Wahrscheinlichkeitsrechnung und Statistik für Studierende der Informatik

Vorlesung 12

Wahrscheinlichkeitsrechnung WiSe 24/25

Teilnehmer/innen

Auszeichnungen

Kompetenzen

Bewertungen

Allgemeines, Forum

Klausur

Skript & weitere Literatur

Übungen, Übungsblätter

Tutorien & LuDi

Wahrscheinlichkeitsrechnung und Statistik für Studierende der Informatik (WiSe 24/25)

Dashboard / Meine Kurse / Wintersemester 2024/2025 / Informatik / Human-centered Computing and Cognitive Science (HCCS) / Wahrscheinlichkeitsrechnung WiSe 24/25

Allgemeines, Forum

Ankündigungen

Online Umfrage zur Veranstaltung:

https://lehrevaluation.zhqe.uni-due.de/evasys/online.php?pswd=VYWWZ

Kursverwaltung

Neuen Kursraum beantragen
Kurs kopieren
Kurs verschieben
Kurs löschen

Hinweise

(English text below)

Grenzwertsätze

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$$

Binomialverteilung reloaded:

Satz 4.3 (Satz von de Moivre-Laplace). *Es sei* $0 < p < 1$ *und* X_n $B(n,p)$*-verteilt sowie* X_n^* *die zu* X_n *gehörende standardisierte Zufallsvariable. Dann gilt für alle* $a < b$*:*

$$\lim_{n\to\infty} P(a \leq X_n^* \leq b) = \phi(b) - \phi(a) \,,$$

wobei ϕ *(wie üblich) die Verteilungsfunktion der Standardnormalverteilung bezeichnet.*

Es folgt:

Ist X eine $B(n,p)$-verteilte Zufallsvariable, dann gilt

$$P(X \leq x) \approx \Phi\left(\frac{x - np + \frac{1}{2}}{\sqrt{np(1-p)}}\right)$$

Faustregel: gute Approximation, wenn $np(1-p) > 9$.

Grenzwertsätze

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$$

Satz 4.6 (Grenzwertsatz von Lindeberg-Levy). *Es sei $X_1, X_2, \ldots$ eine Folge (stochastisch) unabhängiger und identisch verteilter (kurz: u.i.v. oder i.i.d.) Zufallsvariabler mit $\sigma^2 = Var(X_1) > 0$. Setzen wir $\mu := E(X_1)$ und $S_n := X_1 + \ldots + X_n$, so gilt:*

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}} \leq b\right) = \phi(b) - \phi(a) \; .$$

$= S_n^*$

D.h.: große Summen von unabhängigen und gleichverteilten Zufallsvariablen sind immer normalverteilt!

Übersetzung:

$$P(S_n \leq x) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

Grenzwertsätze

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$$

Folgerung 4.7. Betrachten wir in Satz 4.6 das symmetrische Intervall $[-k, k]$, so erhalten wir wegen $E(S_n) = n \cdot E(X_1) = n \cdot \mu$ und $Var(S_n) = n \cdot Var(X_1) = n\sigma^2$ die Beziehung

$$\lim_{n\to\infty} P\left(-k \leq \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \leq k\right) = \phi(k) - \phi(-k) = 2 \cdot \phi(k) - 1 \ ,$$

d.h.

$$\lim_{n\to\infty} P\left(E(S_n) - k\sqrt{Var(S_n)} \leq S_n \leq E(S_n) + k\sqrt{Var(S_n)}\right) = 2 \cdot \phi(k) - 1 \ .$$

Mit Bemerkung 3.13 erhalten wir also, dass die Summe von n unabhängigen und identisch verteilten Zufallsvariablen (als Faustregel für große n) mit einer ungefähren Wahrscheinlichkeit von

