Stochastik

Deskriptive Statistik

Mirko Birbaumer

Hochschule Luzern Technik & Architektur

Der Begriff Wahrscheinlichkeit in der Alltagssprache

- Beispiele, wo der Begriff **Wahrscheinlichkeit** im Alltag auftaucht:
  - "Die *Wahrscheinlichkeit*, dass es heute morgen regnet, liegt bei 60 Prozent"
  - "Die *Wahrscheinlichkeit*, dass ich hundert Jahre alt werde, ist klein."
  - "Wie gross ist die *Wahrscheinlichkeit*, dass Geothermie-Borungen in Basel ein Erdbeben von einer bestimmten Grössenordnung auslösen? "
  - "Wie gross ist die *Wahrscheinlichkeit*, dass ein Geiger-Zähler in den nächsten 10 Sekunden 20 Zerfälle registriert? "
  - "Wie gross ist die *Wahrscheinlichkeit*, dass der Wert vom Euro in diesem Jahr über 1.20 Franken steigt ? "
- **Wahrscheinlichkeiten** geben wir im Zusammenhang mit Vermutungen an, wenn wir uns nicht sicher sind.

Der Begriff Wahrscheinlichkeit in der Alltagssprache

- Wir stellen Vermutungen an, wenn wir eine Aussage oder Vorhersage machen möchten, aber nur über **unvollständige** Informationen oder **unsichere** Kenntnisse verfügen.

- Wir müssen aufgrund unvollständiger Informationen eine Entscheidung fällen:
  - "Soll ich heute morgen einen Regenschirm mitnehmen? "
  - Soll ich eine Bergtour unternehmen, wenn die Wahrscheinlichkeit für Gewitter bei 30% liegt?
  - "Soll ich mich bei einer Bank bewerben, oder selbstversorgender Bio-Bauer werden? "

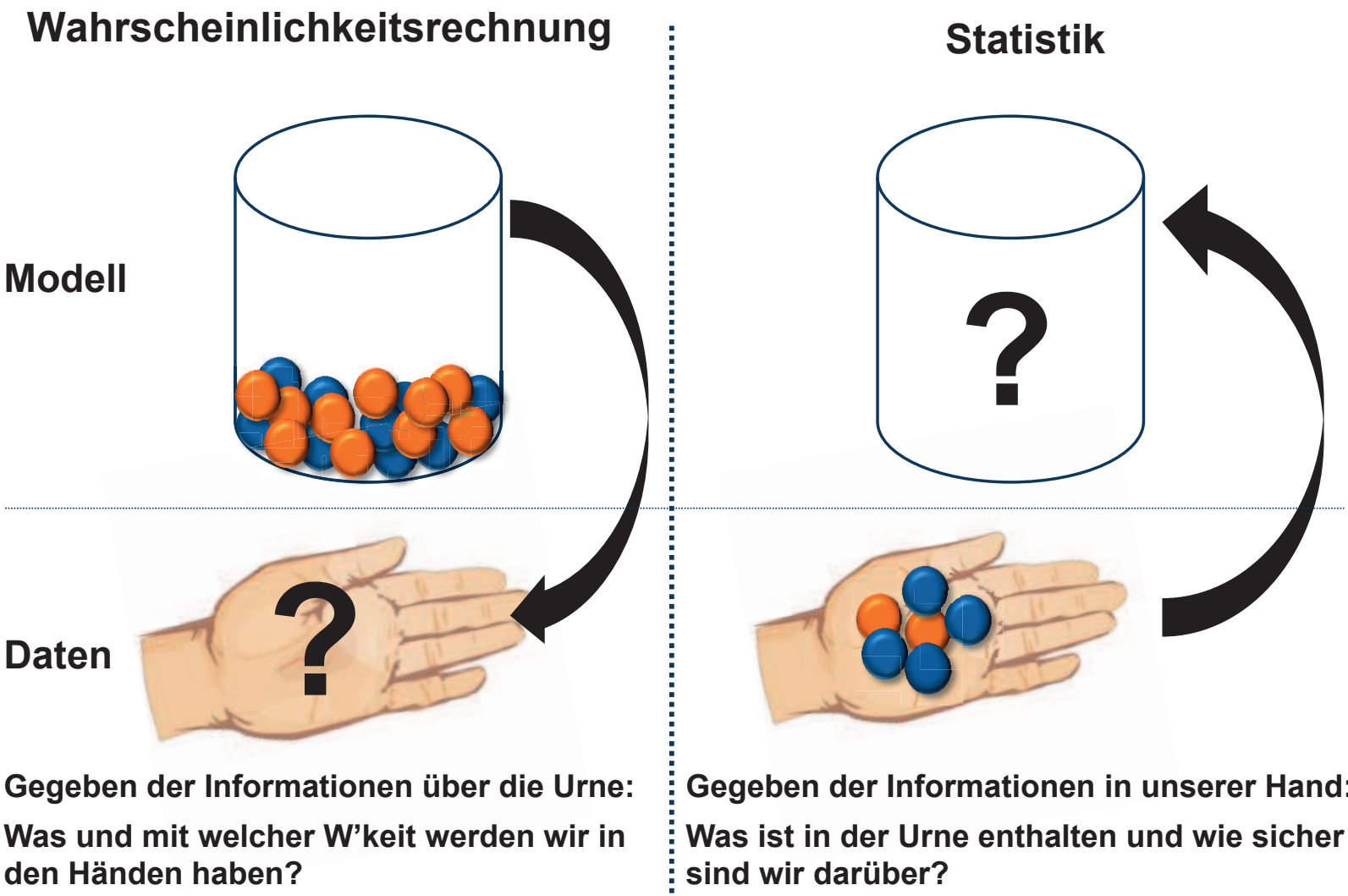
Wozu braucht ein Ingenieur Statistik?

