Deskriptive Statistik (eindimensional)

Peter Büchel

HSLU I

ASTAT: Block 02

Daten

- Daten und Statistiken bestimmen immer mehr unser Leben
- Zeitung: Prognose zur nächsten Abstimmung oder Wahlen
  $\rightarrow$ Befragung
- Googeln: Wie „weiss" Google so genau, was man suchen will?
  $\rightarrow$ Google wertet Suchanfragen aus
- Passkontrolle am Flughafen: Wie erkennt die Software Gesichter?
  $\rightarrow$ Gesichter werden charakterisiert
- Wetterbericht: Wie kommt die Vorhersage zustande? $\rightarrow$ Modell aufgrund früherer Wetterdaten (und Theorie)
- Börsenkurse: Wie lässt sich aus dem Börsenverlauf der letzten paar Tage, der Kurs für die nächsten paar Tage vorhersagen?
  $\rightarrow$ Modellierung aus alten Daten

Datensätze (eindimensional)

- *Liste*: Einfachste Variante eines Datensatzes
- Bsp: Körpergrössen von 5 Personen

  1.75, 1.80, 1.72, 1.65, 1.54
- Solche Listen heissen auch *eindimensionale Datensätze* oder *Messreihen*

Datensätze (zweidimensional)

- Häufigste Form von Datensätzen sind *Tabellen* oder *zweidimensionale Datensätze*
- Bsp:

Person	Grösse	Gewicht	Geschlecht	Nationalität
A	1.82	72	m	CH
B	1.75	82	w	D
C	1.61	70	w	CH
D	1.80	83	m	A
E	1.89	95	w	FL

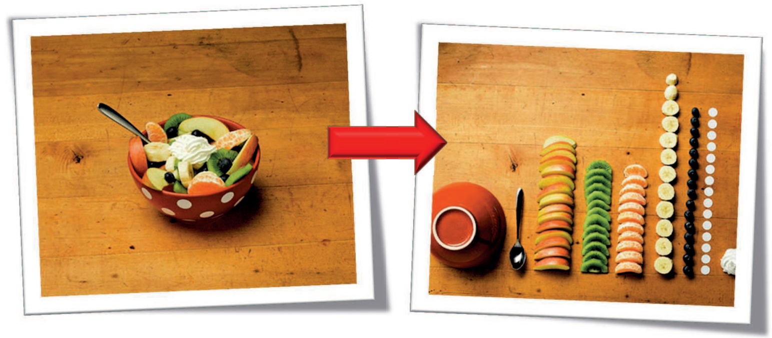
- Grösse und Gewicht: *quantitative* Daten, also (gemessene) Zahlen
- Können, zumindest theoretisch, jeden beliebigen Zahlwert annehmen
- Geschlecht und Nationalität: *qualitative* Daten
- Diese können nur eine bestimmte Anzahl Werte annehmen (müssen auch keine Zahlen sein)

Deskriptive Statistik

- Deskriptive Statistik befasst sich mit der Darstellung von Datensätzen (Zusammenstellung verschiedener Daten)
- Datensätze durch gewisse Zahlen charakterisieren (z.B. Mittelwert)
- *Und* graphisch darstellen
- Zunächst *eindimensionale* Daten: *Eine* Messgrösse wird an einem Untersuchungsobjekt ermittelt (zweidimensionale später)

Ziele der Deskriptiven Statistik

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



Beispieldatensatz

- Messungen Körpergewicht
- Erfahrung: Man steht am Morgen auf Waage und merkt sich Gewicht
- Steht nochmals auf die Waage und erhält leicht anderes Resultat
- Wir wollen es genauer wissen
- Nehmen 80 Kilogramm schweren Metallblock, der geeicht ist, d.h. er hat mit sehr grosser Genauigkeit 80 kg
- Gewicht dieses Metallblocks wird mehrere Male mit zwei Waagen, Waage A und Waage B, gemessen
- Zwei *Datensätze* mit Gewichten (in kg; auf 10 g genau gemessen)

- Tabelle:

Waage A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Waage A	80.03	80.02	80.00	80.02					
Waage B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

