightarrow$ Option `kind="hist"`
- Mit dem `Python` -Befehl lassen sich auch die Anzahl Klassen festlegen, Überschriften ändern, usw. (siehe Übungen)

Histogramm: Dichte

- Histogramm oben: Höhe der Balken entspricht Anzahl Beobachtungen in Klasse
- Andere Form des Histogramms

```
methodeA.plot(kind="hist", normed=True,
edgecolor="black")
```



- *Gesamtfläche* der Balken ist 1
- Auf der vertikalen Achse sind nun die *Dichten* angegeben
- Herauslesen: über

$$(80.018 - 80.026) \cdot 28.846 = 0.23$$

  also etwa 23 % der Daten zwischen 80.018 und 80.026 befinden
- Balkenhöhe: Anzahl Beobachtungen in einem Balken mit $\frac{1}{n}$ multiplizieren, diese Zahl durch die Balkenbreite dividiert
- Unser Beispiel: 3 Beobachtungen im Intervall $[80.018, 80.026]$
- Balkenhöhe:

$$\frac{\frac{1}{13} \cdot 3}{0.008} = 28.8462$$

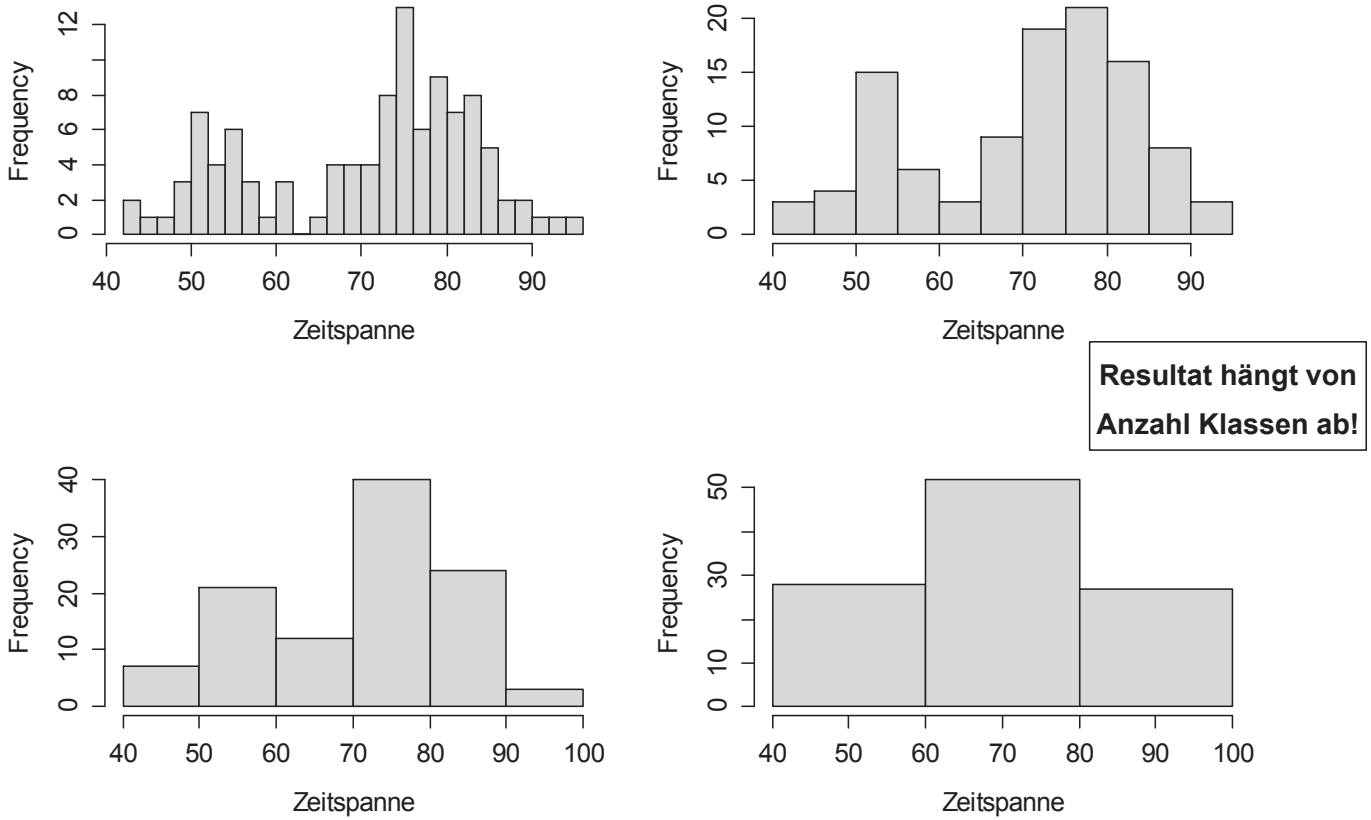
- Vorteil: Messungen mit unterschiedlichen Umfängen besser miteinander vergleichbar