- Sie erhalten als Ingenieur den Auftrag, die Höhe eines Dammes zu berechnen.
- Sie wissen nicht mit Sicherheit, wie gross in den nächsten (z.B.) 100 Jahren der **maximale Wasserstand** sein wird.
- D.h. die zukünftigen Ereignisse unterliegen dem **Zufall**.
- Also müssen Sie eine Entscheidung unter Unsicherheit treffen. Sie müssen daher versuchen, die Unsicherheit zu **quantifizieren**.
- Der Damm soll hoch genug sein (Sicherheit), aber auch die Wirtschaftlichkeit muss gewährleistet sein.

Wie hilft Ihnen Stochastik bei Ihrer Aufgabe?

- Sie verschaffen sich einen Überblick über die Aufzeichnung des Wasserstandes in den letzten hundert Jahren (→**Deskriptive Statistik**)
- Sie wählen ein geeignetes Modell, das die Verteilung des jährlichen maximalen Wasserstandes beschreibt (→**Wahrscheinlichkeitsmodell**)
- Sie schätzen die Parameter Ihres Wahrscheinlichkeitsmodells aus den Daten (→**Parameterschätzung**) und geben deren Unsicherheit an (→**Vertrauensintervall**)
- Aufgrund der Wahrscheinlichkeitsverteilung wählen Sie eine geeignete Dammhöhe. Unter Umständen führen Sie eine Kosten-Nutzen-Rechnung durch (Erwartungswert für Kosten bei Hochwasser→**Erwartungswert**)

Stochastik = Wahrscheinlichkeitsrechnung + Statistik



Gegeben der Informationen über die Urne:

Was und mit welcher W'keit werden wir in den Händen haben?

Gegeben der Informationen in unserer Hand:

Was ist in der Urne enthalten und wie sicher sind wir darüber?

Wozu braucht ein Ingenieur Stochastik Ihrer Meinung nach?

Naturgesetze und Wahrscheinlichkeit

- **Anfangsbedingungen in physikalischem System** nie mit beliebiger Genauigkeit bestimmbar in der Praxis
- Vorhersagen einer physikalischen Grösse in einem Experiment aufgrund von Naturgesetzen sind immer Unsicherheiten/Schwankungen ausgesetzt
- Diese unvollständige Kenntnis der Anfangsbedingung führt zum Begriff der **Wahrscheinlichkeit**.
- Beispiele: Münzwurf, Galtonsches Nagelbrett
- **Quantenmechanik** ist ein Beispiel einer fundamental probabilistischen Theorie; Nutzen: Generieren von **nicht-deterministischen Zufallszahlen** mit radioaktivem Zerfall

Organisation Modul : Flipped Classroom

- Ausführliches **Vorlesungsskript** und **Unterrichtsfolien** stehen Ihnen zur Verfügung
- Sie lesen **vor dem Unterricht** im Selbststudium die Skriptkapitel, die im **Semesterwochenplan** für jede Semesterwoche angegeben werden und beantworten in Maple TA Quizfragen zu den behandelten Themen vor dem Unterricht.
- **Ablauf Unterricht**: es wird jeweils eine Übungsaufgabe zu einem Thema vorgelöst/besprochen, eine weitere Aufgabe zum gleichen Thema wird in Dreier-Gruppen gelöst.

Ziel des Unterrichts: Übungsserie möglichst vollständig lösen mit Unterstützung Ihrer Studienkollegen, Ihres Tutors, Ihres Assistenten und Dozenten.

Organisation Modul : Testatbedingungen und MEP

- **Testatbedingung:** 60% der **Quizfragen** müssen korrekt vor Beginn des Unterrichts gelöst werden.