- Frage: Warum führen verschiedene Messungen, die am gleichen Objekt stattfinden zu unterschiedliche Resultaten?
- Messungen finden nie unter *exakt* denselben Bedingungen statt
- Scheinbar genaue Angaben sind nur *ungefähre* Angaben
  - Kalorienzahl auf einer Packung Schokolade
  - Inhalt 500 ml Pet-Flasche: Keine zwei Pet-Flaschen sind *absolut* gleich
  - Gesichtserkennung am Flughafen: Sie haben *nie* denselben Gesichtsausdruck

Zurück zum Beispiel der Waagen

- Messungen wurden mit grösstmöglichen Sorgfalt durchgeführt
- Trotzdem variieren die Messwerte innerhalb beider Waagen
- Es stellen sich hier nun die folgenden Fragen:
  - Gibt es einen Unterschied zwischen der Waage A und der Waage B?
  - Falls ja, wie können wir diesen Unterschied ermitteln?
- Es fällt auf:
  - Beide Waagen: Messwerte um 80 herum liegen (sollte auch so sein)
  - Waage A: Nur 2 Werte von 13 *unter* 80
  - Waage B: Nur 2 von 8 Werten *über* 80 liegen
  - Werte der Waage A sind also *eher* grösser als die der Waage B

- Was heisst hier aber „eher"?
- Wie kann man die beiden Messreihen miteinander vergleichen?
- Ziel: Messreihen irgendwie *zusammenzufassen*, um die beiden Waagen miteinander vergleichen zu können
- *Deskriptive Statistik* beschäftigt sich damit, auf welche Weisen Daten organisiert und zusammengefasst werden können
- Ziel: Interpretation und darauffolgende statistische Analyse dieser Daten vereinfachen

- Kennzahlen sollen Daten numerisch zusammenfassen und grob charakterisieren
- Bei statistischen Analysen ist es sehr wichtig, nicht einfach blind ein Modell anzupassen oder *ein* statistisches Verfahren anzuwenden
- Daten sollten *immer* mit Hilfe von geeigneten graphischen Mitteln *und* den Kennzahlen dargestellt werden
- Nur auf diese Weise kann man (teils unerwartete) Strukturen und Besonderheiten entdecken

- Aber:

Warnung!!!

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

- Bsp: Noten einer Schulklasse mit 24 Lernenden an einer Prüfung:

  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
- Notendurchschnitt ist 4.51
- Dieser Wert sagt über Klasse als Ganzes etwas aus, aber nichts mehr über die einzelnen Lernenden
- Kennen nur Zahl 4.51: Keine Information mehr, wie die einzelnen Lernenden abgeschnitten haben
- Wissen nicht einmal, wieviele Lernende in der Klasse sind

Bezeichnungen

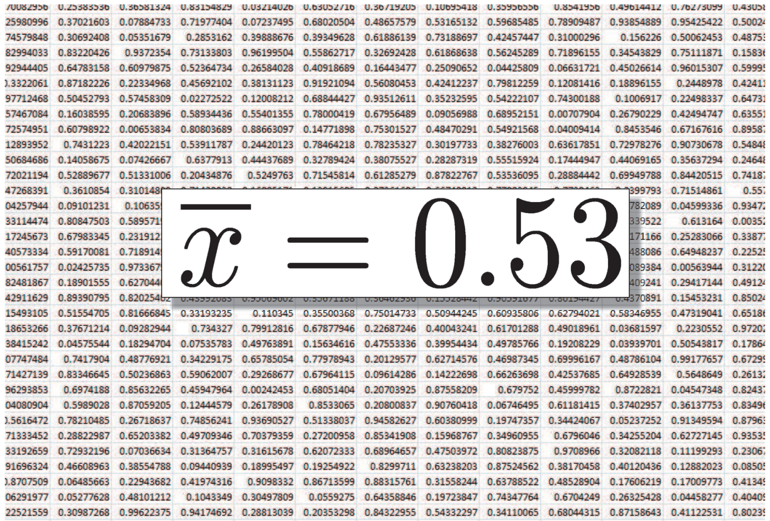
- Standardbezeichnung von Daten mit

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

- n heisst *Umfang* der Messreihe (Daten, Datensatz)
- Beispiel Messreihe der Waage A hat Umfang $n = 13$:

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \ldots, \quad x_{13} = 80.02$$

Kennzahlen



Überblick über die Kennzahlen

- 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

- Umgangsprachlich: *Durchschnitt*
- Man addiert alle Daten und teilt sie durch die Anzahl Daten (Umfang)
- Definition:

Arithmetisches Mittel

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

- Sprechweise: „x quer"
- Beispiel Waage A: Arithmetische Mittel der $n = 13$ Messungen

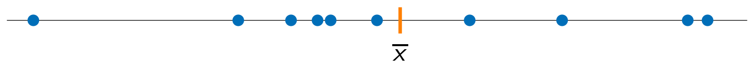
$$\overline{x} = \frac{79.98 + 80.04 + \ldots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.020\,77$$

- R-Befehl für arithmetisches Mittel `mean(...)`

```
waageA <- 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(waageA)
## [1] 80.02077
```

- Arithmetisches Mittel: Anschaulich



Vergleich der Waagen

- Waagen mit dem arithmetischen Mittel miteinander vergleichen:

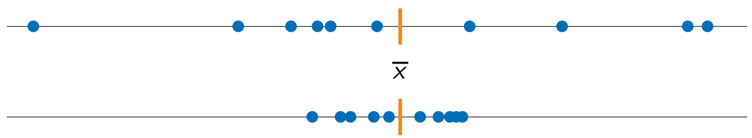
```
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95,
            79.97)

mean(waageB)
## [1] 79.97875
```

- Waage B hat also durchschnittlich die tieferen Werte als die Waage A

Streuung graphisch

- Arithmetisches Mittel sagt schon einiges über einen Datensatz aus
- Arithmetisches Mittel : „Wo ist die „Mitte" der Daten?"
- Aber: Arithmetisches Mittel sagt nicht alles über Messreihe aus
- Graphisches Beispiel:



- Beide Datenreihen haben dasselbe arithmetische Mittel
- Punkte der zweiten Datenreihe sind durchschnittlich viel näher beim Mittelwert $\overline{x}$ als die Punkte der ersten Datenreihe
- Verschiedene *Streuung* der Daten um den Durchschnitt

Streuung numerisch

- 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
- Wie kann man diese Streuung mathematisch beschreiben?

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

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

- Zweiter Fall auch 0:

$$\frac{(4-4)+(4-4)+(4-4)+(4-4)}{4}=\frac{0+0+0+0}{4}=\frac{0}{4}=0$$

- Keine Aussage über unterschiedliche Streuung
- Problem: Unterschiede können *negativ* werden $\rightarrow$ Können sich aufheben

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

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

- D.h.: Noten weichen im Schnitt 1.5 vom Mittelwert ab
- 2. Fall: Dieser Wert natürlich auch 0

$$\frac{\big|(4-4)\big|+\big|(4-4)\big|+\big|(4-4)\big|+\big|(4-4)\big|}{4}=\frac{0+0+0+0}{4}=0$$

- Je grösser dieser Wert (immer grösser gleich 0) , desto mehr unterscheiden sich die Daten bei gleichem Mittelwert untereinander
- Dieser Wert für die Streuung heisst *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

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

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

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
- In einigen Büchern: n anstatt $n-1$: Spielt für grosse n keine Rolle (siehe Jupyter Notebook)
- Standardabweichung ist die Wurzel der Varianz
- Durch 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

- *Wert* der empirische Varianz hat keine physikalische Bedeutung. Man weiss nur, je grösser der Wert umso grösser die Streuung
- Wichtig: *Nur die Standardabweichung s_x lässt sich konkret interpretieren.*
- Beispiel:
  - Daten in cm
  - Varianz wegen der Quadrierung der Daten in cm^2
  - Einheit der ursprünglichen Daten haben $\rightarrow$ Aus Varianz die Wurzel ziehen
- Für normalverteilte Daten hat die Standardabweichung noch eine schöne geometrische Interpretation (siehe später)

Beispiel: Waage *A*

- 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 kg ist 0.024 kg

- Von Hand sehr mühsam. Mit **R**:

```
var(waageA)
## [1] 0.000574359
sd(waageA)
## [1] 0.02396579
```

