Mathe III

Ralf Herbrich

Explorative Datenanalyse

Überblick

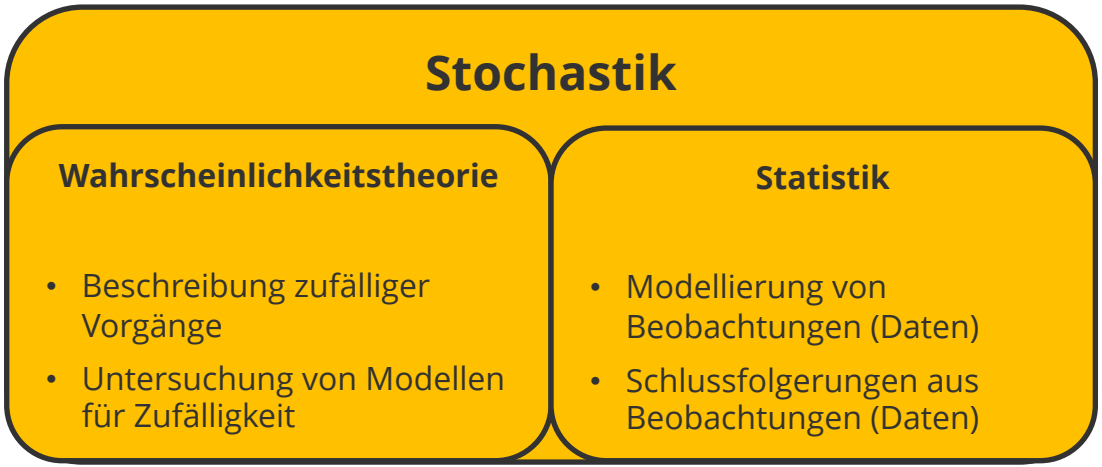
1. Empirische Daten
2. Visualisierung empirischer Daten

Überblick

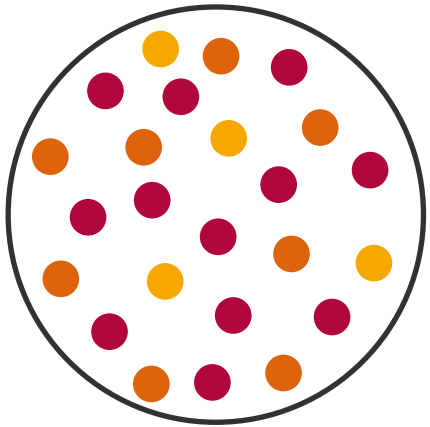
1. **Empirische Daten**
2. Visualisierung empirischer Daten

Wahrscheinlichkeitstheorie vs. Statistik

- **Wahrscheinlichkeitstheorie**: Wir haben uns bis jetzt hauptsächlich mit den Grundlagen der Wahrscheinlichkeitstheorie beschäftigt.
- **Statistik**: Statt nur Modelle zu beschreiben, wollen wir in der Statistik nun auch lernen, wie man reale Zufallsversuche, für die kein Modell bekannt ist, beschreiben kann.