Old Faithful Geysir (Yellowstone NP): Daten

- *Zeitspanne* [min] zwischen Ausbrüchen
- *Eruptionsdauer* [min]
- Daten finden Sie auf ILIAS

	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	

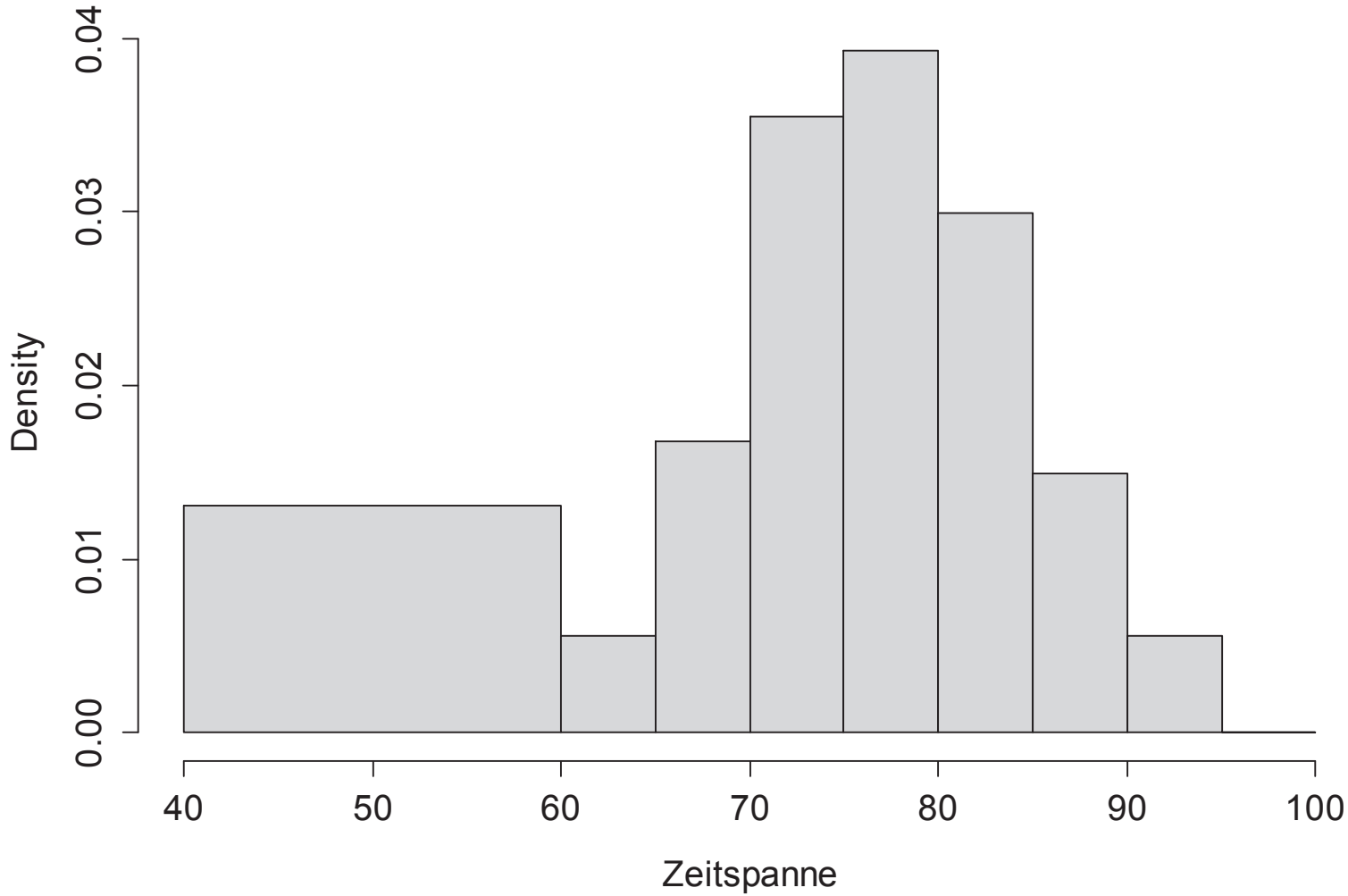
Histogramme der Zeitspanne (verschiedene Anzahl Klassen)