- **sd**: standard deviation

Median

- Ein weiteres Lagemass für die „Mitte": *Median*
- Sehr vereinfacht: Wert, bei dem die Hälfte der Messwerte unter oder gleich diesem Wert sind
- Andere Hälfte ist gleich diesem Messwert oder darüber
- Beispiel: Prüfung in der Schule ist Median 4.6
  - D.h.: Hälfte der Klasse hat *diese Note oder ist schlechter*
  - Umgekehrt: Andere Hälfte der Klasse hat *diese Note oder ist besser*
- Obige Interpretation: Medians sehr vereinfacht dargestellt
- Exakte Definition folgt nun

Geordnete Strichprobe

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

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

- Runde Klammern im Index: Daten geordnet
- Bestimmung *Median*: Daten zuerst der Grösse nach ordnen:

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

Beispiel Waage *A*

- Daten Waage *A* geordnet:

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 $\rightarrow$ Wert der mittleren Beobachtung
- 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 des Datensatzes der Waage *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

Beispiel Waage *B*

- 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 Waage *B* hat 8 Beobachtungen
- Ordnen den Datensatz: Median Durchschnitt von der 4. und 5. Beobachtung

79.94; 79.95; 79.97; 79.97; 79.97; 79.94; 80.02; 80.03

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

- R–Befehl

```
median(waageA)
## [1] 80.03
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97,
            80.03, 79.95, 79.97)

median(waageB)
## [1] 79.97
```

- Als Median kann Wert auftreten, der in Messreihe nicht ist
- Annahme: Mittlere Beobachtungen der Waage 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:
  - Kommt auf die jeweilige Problemstellung an
  - Am besten werden beide Masse gleichzeitig betrachtet
- 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

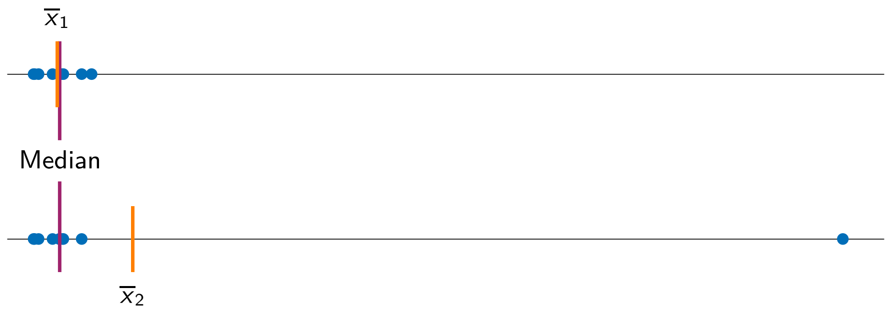
$$\overline{x} = 135.44$$

- Der Median ist aber nach wie vor

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

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

Graphisch



Beispiel

- Untersuchen typisches Haushaltseinkommen von Vororten von Seattle um Lake Washington
- Durchschnittliches Einkommen von Medina und Windermere wird sehr unterschiedlich sein
- Grund: Bill Gates lebt in Medina
- Grundsätzlich: Für das mittlere Einkommen wird praktisch immer der Median genommen und nicht der Durchschnitt, da dies gerechter ist
- Zügelt Bill Gates von Medina nach Windermere, so ändert sich das Durchschnitteinkommen, aber niemand hat was davon

Bemerkung

- Median: Auch *Zentralwert* oder *mittlerer Wert* (nicht zu verwechseln mit dem Mittelwert) genannt
- Exakte Interpretation des Medians ist noch erstaunlich schwierig
- Für uns ausreichend: Hälfte der Werte kleiner oder gleich und die andere Hälfte grösser oder gleich dem Median

Quartile

- Median: Wert, wo die Hälfte der Beobachtungen kleiner (oder gleich) wie dieser Wert sind
- Analoge "Uberlegung: 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: Waage

- Waage A hat $n = 13$ Messpunkte $\rightarrow$ 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

- Waage B: 25 % der Werte 2 $\rightarrow$ ganze Zahl
- Man *wählt* Durchschnitt von $x_{(2)}$ und $x_{(3)}$ als unteres Quartil
- Dann sind 2 Beoachtungen kleiner und 6 Beobachtungen grösser als dieser Wert