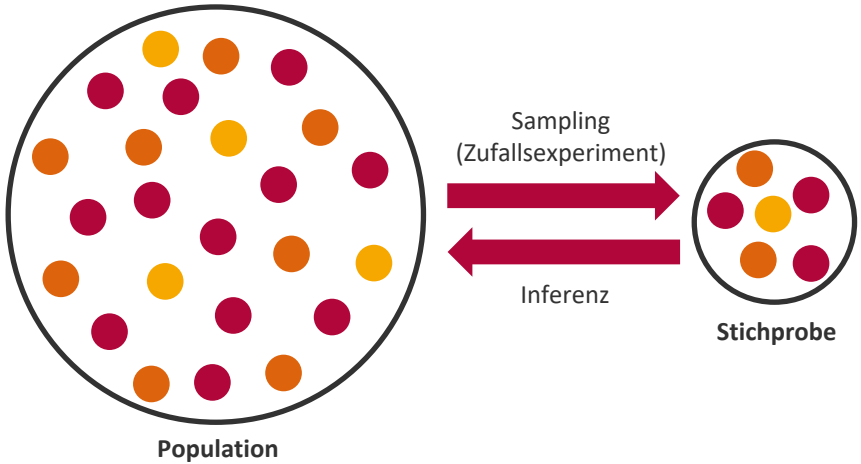
Population, Stichprobe und Modell



Population

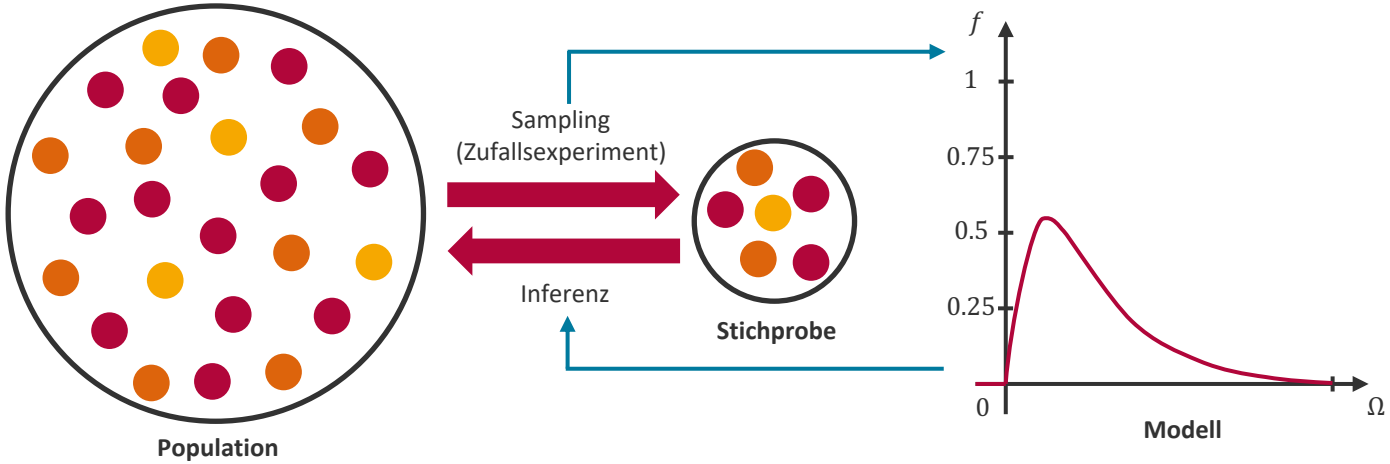
- Eine Population beschreibt eine **Grundgesamtheit** von Objekten, welche **Merkmale** tragen.
- Wir wollen Aussagen über die Merkmalsausprägung in der Population treffen.
- In der Regel ist die Population zu groß, um sie vollständig zu vermessen.
- Daher möchte man zumindest eine „zuversichtliche“ Aussage treffen können.

Population, Stichprobe und Modell



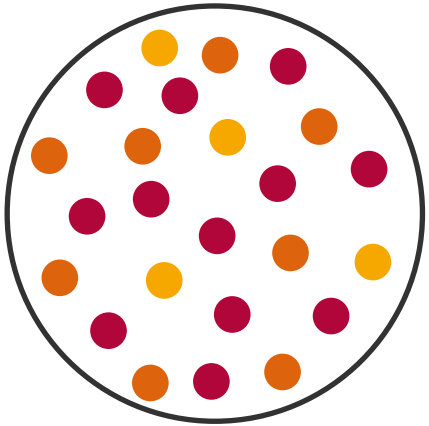
- Die **Stichprobe** ist eine Auswahl von Objekten aus einer Population.
- Die Auswahl der Stichprobe ist ein **Zufallsexperiment**.
- Wir können die Objekte der Stichprobe mit Kennzahlen charakterisieren (Deskriptive Statistik)
- Wir wollen damit Aussagen über die Population treffen (**Inferenz**).

Population, Stichprobe und Modell



- Das **stochastische Modell** beschreibt das Zufallsexperiment der Stichprobenwahl.
- Wir nutzen dieses, um von Stichprobenmerkmalen auf die Populationsmerkmale zu schließen.
- Beim Aufstellen eines Modells trifft man Annahmen – stimmen diese nicht, sind die Ergebnisse hinfällig.

Probleme von Stichproben



Population

Repräsentative Stichprobe

Nicht-repräsentative Stichprobe

- Damit die Eigenschaften einer Stichprobe generalisierbar auf die Population sind, muss sie **repräsentativ** sein.
- Nicht-repräsentative Stichproben besitzen nur eine eingeschränkte Aussagekraft.