- **Zugelassene Hilfsmittel an MEP:**
  - eine 15-einseitige eigenhändig von Hand geschriebene Zusammenfassung
  - eine Formelsammlung (z.B. von Papula)
  - die **R**-Referenzkarte (unter Umständen mit Ihren eigenen Anmerkungen)
  - Software **R** auf einem Prüfungslaptop

- **Ablauf der MEP:**
  - Sie schreiben alle Ihre Lösungen zu den Aufgaben mit vollständigen Zwischenschritten auf Papier nieder
  - Sie schreiben alle von Ihnen benützten **R**-Befehle in ein **R**-Script-File, das Sie mit *nachname_name.Rnw* benennen
  - Sie speichern diese Datei auf dem Prüfungs-USB-Stick ab

Organisation Modul: Tutorat

- Herr Christoph Zaugg wird Ihnen als Tutor am Dienstag von 16:45 bis 18:00 im Raum E201 zur Verfügung stehen.
- Sie werden mit Unterstützung des Tutors die Theorie des bevorstehenden Unterrichtsblocks besprechen und die Quizaufgaben lösen
- Sie können Fragen zu den Übungsaufgaben stellen, die Sie nicht verstanden haben
- Der Besuch des Tutorats unterstützt Sie dabei, die Theorie für den bevorstehenden Übungsblock zu verstanden
- Wir empfehlen Ihnen deswegen nachdrücklich, das Tutorat zu besuchen.

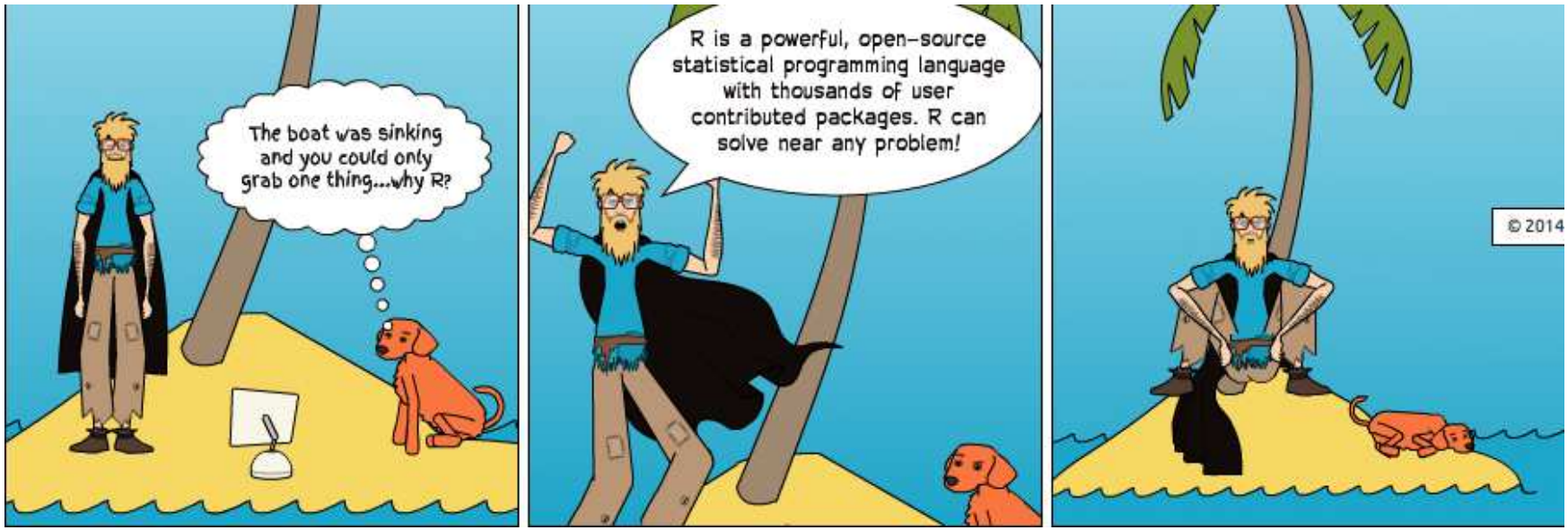
Organisation Modul : Statistik-Software **R**

- Wir werden die Statistiksoftware **R** verwenden, insbesondere `R Studio`.
- Die Beherrschung von **R** ist wesentlicher Bestandteil dieses Stochastik Moduls und auch prüfungsrelevant.
- Zusammenfassung der wichtigsten **R**-Befehle zusammengestellt in der **R-Referenzkarte**.
- Hervorragendes Nachschlagewerk zur Benützung von **R** (auf Ilias):

  Peter Dalgaard, *Introductory Statistics with R*, 2008, 2nd Edition, Springer

Entwarnung: Sie werden in diesem Modul bestimmt nicht an der Statistiksoftware **R** scheitern!

Einführung in Statistiksoftware **R**



- **R** ist eine frei erhältliche Programmiersprache für **statistisches Rechnen** und **statistische Graphiken**
- **R** ist eine interpretierte Programmiersprache; es existieren zahlreiche Benutzeroberflächen wie `R Studio`
- **R** ist mittlerweilen die bedeutendste Statistiksoftware in vielen Gebieten wie der Finanzmathematik und Bioinformatik.