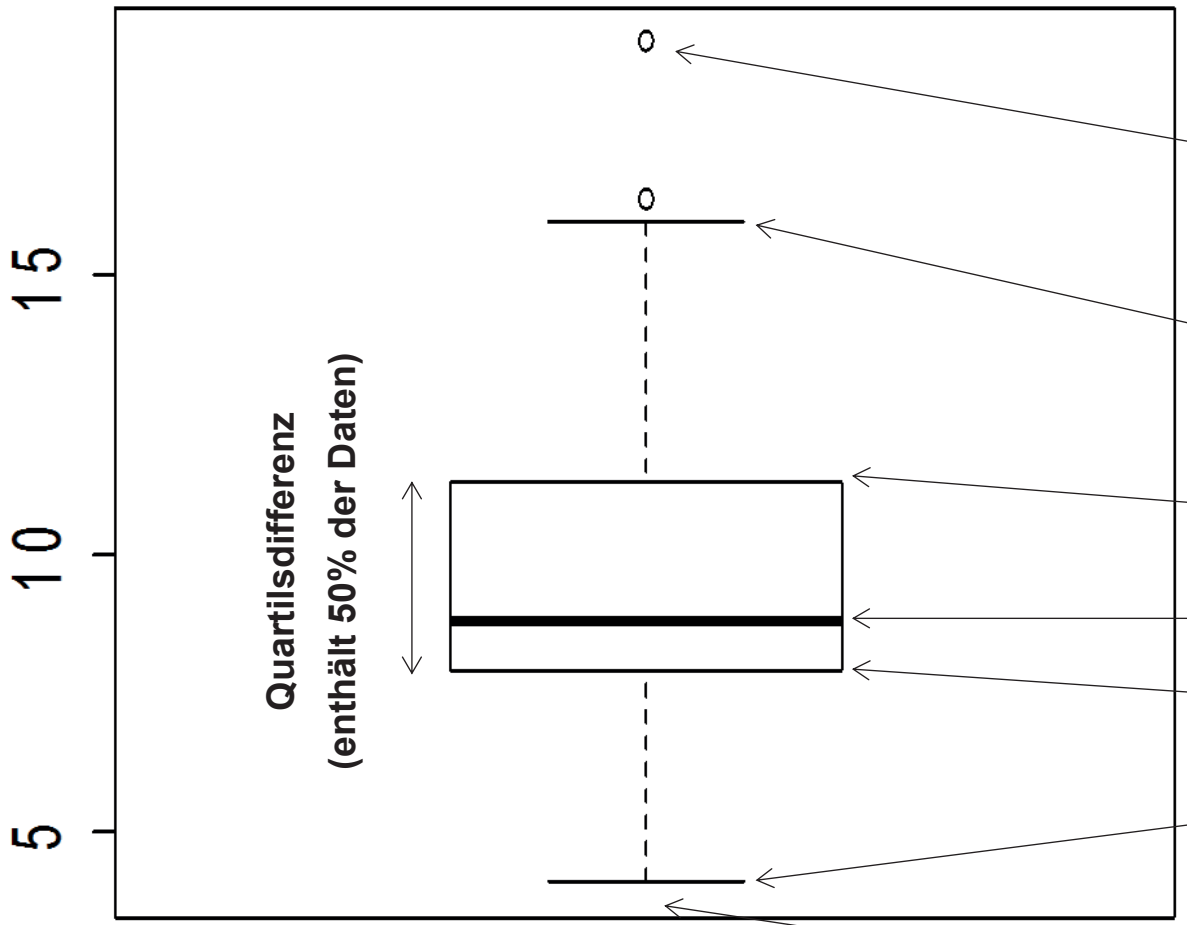
Interaktiv Klassen verändern (bei anderen Daten):