Statistiken: Mittelwert

- **Definition (Mittelwert)**. Für eine Sequenz $\boldsymbol{x}$ von Beobachtungen $x_1, \ldots, x_n \in \mathbb{R}$ ist der **Mittelwert** $\overline{\boldsymbol{x}}$ definiert als

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

- **Beispiel (Mittelwert)**. Wir betrachten die Neupreise einer zufälligen Stichprobe von 4 Neuwagen:

$\boldsymbol{x}$	Gezogenes Modell	Neupreis in 1000 €
x_1	BMW i5	72
x_2	smart #1	42
x_3	VW ID.3	40
x_4	Mercedes EQA	50

Hier beträgt der Mittelwert

$$\overline{\boldsymbol{x}} = \frac{1}{4} \cdot (72 + 42 + 40 + 50) = \frac{204}{4} = 51$$

Statistiken: Quantile und Median

- **Definition (Quantil)**. Für eine Sequenz $\boldsymbol{x}$ von Beobachtungen $x_1, \dots, x_n \in \mathbb{R}$ und ein $p \in (0, 1)$ bezeichnen wir die Beobachtung an der Position $\lceil n \cdot p \rceil$ nach Sortieren der Werte $x_1, \dots, x_n$ als p-Quantil.

- **Definition (Median)**. Für eine Sequenz $\boldsymbol{x}$ von Beobachtungen $x_1, \dots, x_n \in \mathbb{R}$ bezeichnen wir das $\frac{1}{2}$-Quantil als Median $\tilde{\boldsymbol{x}}$ von $\boldsymbol{x}$.

- **Beispiel (Median und Quantil)**. Sei $\boldsymbol{x} = (1, 2, 3, 4, 5)$.
  - Der Median $\tilde{\boldsymbol{x}}$ von $\boldsymbol{x}$ ist 3.
  - Das 0.25-Quantil von $\boldsymbol{x}$ ist 2.
  - Das 0.99-Quantil von $\boldsymbol{x}$ ist 5.

Statistiken: Quantile und Median

- **Bemerkungen (Quantile)**
  - Es existieren verschiedene Definitionen des Quantils, z. B. als arithmetisches Mittel zwischen dem größten Wert der $\lceil n \cdot p \rceil$ kleinsten Werte und dem kleinsten Wert der $\lfloor n \cdot (1-p) \rfloor$ größten Werte.
  - Die von uns verwendete Definition wird auch Untermedian genannt.
  - Während der Mittelwert **ausreißeranfällig** ist, sind Quantile **robust**.

- **Beispiel (Robustheit des Medians)**. Sei $\boldsymbol{x} = (1, 2, 3, 4, 5)$ und $\boldsymbol{y} = (1, 2, 3, 4, 100)$
  - Für die Mittelwerte der Stichproben gilt $\overline{\boldsymbol{x}} = 3$ und $\overline{\boldsymbol{y}} = 22$.
  - Der Median für beide Stichproben hingegen ist unverändert 3.
  - Der Ausreißer in $\boldsymbol{y}$ hat demnach starken Einfluss auf den Mittelwert, jedoch kaum (bzw. hier keinen) Einfluss auf den Median.

Statistiken: *Median Absolute Deviation*

- **Definition (*Median Absolute Deviation* – MAD)**. Für eine Sequenz $\boldsymbol{x}$ von Beobachtungen $x_1, \dots, x_n \in \mathbb{R}$ ist die *median absolute deviation* (MAD) definiert als

$$\mathrm{MAD}(\boldsymbol{x}) = \mathrm{median}(|x_1 - \tilde{\boldsymbol{x}}|, |x_2 - \tilde{\boldsymbol{x}}|, \dots, |x_n - \tilde{\boldsymbol{x}}|)$$

- **Beispiel (*Median Absolute Deviation*)**. Sei $\boldsymbol{x} = (1, 2, 3, 4, 5)$.
  - Der Median $\tilde{\boldsymbol{x}}$ von $\boldsymbol{x}$ ist 3.
  - Die absoluten Abweichungen zum Median betragen (2, 1, 0, 1, 2).
  - Folglich gilt $\mathrm{MAD}(\boldsymbol{x}) = 1$.
- **Bemerkungen (*Median Absolute Deviation*)**
  - Der MAD gilt ebenso wie der Median als besonders robustes Maß einer Stichprobe.
  - Es gibt andere Definitionen für „MAD“, welche stattdessen den Mittelwert (Mean) nutzen – hier besteht Verwechslungsgefahr!

$$\overline{(|x_i - \overline{\boldsymbol{x}}|)}$$

Statistiken: Empirische Varianz, Kovarianz und Korrelation

- **Definition (Empirische Varianz und Standardabweichung)**. Für eine Sequenz $\boldsymbol{x}$ von Beobachtungen $x_1, \ldots, x_n \in \mathbb{R}$ heißt

$$V[\boldsymbol{x}] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{\boldsymbol{x}})^2$$

  die empirische Varianz von $\boldsymbol{x}$, und $\sqrt{V[\boldsymbol{x}]}$ die empirische Standardabweichung von $\boldsymbol{x}$.