Einführung in Statistiksoftware **R** : Start und Hilfe

- Download von *RStudio* und **R** unter `http://www.rstudio.com/ide/download/`
- **R** besteht aus einem Grundprogramm mit vielen Zusätzen, den sogenannten *packages* oder *Paketen*
- **R** bietet eine Vielzahl frei verfügbarer Pakete (≈ 4200)
- Ein Paket enthält unterschiedlichste, spezielle Funktionen
- Beim Start von **R** ist nur eine Grundausstattung geladen, alle anderen Pakete müssen zusätzlich geladen werden
- Jeder kann sein eigenes Paket schreiben

R als Taschenrechner

R-Befehl: Wertzuweisung mit $<-$

```
> a <- 5
> b <- 3
> a
[1] 5
> a+b
[1] 8
> a-b
[1] 2
> a*b
[1] 15
> a/b
[1] 1.666667
> sqrt(a)
[1] 2.236068
> sin(b)
[1] 0.14112
```

Vektoren in **R**

R-Befehl: length(),sort(), order()

```
> a <- c(5,4,6)
> a
[1] 5 4 6
> length(a)
[1] 3
> a[1]
[1] 5
> a[2]
[1] 4
> 3*a
[1] 15 12 18
> sort(a)
[1] 4 5 6
> order(a)
[1] 2 1 3
```

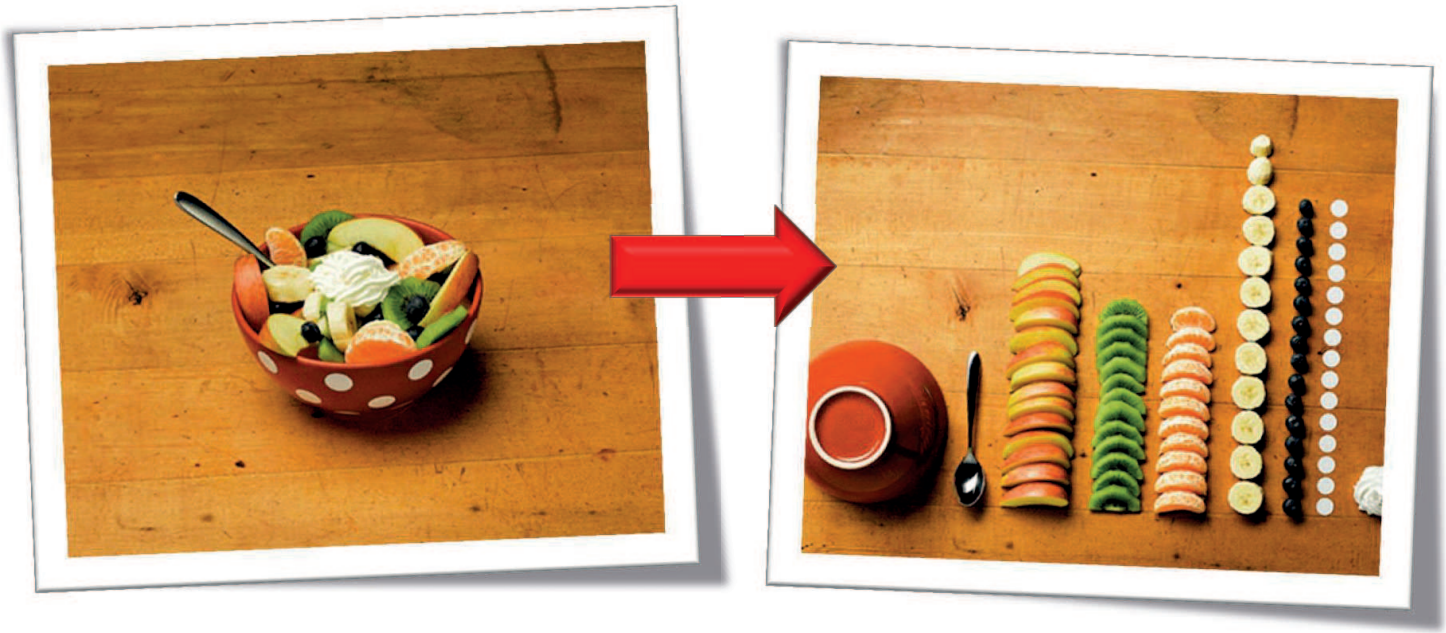
Matrizen in **R**

R-Befehl: matrix()

```
> mat <- matrix(c(1,0,0,0,2,0,0,0,3),nrow=3)
> mat
[,1] [,2] [,3]
[1,] 1 0 0
[2,] 0 2 0
[3,] 0 0 3
> mat[1,1]
[1] 1
> mat[,2]
[1] 0 2 0
> mat[3,]
[1] 0 0 3
> mat[1:3,2]
[1] 0 2 0
> mat[-1,]
```