`http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html`

Histogramm der Zeitspanne mit unterschiedlicher Intervallbreite



Boxplot: Schematischer Aufbau



Ausreisser
(falls vorhanden)

Grösste «normale» Beobachtung

Oberes Quartil

Median

Unteres Quartil

Kleinste «normale» Beobachtung

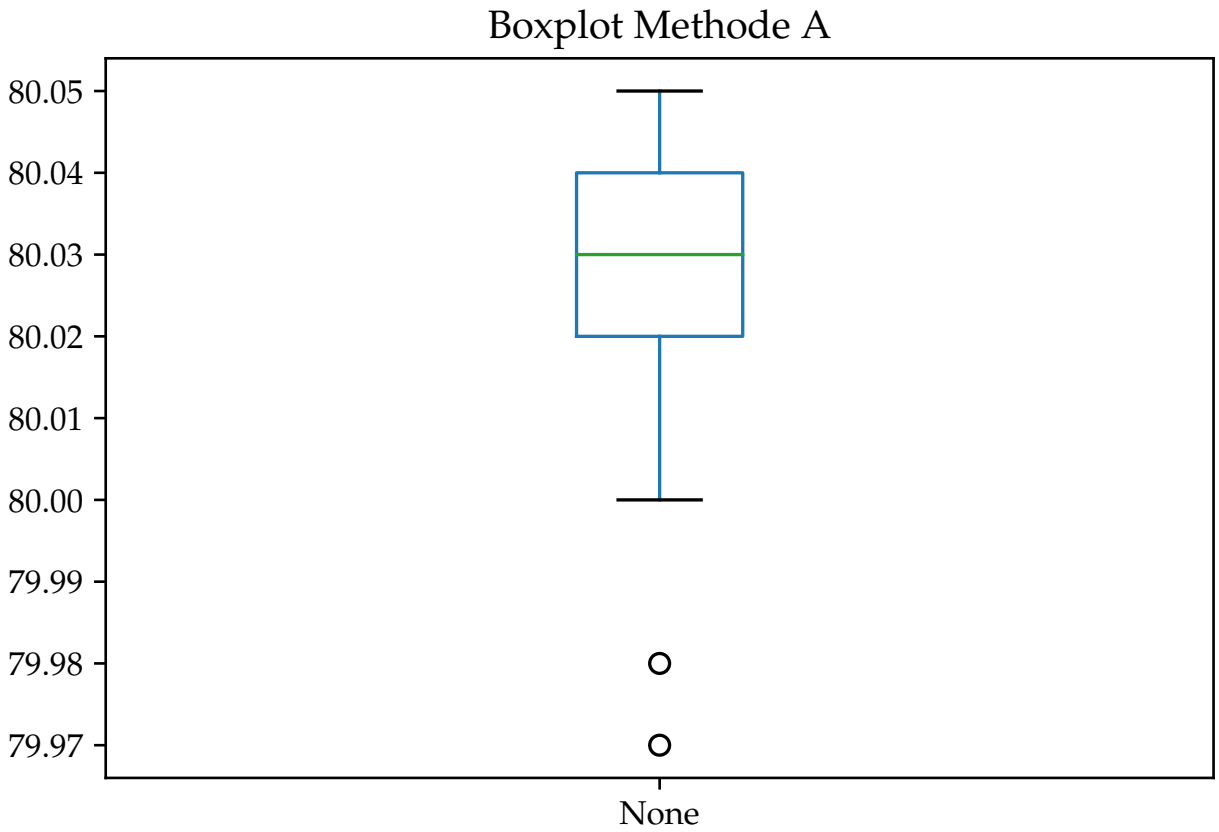
Ausreisser
(falls vorhanden)