- **Definition (Empirische Kovarianz und Korrelation)**. Für die Sequenzen $\boldsymbol{x}$ und $\boldsymbol{y}$ von Beobachtungen $x_1, \ldots, x_n \in \mathbb{R}$ und $y_1, \ldots, y_n \in \mathbb{R}$ heißt

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{\boldsymbol{x}}) \cdot (y_i - \overline{\boldsymbol{y}})$$

  die empirische Kovarianz von $\boldsymbol{x}$ und $\boldsymbol{y}$. Die empirische Korrelation ergibt sich als

$$\mathrm{Corr}[\boldsymbol{x}, \boldsymbol{y}] = \frac{\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}]}{\sqrt{V[\boldsymbol{x}] \cdot V[\boldsymbol{y}]}}$$

Statistiken: Empirische Varianz, Kovarianz und Korrelation

- **Beispiel (Empirische Korrelation)**. Wir betrachten die Neupreise und Reichweiten einer zufälligen Stichprobe $\boldsymbol{x}$ und $\boldsymbol{y}$ von 4 Neuwagen:

Gezogenes Modell	Neupreis in 1000 €	Reichweite in km	$(x_i - \overline{x})$	$(y_i - \overline{y})$
BMW i5	$x_1 = 72$	$y_1 = 570$	21	45
smart #1	$x_2 = 42$	$y_2 = 440$	-9	-85
VW ID.3	$x_3 = 40$	$y_3 = 560$	-11	35
Mercedes EQA	$x_4 = 50$	$y_4 = 530$	-1	5

  - Es gilt $\overline{\boldsymbol{x}} = \frac{1}{4} \cdot (72 + 42 + 40 + 50) = 51$ und $\overline{\boldsymbol{y}} = \frac{1}{4} \cdot (570 + 440 + 560 + 530) = 525$.
  - Daher gilt

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \frac{1}{4} \cdot (21 \cdot 45 + (-9) \cdot (-85) + (-11) \cdot 35 + (-1) \cdot 5) = 330$$

  - Es gilt $V[\boldsymbol{x}] = \frac{1}{4} \cdot (21^2 + 9^2 + 11^2 + 1^2) = 161$ und $V[\boldsymbol{y}] = \frac{1}{4} \cdot (45^2 + 85^2 + 35^2 + 5^2) = 2625$.
  - Daher gilt

$$\mathrm{Corr}[\boldsymbol{x}, \boldsymbol{y}] = \frac{330}{\sqrt{161 \cdot 2625}} \approx 0.5$$

Empirische Kenngrößen

- **Bemerkungen (Empirische Kenngrößen)**
  - Die neu eingeführten Kenngrößen können genutzt werden, um gezogene Stichproben aus einer Population mit einer unbekannten Verteilung zu beschreiben.
  - Die Kenngrößen sind die äquivalenten Momente von ein- und mehr-dimensionalen Zufallsvariablen wenn die **empirische Verteilung** benutzt wird mit

$$p_{\{x_1,\dots,x_n\}}(x) = \frac{|\{x_i \in \{x_1, \dots, x_n\} \mid x_i = x\}|}{n}$$

  - Mithilfe dieser Kenngrößen können Rückschlüsse auf die tatsächliche Verteilung getroffen werden, z. B.
    - Mittelwert, Median → Erwartungswert
    - Empirische Varianz → Varianz
    - Empirische Standardabweichung, MAD → Standardabweichung
    - Empirische Kovarianz → Kovarianz
    - Empirische Korrelation → Korrelation