Ziele der Deskriptiven Statistik

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



Daten

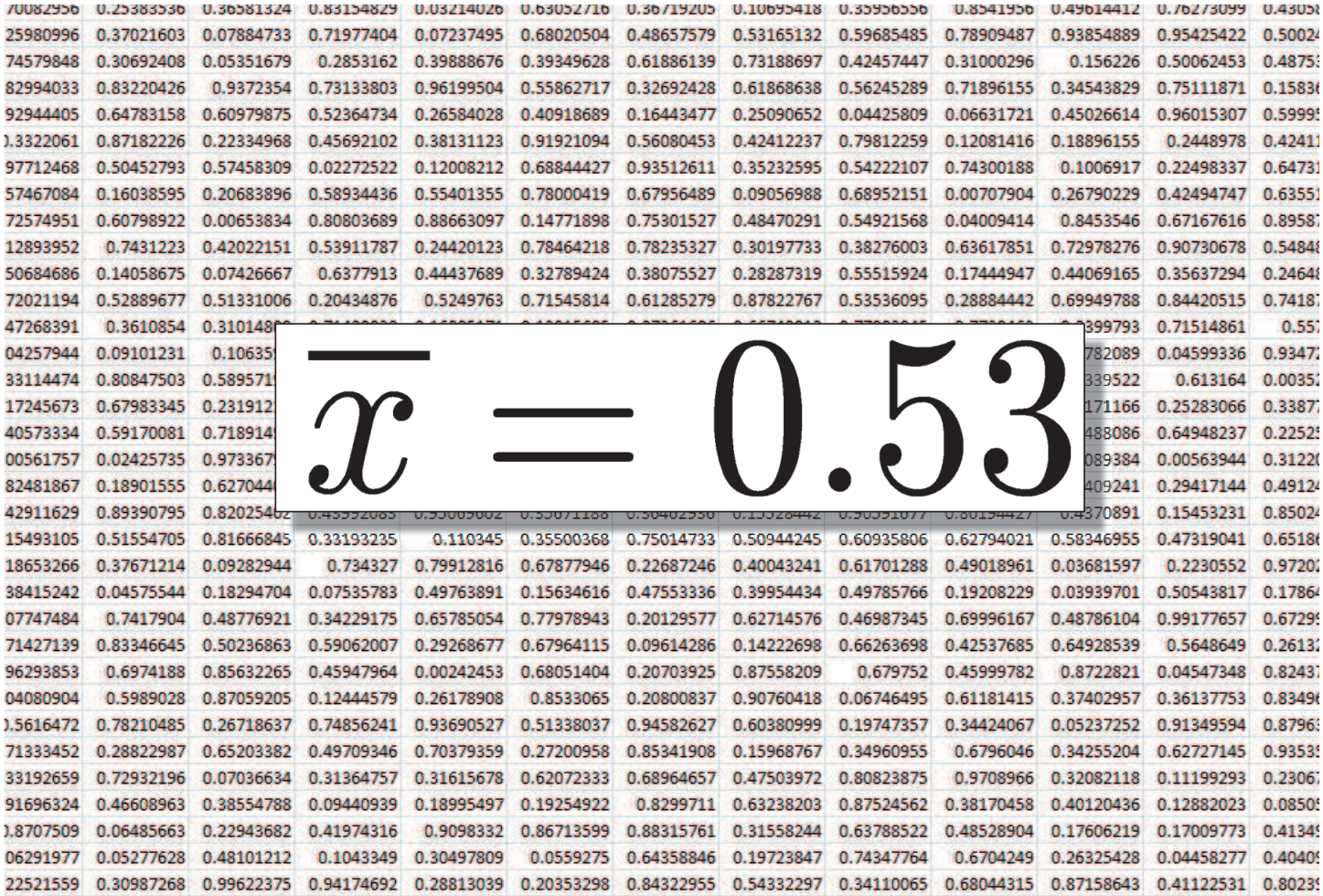
- Wir betrachten im folgenden **reale Daten**

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

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

- Basierend auf den Daten können wir diverse **Kennwerte** berechnen bzw. die Daten **graphisch darstellen**.

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


$\overline{x} = 0.53$

Überblick

- Wir haben n beobachtete Datenpunkte (Messungen)

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

(z.B. die $n = 13$ Messungen der Schmelzwärme mit Methode A)

- Wir unterscheiden 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

- **Arithmetisches Mittel**

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

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

$$\overline{x} = \frac{79.98 + 80.04 + \ldots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.02077$$

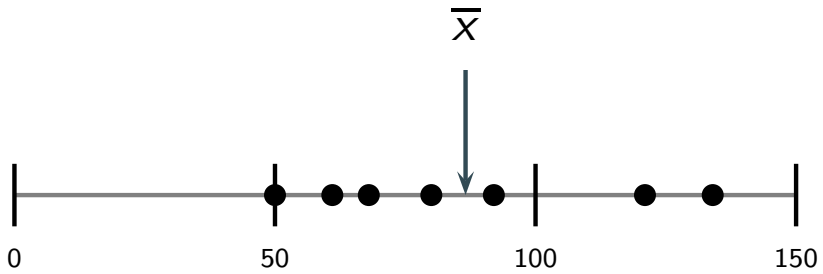
- R-Befehl

R-Befehl: mean()

```
> methodeA <- c(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)
> mean(methodeA)
[1] 80.02077
```

Arithmetisches Mittel

- Arithmetische Mittel: anschaulich



Schwerpunkt der Daten

Streuung

- Arithmetisches Mittel sagt einiges über Datensatz aus: „Wo ist die Mitte?“
- 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
- Im ersten Fall gibt zwei gute und zwei schlechte Schüler und im zweiten Fall sind alle Schüler gleich gut
- Wir sagen, die Datensätze haben eine verschiedene *Streuung* um die Mittelwerte

Streuung

- Streuung numerisch: Erste 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