- 0.6826 in den Grenzen $E(S_n) \pm 1 \cdot \sqrt{Var(S_n)}$
- 0.9544 in den Grenzen $E(S_n) \pm 2 \cdot \sqrt{Var(S_n)}$
- 0.9974 in den Grenzen $E(S_n) \pm 3 \cdot \sqrt{Var(S_n)}$

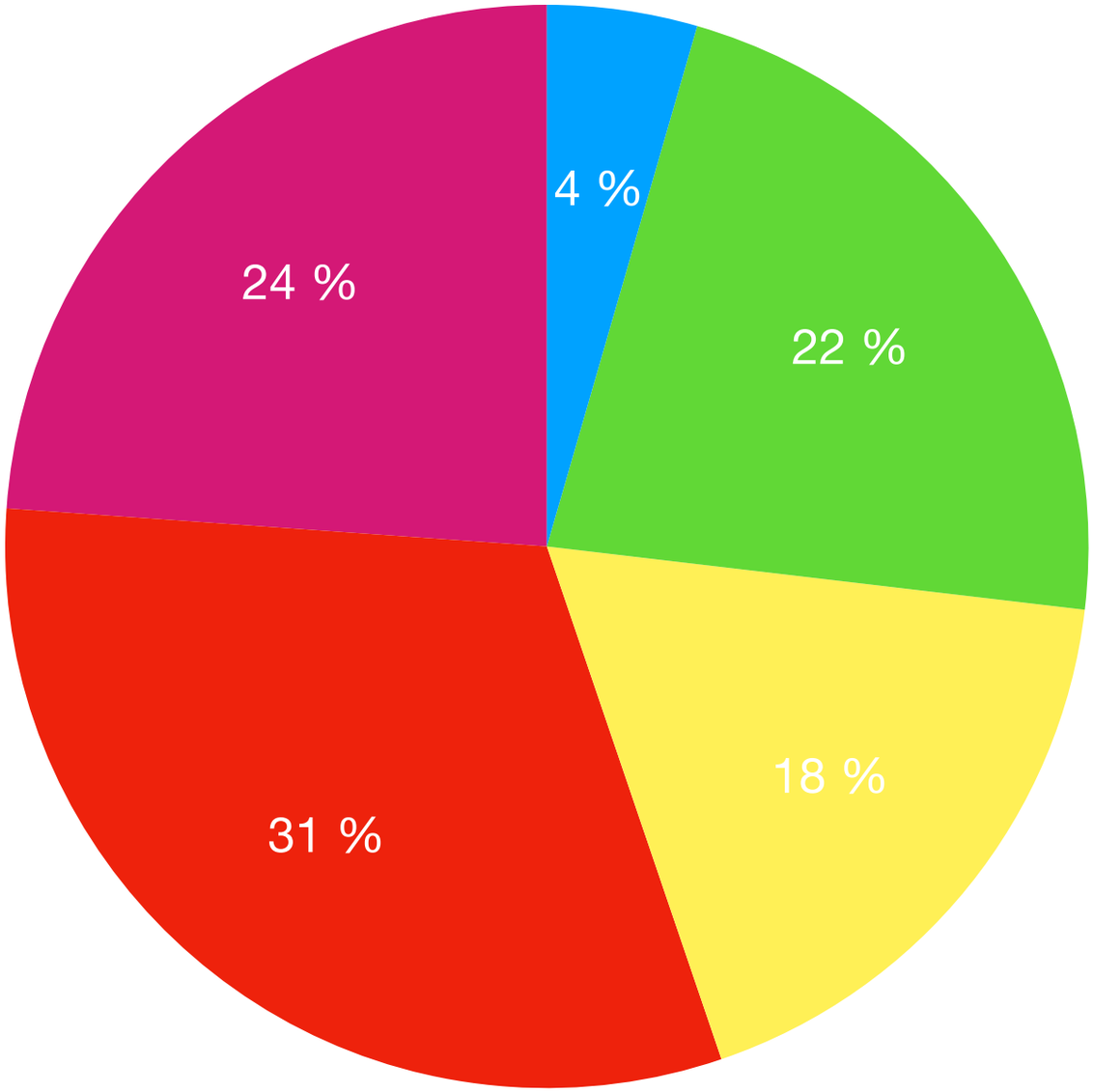
liegt.