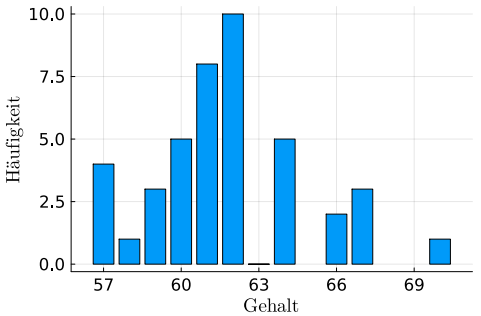
Überblick

1. Empirische Daten
2. **Visualisierung empirischer Daten**

Stichprobenvisualisierung

- **Problem**: Wie können wir einen **schnellen** Überblick über charakteristische Eigenschaften der Stichprobe bekommen?
  - Mittelwert
  - Streuung
  - Symmetrie oder Schiefe
  - Ausreißer
  - Verteilungsannahmen
- **Lösung**: Mittels Visualisierung der Stichprobe („ein Bild sagt mehr als tausend Worte“)
  - Schnelle Qualitätskontrollen für automatisierte Verfahren
  - Eigenschaften der Stichprobe an Endnutzer kommunizieren
- Während es sehr viele Visualisierungsarten gibt, werden wir nur eine Auswahl davon behandeln.

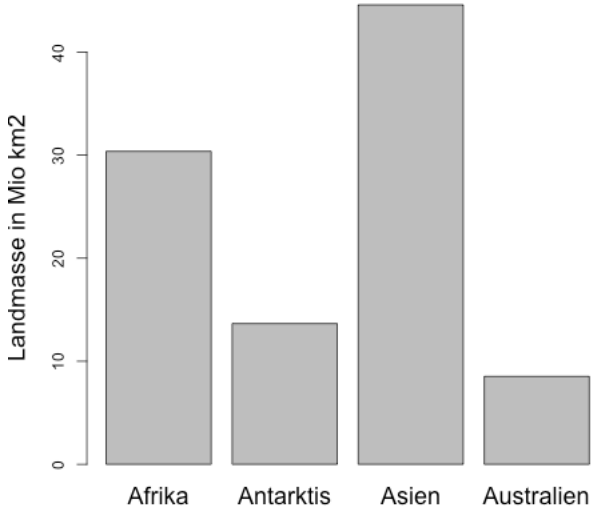
Gehalt (in Tausend €)	Häufigkeit
57	4
58	1
59	3
60	5
61	8
62	10
63	0
64	5
66	2
67	3
70	1



Säulendiagramm (*bar plot*)

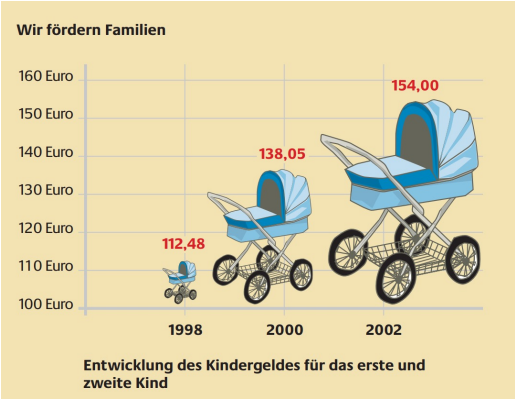
- Ein **Säulendiagramm** ist die Darstellung einer kontinuierlichen Metrik für verschiedene Objekte oder Objektkategorien in der Stichprobe.
- Es eignet sich zum Vergleich der Objekte bzw. Objektkategorien in dieser Metrik.
- **Beispiel (Säulendiagramm)**

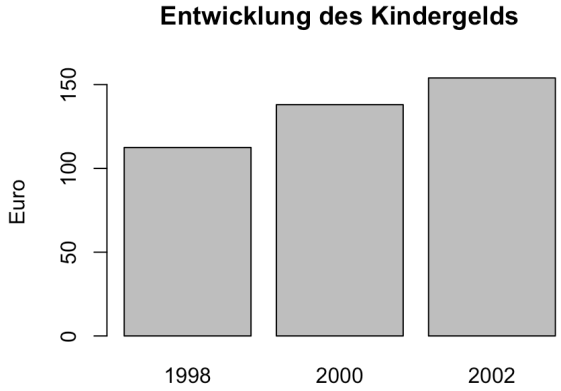
```
data <- data.frame(
  continent = c("Afrika", "Antarktis", "Asien", "Australien"),
  landmass = c(30.37, 13.66, 44.58, 8.53)
)

barplot(height = data$landmass, names = data$continent,
        ylab = "Landmass in Mio km2")
```



Säulendiagramm

- **Achtung**: Säulendiagramme können den Betrachter täuschen, wenn der Achsenschnittpunkt schlecht platziert ist oder die Säulen unterschiedlich breit sind.
- **Beispiel (Täuschung im Säulendiagramm)**





Agenda 2010 – Deutschland bewegt sich. Antworten zur Agenda 2010. 4. Auflage. Presse- und Informationsamt der Bundesregierung, Berlin 2003.