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

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

Streuung

- D.h.: Noten weichen im Schnitt 1.5 vom Mittelwert ab
- Im zweiten Fall ist dieser Wert natürlich 0
- Je grösser dieser Wert (immer grösser gleich 0) , desto mehr unterscheiden sich die Daten bei gleichem Mittelwert untereinander
- Dieser Wert für die Streuung heisst auch *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

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

$$s_x = \sqrt{\mathrm{var}(\mathrm{x})} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\,.$$

Eigenschaften der Varianz

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

Beispiele: Schmelzwärme

- Arithmetische Mittel der $n = 13$ Messungen ist $\overline{x} = 80.02$ (siehe oben)
- Die empirische Varianz ergibt

$$\begin{aligned}\mathrm{var}(x) &= \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \cdots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1} \\ &= 0.000\,574\end{aligned}$$

- Die empirische Standardabweichung ist dann

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

- D.h.: die „mittlere“ Abweichung vom Mittelwert 80.02 cal/g ist 0.024 cal/g
- Von Hand sehr mühsam. Mit R:

R-Befehl: var(), sd()

```
> var(methodeA)
[1] 0.000574359
> sd(methodeA)
[1] 0.02396579
```

Median

- Ein weiteres Lagemass für die „Mitte“ ist der *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.: die 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.

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

- Der Median ist nun sehr einfach zu bestimmen
- Er ist unter diesen 13 Messungen dann der 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

- Der Median des Datensatzes der Methode A ist 80.03
- Das bedeutet, dass knapp die Hälfte der Messwerte, nämlich 6 Beobachtungen kleiner oder gleich 80.03 sind
- Ebenso sind 6 Messwerte grösser oder gleich dem Median
- Da es eine ungerade Anzahl Messungen sind, können wir auch nicht von *genau* der Hälfte der Messwerte sprechen

Median

- Vorher: Anzahl der Daten ungerade und damit ist die mittlere Beobachtung eindeutig bestimmt
- Ist die Anzahl der Daten gerade, so gibt es *keine* mittlere Beobachtung
- Wir *definieren* den Median: Mittelwert der beiden mittleren Beobachtungen
- Beispiel: Datensatz der Methode B hat 8 Beobachtungen
- Wir ordnen den Datensatz und für den Median nehmen wir den 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

- R–Befehl

R-Befehl: median()

```
> median(methodeA)
[1] 80.03
> methodeB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03,
79.95, 79.97)
> median(methodeB)
[1] 79.97
```

- Als Median kann Wert auftreten, der in der Messreihe gar nicht vorkommt

- Wären die beiden mittleren Beobachtungen der Methode B die Werte 79.97 und 79.98, so wäre der Median der Durchschnitt dieser Werte:

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

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: 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