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

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

- Das untere Quartil der Methode B ist also 79.96

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 aufgerundet, abgerundet oder gerundet wird
- Aber: Es gibt keine einheitliche Definition für die Quartile

- Software **R** kennt keine eigenen Befehle für die Quartile
- Allgemeinerer Befehl **quantile** (Quantile kommen gleich)
- **R**: Quartile nach unserer Definition $\rightarrow$ Option **type=2**

```r
# Syntax für das untere Quartil: p=0.25

quantile(waageA, p = 0.25, type = 2)
##   25%
## 80.02
quantile(waageB, p = 0.25, type = 2)
##   25%
## 79.96
# Syntax für das obere Quartil: p=0.75

quantile(waageA, p = 0.75, type = 2)
##   75%
## 80.04
```


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

$$80.04 - 80.02 = 0.02$$

- R: Quartilsdifferenz (interquartile range) $\rightarrow$ IQR

```
IQR(waageA, type=2)
## [1] 0.02
```

- Die (oder ungefähr die) Hälfte der Messwerte liegt also in einem Bereich der Länge 0.02
- Beim Boxplot: Anschauliche Interpretation der Quartilsdifferenz

Beispiel

- Schulklasse mit 24 Lernenden: Noten an einer Prüfung:

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

- Berechnen mit R die Quartile und die Quartilsdifferenz:

```
noten <- 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, p = c(0.25, 0.75), type = 2)
##  25%  75%
## 3.80 5.35
IQR(noten, type = 2)
## [1] 1.55
```

- Hälfte der Lernenden liegen innerhalb von 1.55 Noten, nämlich zwischen 3.8 und 5.35
- 25 % der Klasse 3.8 oder weniger; rund 25 % der Klasse 5.35 und mehr

Quantile

- Quartile auf jede andere Prozentzahl verallgemeinern $\rightarrow$ Quantile
- 10 %-Quantil: Wert, wo 10 % der Werte kleiner oder gleich und 90 % der Werte grösser oder gleich diesem Wert sind
- Definition analog wie bei Quartilen
- Median ist 50 %–Quantil
- 25 %–Quantil ist unteres Quartil
- 75 %–Quantil ist oberes Quartil

- R: 10 %– und 70 %–Quantil der Waage A:

```
quantile(waageA, p = .1, type = 2)
##   10%
## 79.98
quantile(waageA, p = .7, type = 2)
##   70%
## 80.04
```

- Knapp 10 % der Messwerte sind kleiner oder gleich 79.97
- Entsprechend: Knapp 70 % der Messwerte kleiner oder gleich 80.04

Beispiel

- Noten an Prüfung in Schulklasse mit 24 Lernenden:

  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

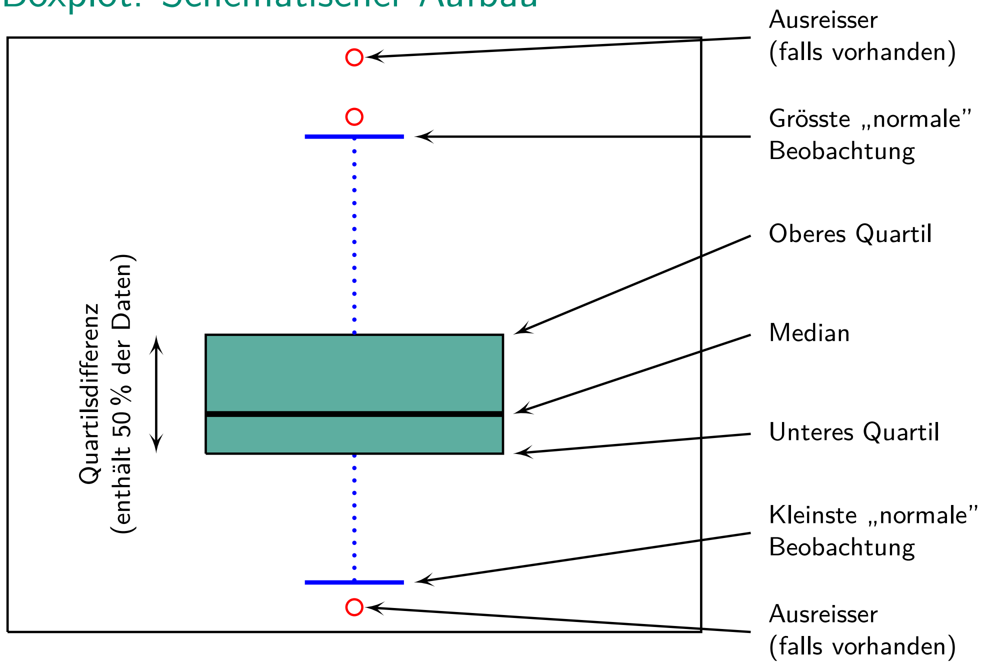
- Verschiedene Quantile mit **R**:

```
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, p = seq(from = .2, to = 1, by = .2), type = 2)
##  20%  40%  60%  80% 100%
##  3.6  4.2  5.0  5.6  6.0
```