Histogramm

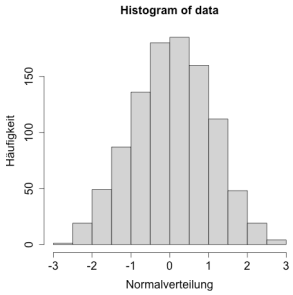
- Ein Säulendiagramm kommt schnell an seine Grenzen, falls
  - die Objektkategorien aus einer zu großen Menge (>10 Kategorien) stammen.
  - die Objektkategorien aus dem Raum der Zahlen kommen (d.h. Distanzen haben Bedeutung)
- Hierbei können Histogramme helfen.
  - Es zählt, wie oft eine Größe vorkommt und bündelt die Ergebnisse in Klassen (*bins*).
  - Ein Histogramm wird durch Ursprung x_0 und Klassenbreite (*bin width*) h bestimmt.
  - Für alle $i \in \mathbb{Z}$ zählt die i-te Klasse die Vorkommen im Intervall

    $$[x_0 + i \cdot h, x_0 + (i+1) \cdot h)$$

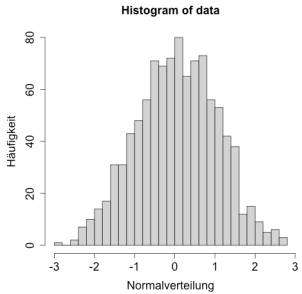
  - Jede Klasse ist eine Säule über ihrem Intervall mit der Häufigkeit als Höhe.

```
data <- rnorm(1000, mean = 0, sd = 1)

hist(data, xlab = "Normalverteilung",
     ylab = "Häufigkeit", breaks = 10)
```

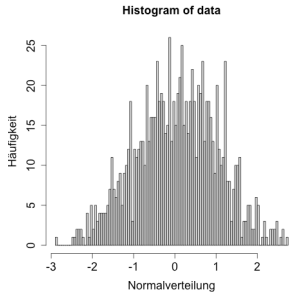
10 bins



30 bins



100 bins

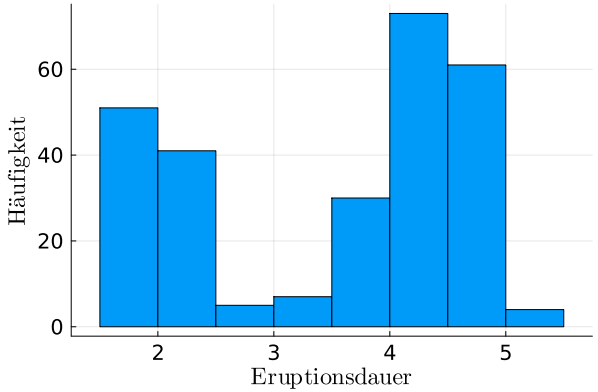


Histogramm

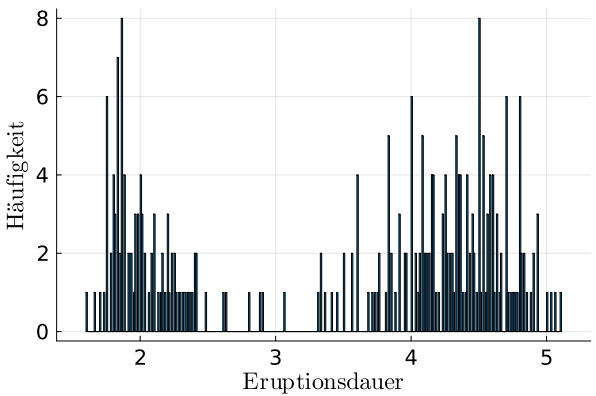
- **Achtung:** Ein Histogramm kann degenerieren, wenn
  1. die Klassenbreite zu schmal ist und viele Klassen kein oder wenige Vorkommen enthalten
  2. die Klassenbreite zu weit ist und eine zu starke Bündelung die Aussagekraft trübt.