Statistik

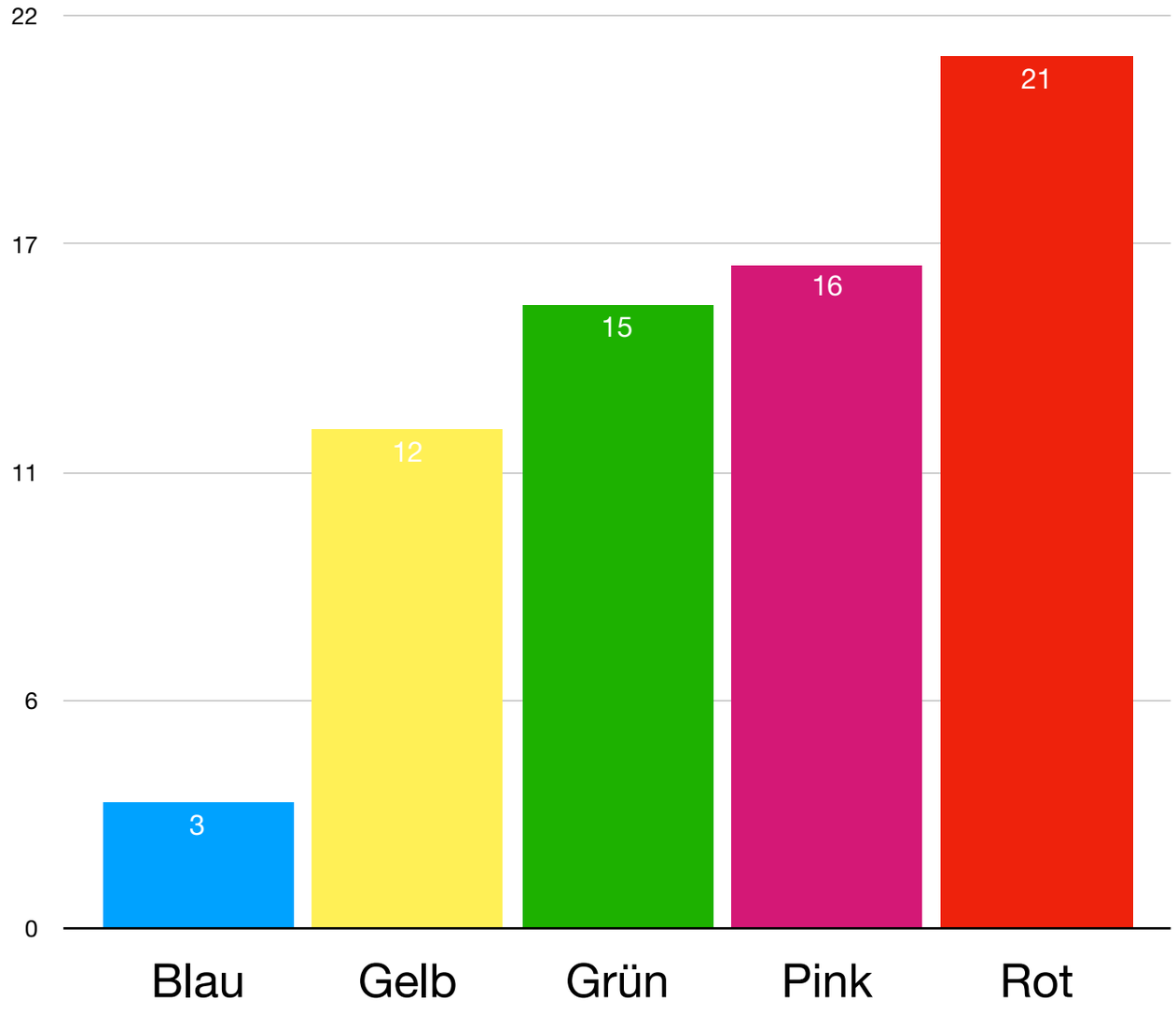
Darstellung von Daten

Lieblingsfarbe in einer Kindergartengruppe

Kreisdiagramm



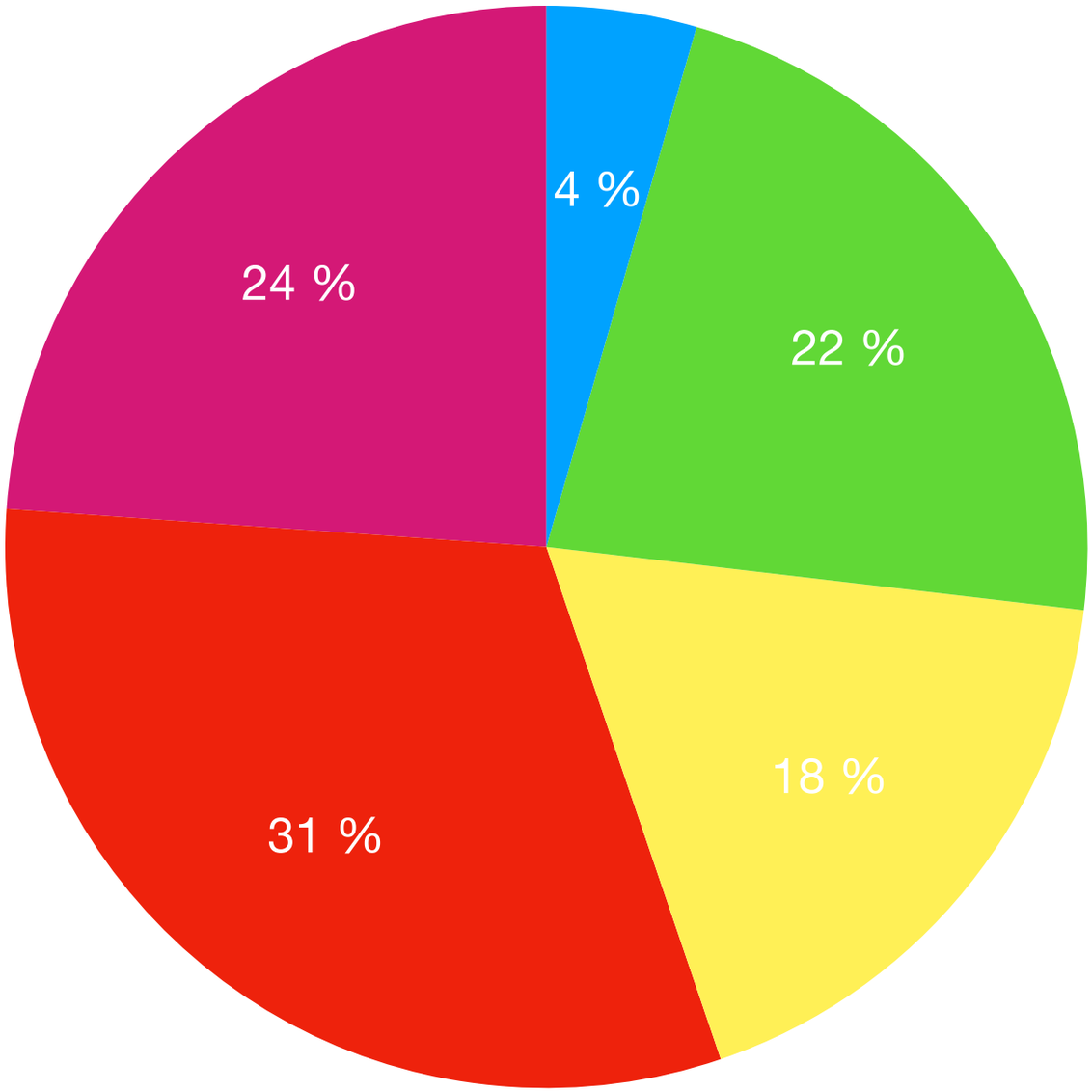
Säulendiagramm



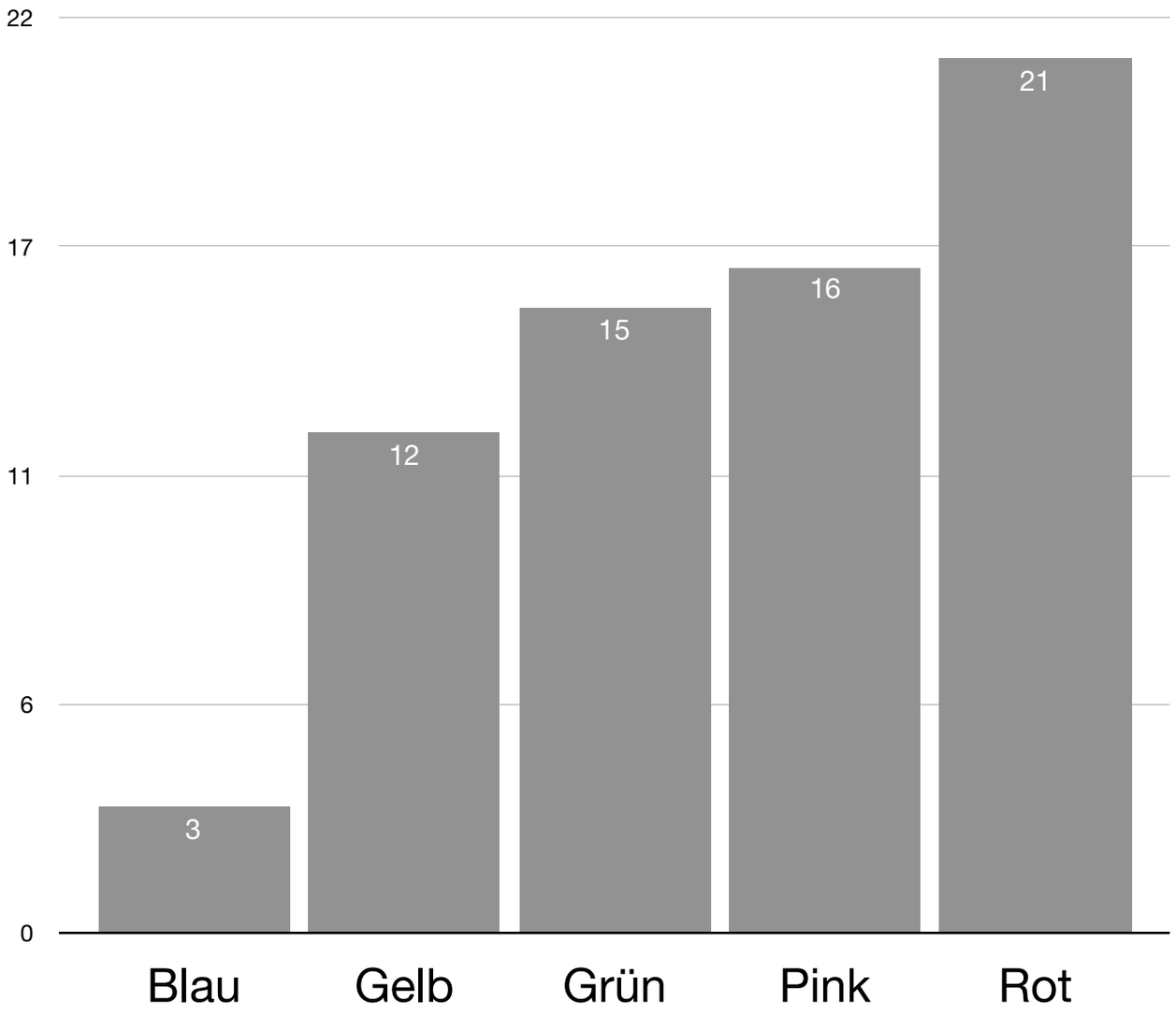
Darstellung von Daten

Lieblingsfarbe in einer Kindergartengruppe