- D.h.: Knapp 20 % der Lernenden sind schlechter als 3.6
- Genau 20 % der Lernenden nicht möglich → 4.8 Lernende
- 60 %–Quantil: (Knapp) diese Anzahl Prozent der Lernenden waren schlechter oder gleich einer 5

Boxplot: Schematischer Aufbau



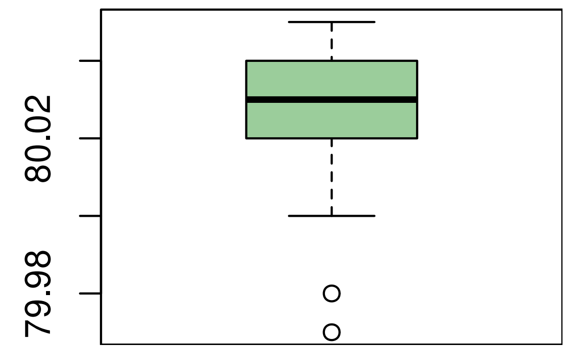
Boxplot: Schematischer Aufbau

- Rechteck, dessen Höhe vom empirischen 25 %- und vom 75 %-Quantil begrenzt wird → Box
- Horizontaler Strich für den Median in Box (schwarz)
- Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten „normalen" Wert führen (blau eingezeichnet)
  - Definition: „Normaler" Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt
- Ausreisser: Kleine Kreise (rot)

R

```
boxplot(waageA,
        col = "darkseagreen3"

)
```



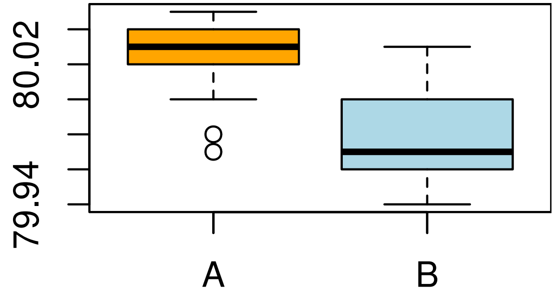
- Hälfte der Beobachtungen befindet sich zwischen dem oberen Quartil 80.04 und dem unteren Quartil 80.02, mit Quartilsdifferenz 0.02
- Median liegt bei 80.03
- Der „normale" Bereich der Werte liegt zwischen 80.00 und 80.05
- Zwei Ausreisser: 79.97 und 79.98
- Erste beiden Punkte: Früher berechnet
- Boxplot: Graphische Darstellung von Median und Quartilen

Vergleich von Datensätzen

- Boxplot: Darstellungen von verschiedenen Gruppen

```
boxplot(waageA, waageB,
        xlab = "Waage",
        col = c("orange", "lightblue")

)
axis(side = 1, at = c(1, 2), labels = c("A", "B"))
```



- Waage *A* grössere Werte als Waage *B* $\rightarrow$ Median von *A* grösser ist
- Daten von Waage *A* haben weniger Streuung als die Daten von Waage *B* $\rightarrow$ Rechteck weniger hoch (Quartilsdifferenz!)
- R-Code `axis(...)`: Siehe Skript