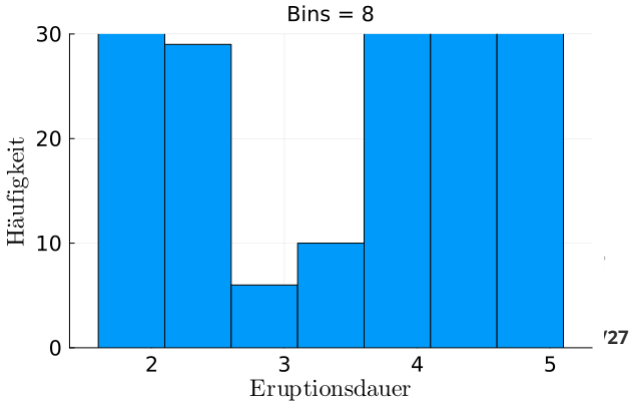
- **Beispiel (Verschiedene Klassenbreiten im Histogramm)**. Eruptionsdauer des Old Faithful Geysirs im Yellowstone-Nationalpark (in Minuten)

8 bins



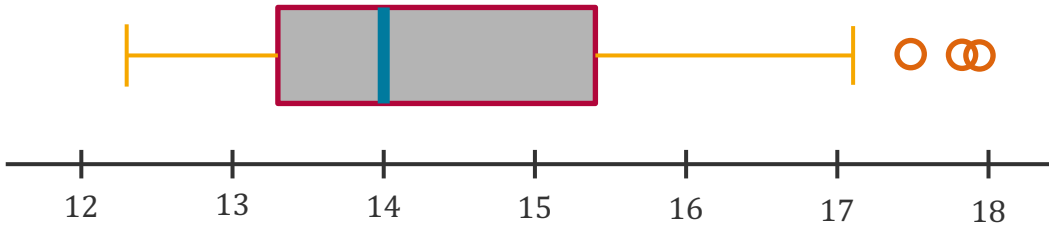
500 bins





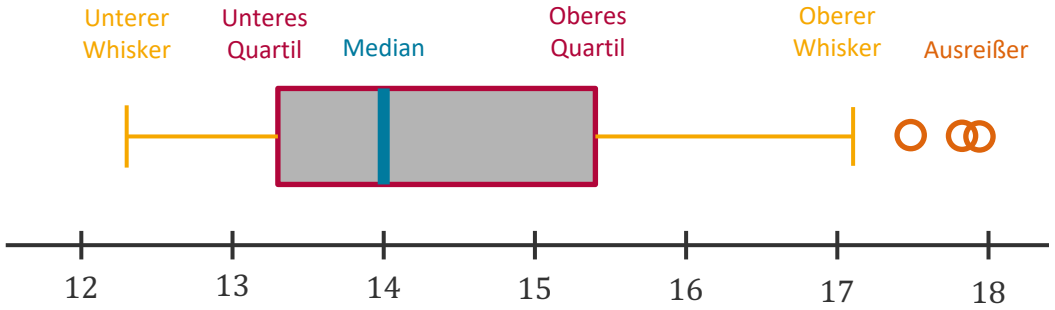
Boxplot

- Für eine besonders kompakte Darstellung der Verteilung von Werten kann ein Boxplot verwendet werden.
- Ein Boxplot enthält weniger Detailstufen als ein Histogramm, allerdings sind Kenngrößen wie Median oder Quantil auf einen Blick ersichtlich.
- Damit können mehrere Verteilungen effizient bezüglich dieser Kenngrößen verglichen werden.



Boxplot

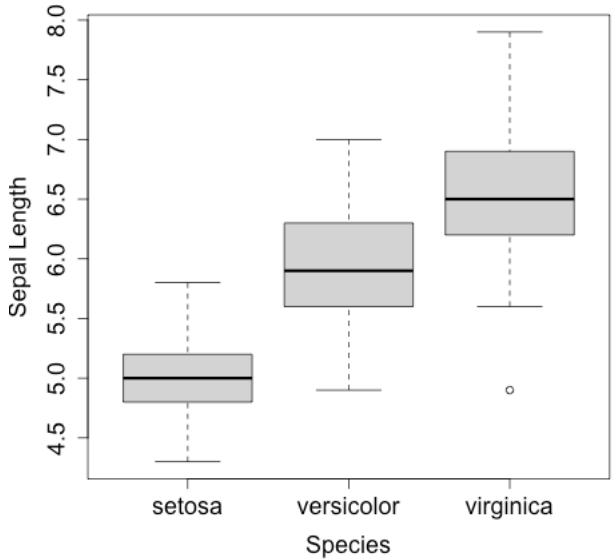
- Das untere **Quartil** entspricht dem 0.25-Quantil, das obere dem 0.75-Quantil.
- Die **Whisker** sind im Allgemeinen so lang wie die Wertspanne, jedoch nicht länger als der anderthalbfache Interquartilabstand.
- Werte außerhalb der Whisker (weiter als der anderthalbfache Interquantilabstand vom Median entfernt) werden als **Ausreißer** gekennzeichnet.
- **Bemerkung**. Die Definition der Whisker ist in der Literatur nicht konsistent.