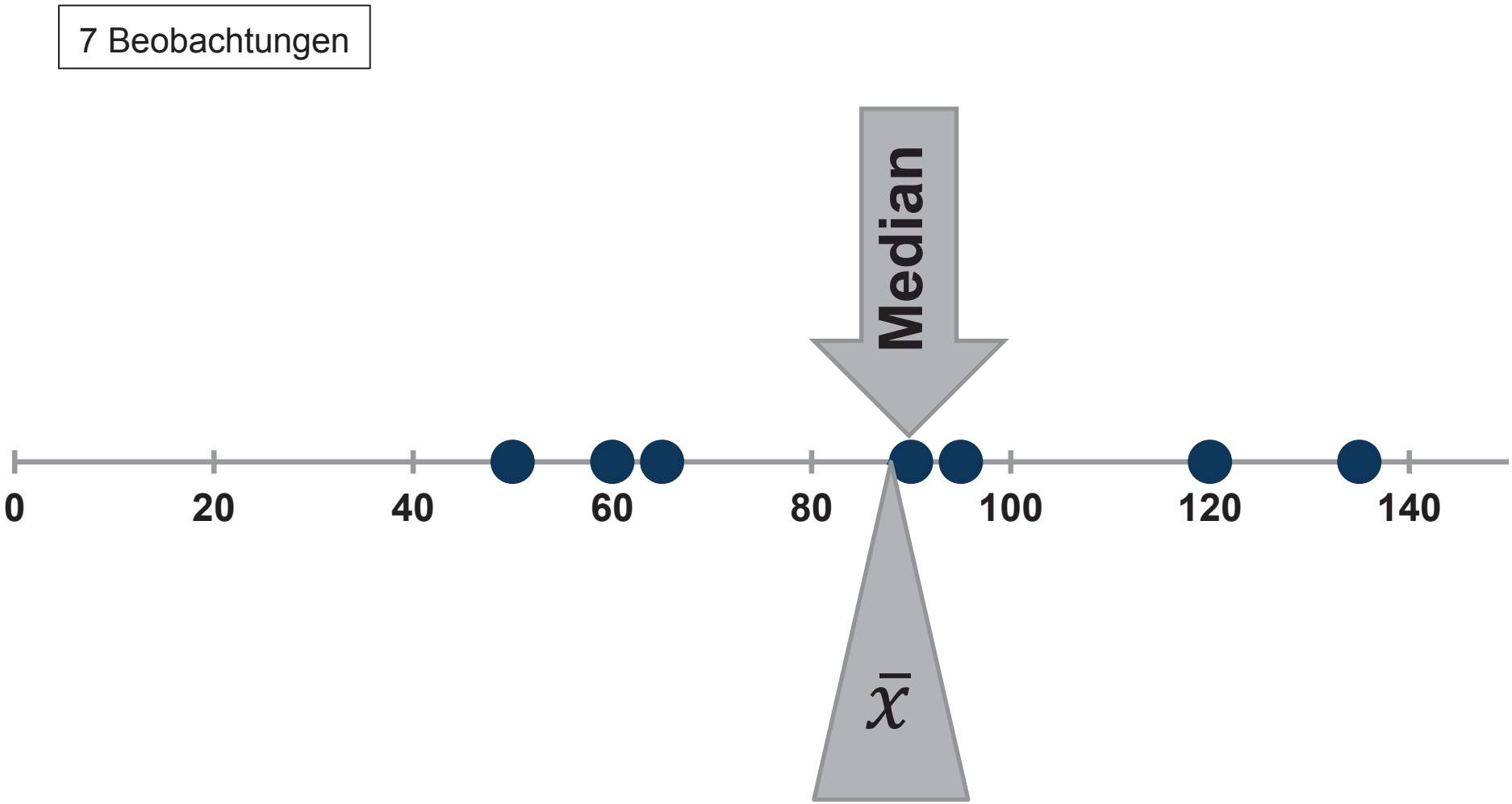
$$\overline{x} = 135.44$$

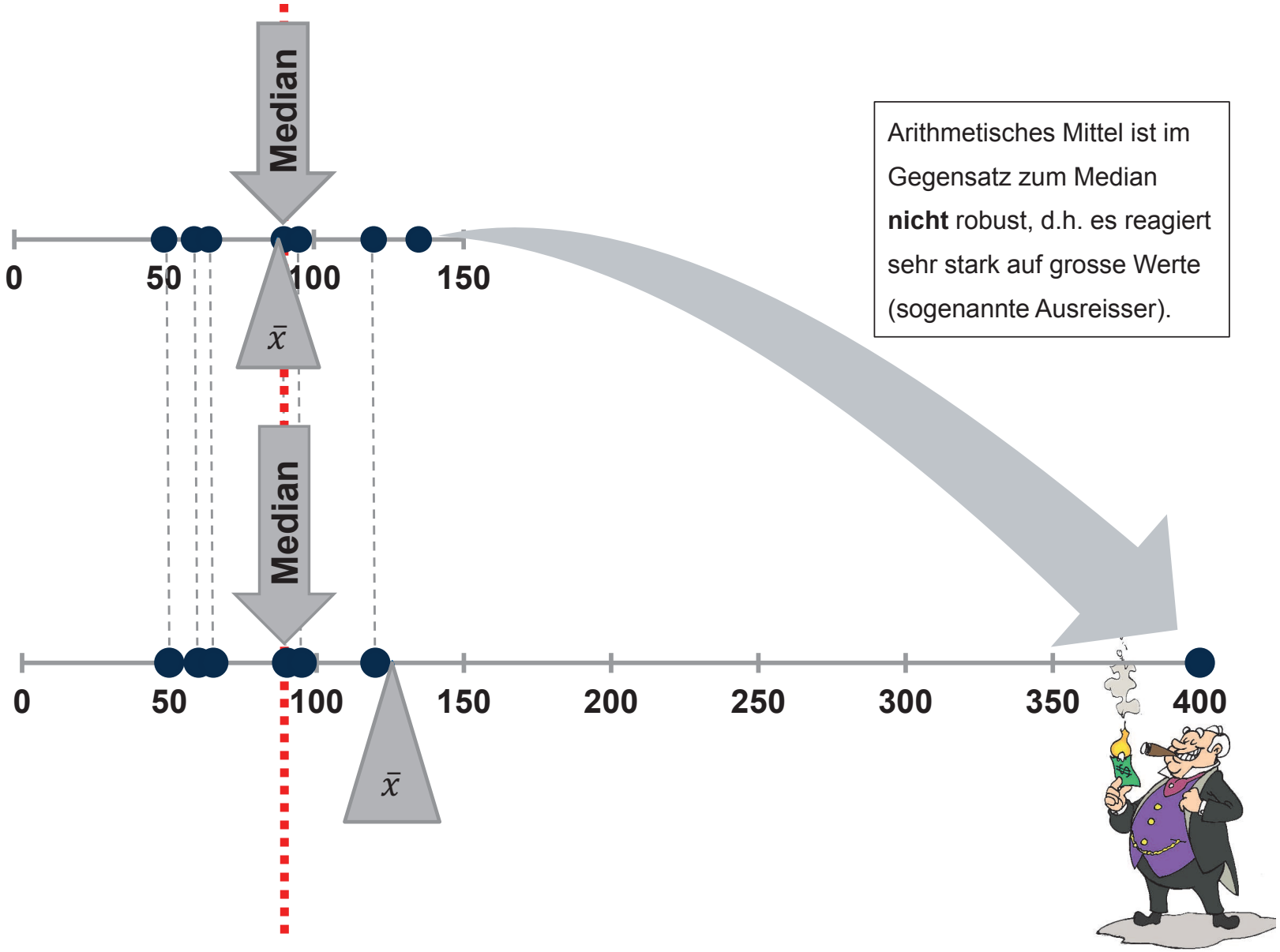
- Der Median ist aber nach wie vor

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

- Das arithmetische Mittel wird also durch Veränderung einer Beobachtung sehr stark beeinflusst, während der Median hier gleich bleibt – er ist robust.