Kreisdiagramm



Säulendiagramm



Darstellung von Daten

Mit Farben lässt sich schlecht rechnen. Sie lassen sich auch nicht ordnen

Rot > Grün oder Grün > Rot?

Wir interessieren uns im Folgenden für quantitative Daten, wie Körpergröße, Anzahl von …

Darstellung von Daten

Definition 5.5. Wir betrachten die Daten $x_1, \ldots, x_n$. Wir teilen diese Daten in s disjunkte Klassen auf, indem wir s halboffene Intervalle

$$[a_1, a_2[, [a_2, a_3[, \ldots, [a_s, a_{s+1}[$$

mit $a_1 < a_2 < \ldots < a_{s+1}$ betrachten, in denen alle Daten liegen. Nun bilden wir über jedem Teilintervall $[a_i, a_{i+1}[$ ein Rechteck der Höhe d_i, wobei

$$d_i \cdot (a_{i+1} - a_i) = k_i \quad \text{für } 1 \leq i \leq s$$

ist mit

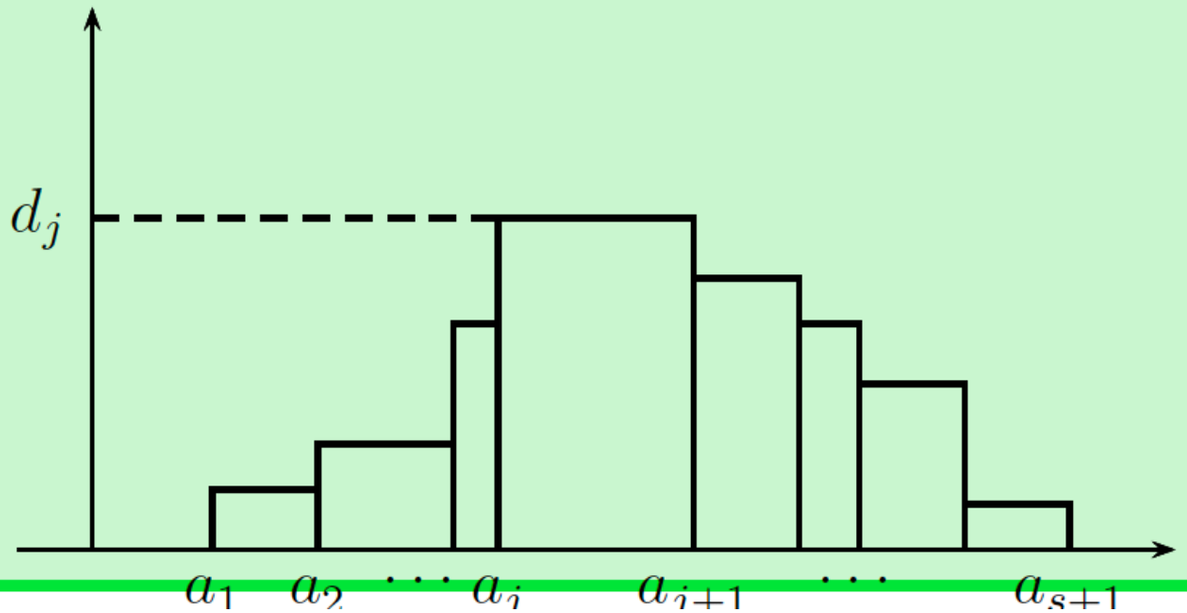
$$k_i = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{a_i \leq x_j < a_{i+1}\}}$$

Wahrscheinlichkeit, dass ein zufällig gewähltes x_k im Intervall $[a_i, a_{i+1})$ liegt.