Boxplot

- **Beispiel (Vergleich von Daten mit Boxplots)**
  - Im Vergleich dargestellt sind drei Boxplots der Messwerte von Kelchblattlänge in Zentimetern für die Schwertlilien-Spezien *Iris setosa* , *Iris versicolor* und *Iris virginica.*
  - Auch wenn die Wertebereiche nicht disjunkt sind, ist eine Tendenz klar erkennbar.

```
data(iris)

boxplot(iris$Sepal.Length ~ iris$Species,
        xlab = "Species",
        ylab = "Sepal Length")
```



Iris setosa

Iris versicolor

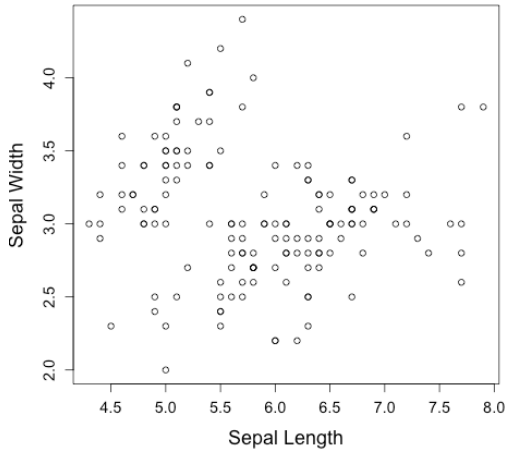
Iris virginica

Streudiagramm

- Ein Streudiagramm (*scatter plot*) stellt zwei (oder mehr) kontinuierliche Eigenschaften einer Zufallsbeobachtung dar, welche gleichzeitig aufgetreten sind.
- Damit können Korrelationen gut sichtbar gemacht werden.

- **Beispiel (Streudiagramm)**

```
data(iris)
plot(iris$Sepal.Length, iris$Sepal.Width,
     xlab = "Sepal Length",
     ylab = "Sepal Width")
```

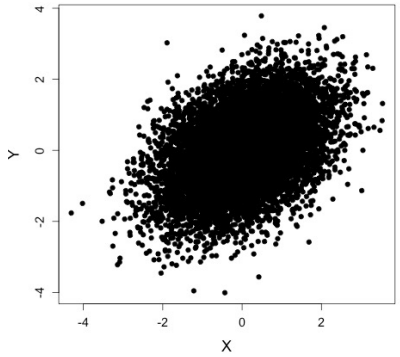


Streudiagramm

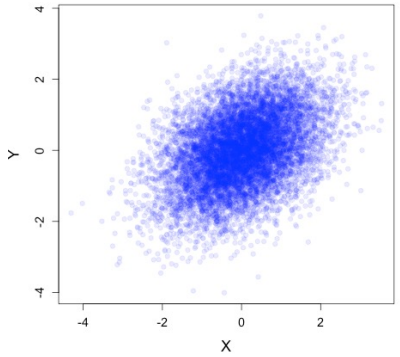
- **Achtung:** Überlagern sich mehrere Datenpunkte im Streudiagramm, kann die optische Wahrnehmung verfälscht werden.
  - Eine Darstellung der Messwert-Dichte statt der Einzelmesspunkte kann dies beheben.

- **Beispiel (Überlagerung in Streudiagrammen)**

```
plot(data[, 1], data[, 2],
     xlab = "X", ylab = "Y", pch=19)
```




```
plot(data[, 1], data[, 2],
     xlab = "X", ylab = "Y", pch=19,
     col = rgb(0, 0, 1, 0.1))
```



Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Viel Spaß bis zur nächsten Vorlesung!