Arithmetisches Mittel vs. Median: Einkommen [k CHF]





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?"

Oberes und unteres Quartil

- *Repetition*: Der **Median** ist derjenige Wert, wo die Hälfte der Beobachtungen kleiner (oder gleich) wie dieser Wert sind.

- Ähnlich zum Median gibt es noch das **untere** und **obere 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: für die meisten Datensätze sind es nicht *exakt* 25 % der Anzahl Beobachtungen, die kleiner als das untere Quartil sind

Beispiel: Schmelzwärme Methode A

- Methode A hat $n = 13$ Messpunkte und 25 % dieser Anzahl ist 3.25.
- **Unteres Quartil:** nächstgrösserer Wert $x_{(4)}$

  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 also 80.02: rund ein Viertel der Messwerte ist kleiner oder gleich diesem Wert
- **Oberes Quartil**: wir 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

- Oberes Quartil ist 80.04: rund drei Viertel aller Messwerte sind also kleiner oder gleich diesem Wert

Beispiel: Schmelzwärme Methode B

- Messwerte mit Methode B:

$$79.94;\ 79.95;\ 79.97;\ 79.97;\ 79.97;\ 79.94;\ 80.02;\ 80.03$$

- 25 % der Werte ist $0.25 \cdot 8 = 2$

- 2 ist eine ganze Zahl : wir *wählen* den Durchschnitt von $x_{(2)}$ und $x_{(3)}$ als **unteres Quartil**

- Dann sind 2 Beoachtungen kleiner und 6 Beobachtungen grösser als dieser Wert

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

$$\frac{79.95 + 79.97}{2} = 79.96$$

- **Unteres Quartil** der Methode B ist also 79.96

Berechnung der Quartile/Quantile mit R

- Die Software R kennt keine eigenen Befehle für die Quartile

R-Befehl: quantile()

```
> # Syntax für das untere Quartil: p=0.25
> quantile(methodeA,0.25,type=2)
[1] 80.02
> quantile(methodeB,0.25,type=2)
[1] 79.96
> # Syntax für das obere Quartil: p=0.75
> quantile(methodeA,0.75,type=2)
[1] 80.04
```

- Damit R die Quartile nach unserer Definition berechnet, müssen wir die Option type=2 hinzufügen

Quartilsdifferenz

- Die **Quartilsdifferenz** ist ein robustes **Streuungsmass** für die Daten

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

- Quartilsdifferenz 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
- Quartilsdifferenz der Methode A : $80.04 - 80.02 = 0.02$

R-Befehl: IQR()

```
> IQR(methodeA,type=2)
[1] 0.02
```

- Rund die Hälfte der Messwerte liegt also in einem Bereich der Länge 0.02

Quantile

- *Beispiel*: 10 %-Quantil, derjenige Wert, wo 10 % der Werte kleiner oder gleich und 90 % der Werte grösser oder gleich diesem Wert sind

- Empirischer Median ist empirisches 50 %–Quantil; empirisches 25 %–Quantil ist unteres Quartil; empirisches 75 %–Quantil ist oberes Quartil

- Das **empirische α-Quantil** ist anschaulich gesprochen der Wert, bei dem $\alpha \times 100\,\%$ der Datenpunkte kleiner oder gleich und $(1-\alpha) \times 100\,\%$ der Punkte grösser oder gleich sind

Empirische α-Quantile

$$\frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)}) \quad , \text{ falls } \alpha \cdot n \text{ eine natürliche Zahl ist,}$$

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

Quantile mit R

R-Befehl: quantile()

```
> quantile(methodeA,.1,type=2)
  10%
  79.98
> quantile(methodeA,.7,type=2)
  70%
  80.04
```

- Rund 10 % der Messwerte sind kleiner oder gleich 79.97.
- Entsprechend sind rund 70 % der Messwerte kleiner oder gleich 80.04

Beispiel: Notenstatistik

- In Schulklasse mit 24 SchülerInnen gab es an Prüfung folgende Noten:

  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

R-Befehl: Quantile

```
> noten.1 <- c(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)
> quantile(noten.1,seq(.2,1,.2),type=2)
  20%   40%   60%   80%   100%
  3.6   4.2   5.0   5.6   6.0
```

- Rund 20 % der SchülerInnen schlechter als 3.6 sind
- Das 60 %–Quantil besagt, dass rund 60 % der SchülerInnen schlechter oder gleich einer 5 waren
- Oder 40 % haben eine 5 oder sind besser