Boxplot: Schematischer Aufbau

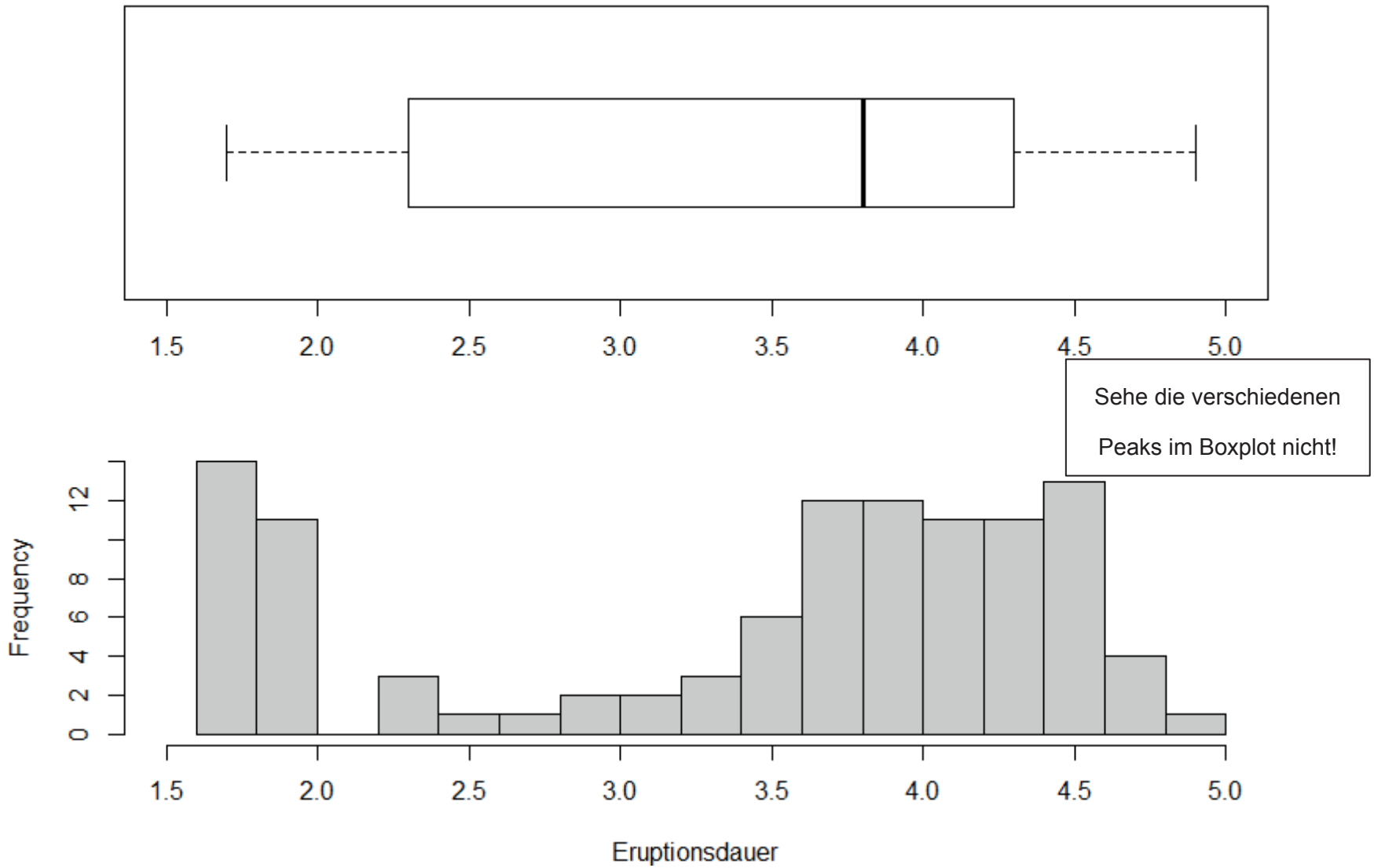
- *Grösste normale Beobachtung*: Grösste Beobachtung, die höchstens $1.5 \cdot r$ vom oberen Quartil entfernt ist (r: *Quartilsdifferenz*)
- *Kleinste normale Beobachtung*: Analog definiert mit dem unteren Quartil
- *Ausreisser* sind Punkte, die ausserhalb dieser Bereiche liegen

Python

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

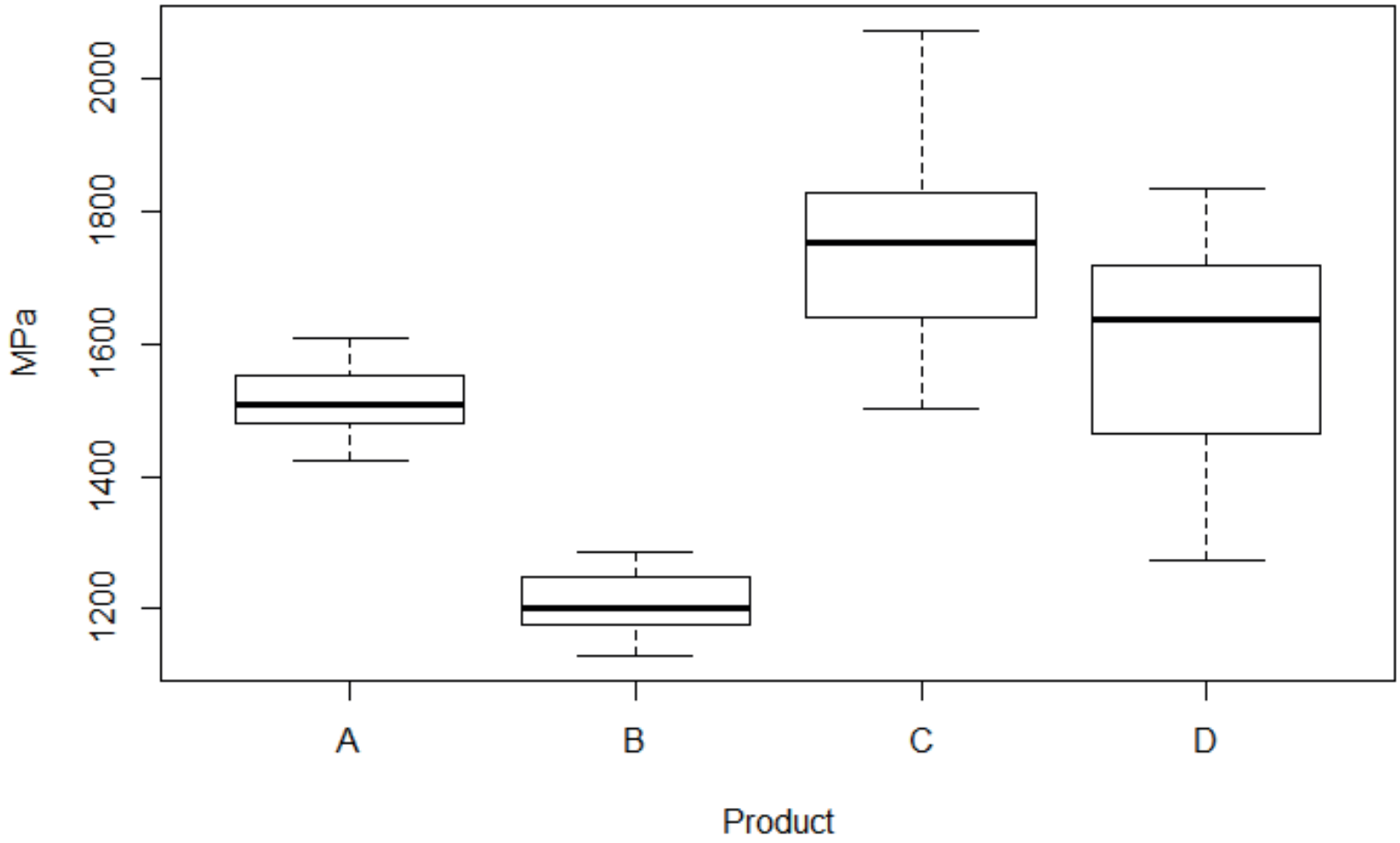


Boxplot und Histogramm der Eruptionsdauer



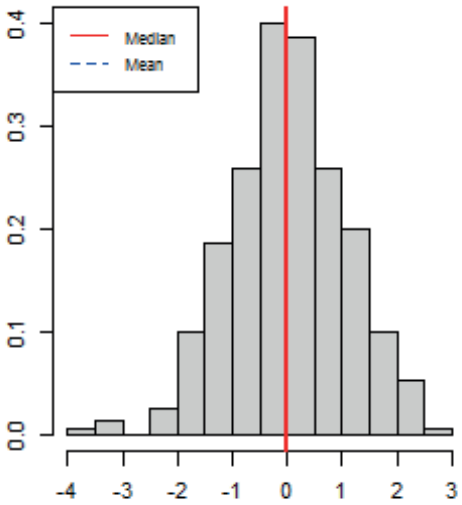
Mehrere Boxplots

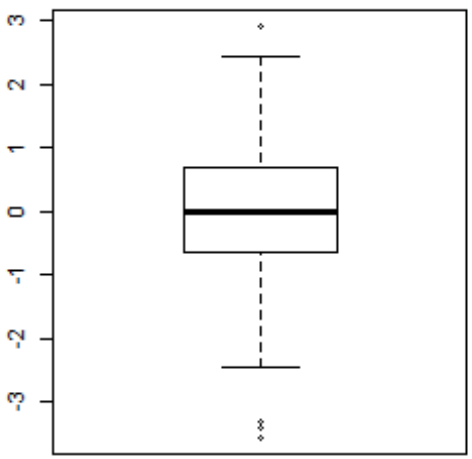
Mit mehreren Boxplots kann man einfach und schnell die Verteilung von verschiedenen Gruppen (Methoden, Produkte, ...) vergleichen



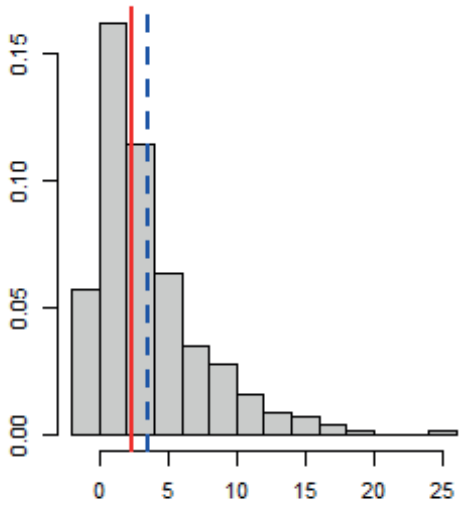
Schiefe

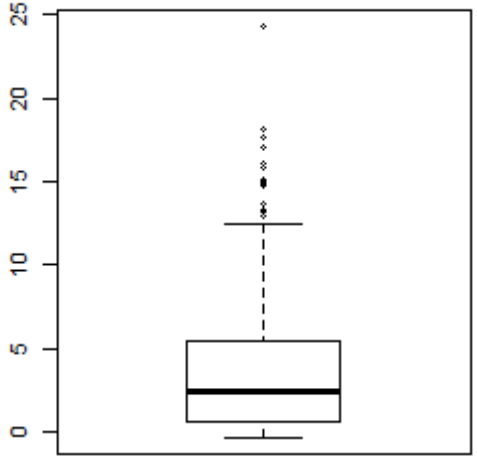
symmetrisch



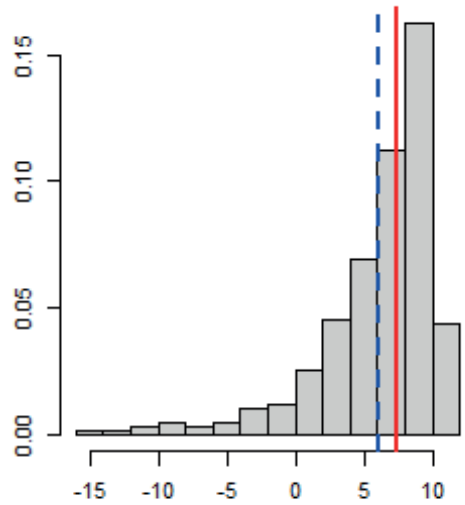


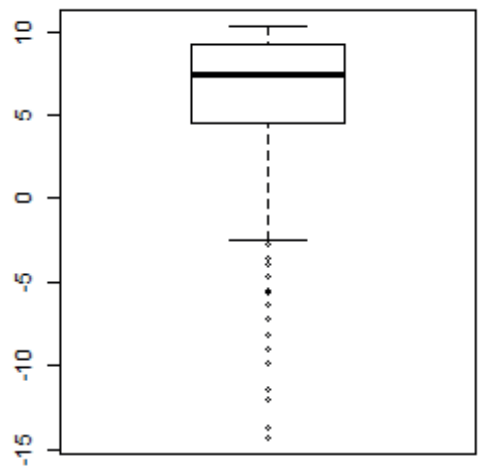
rechtsschief





linksschief





Boxplot: Bemerkungen

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