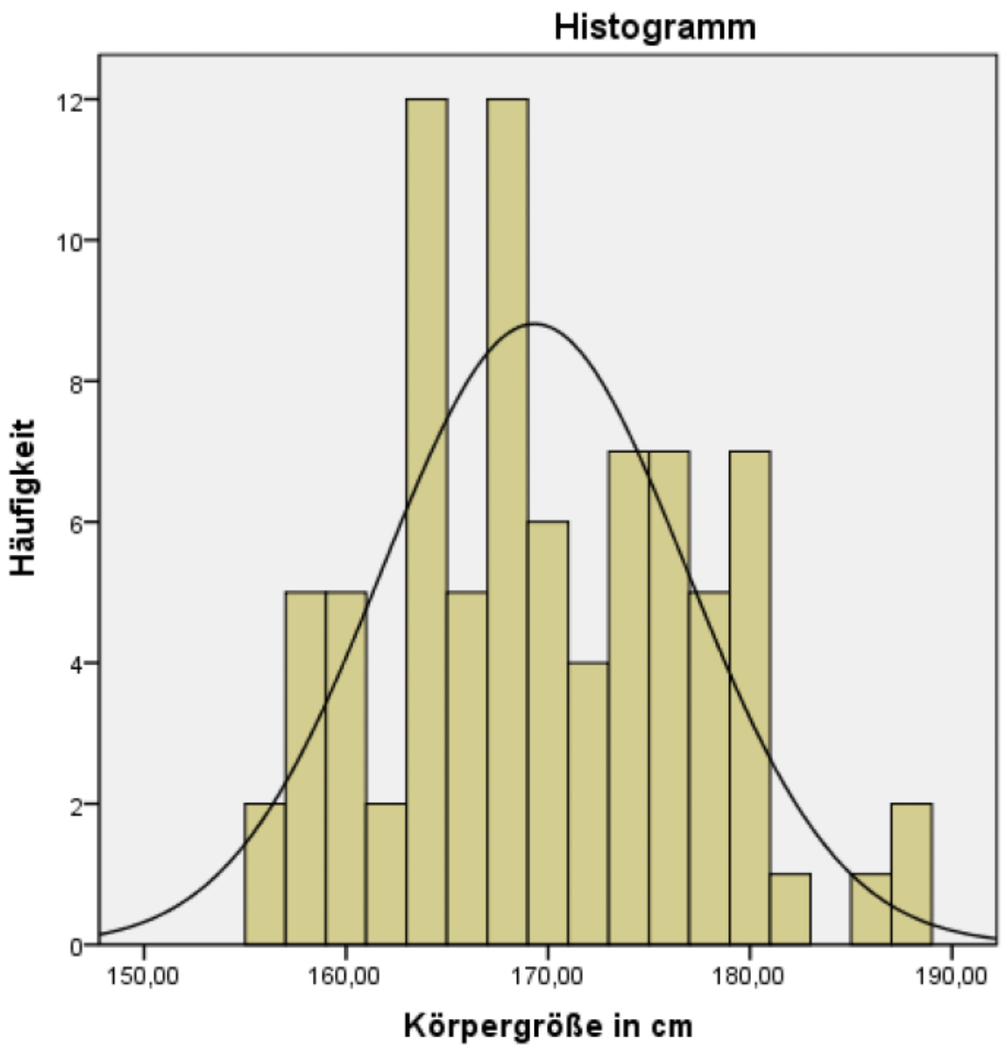
Anzahl von Daten in $[a_i, a_{i+1})$

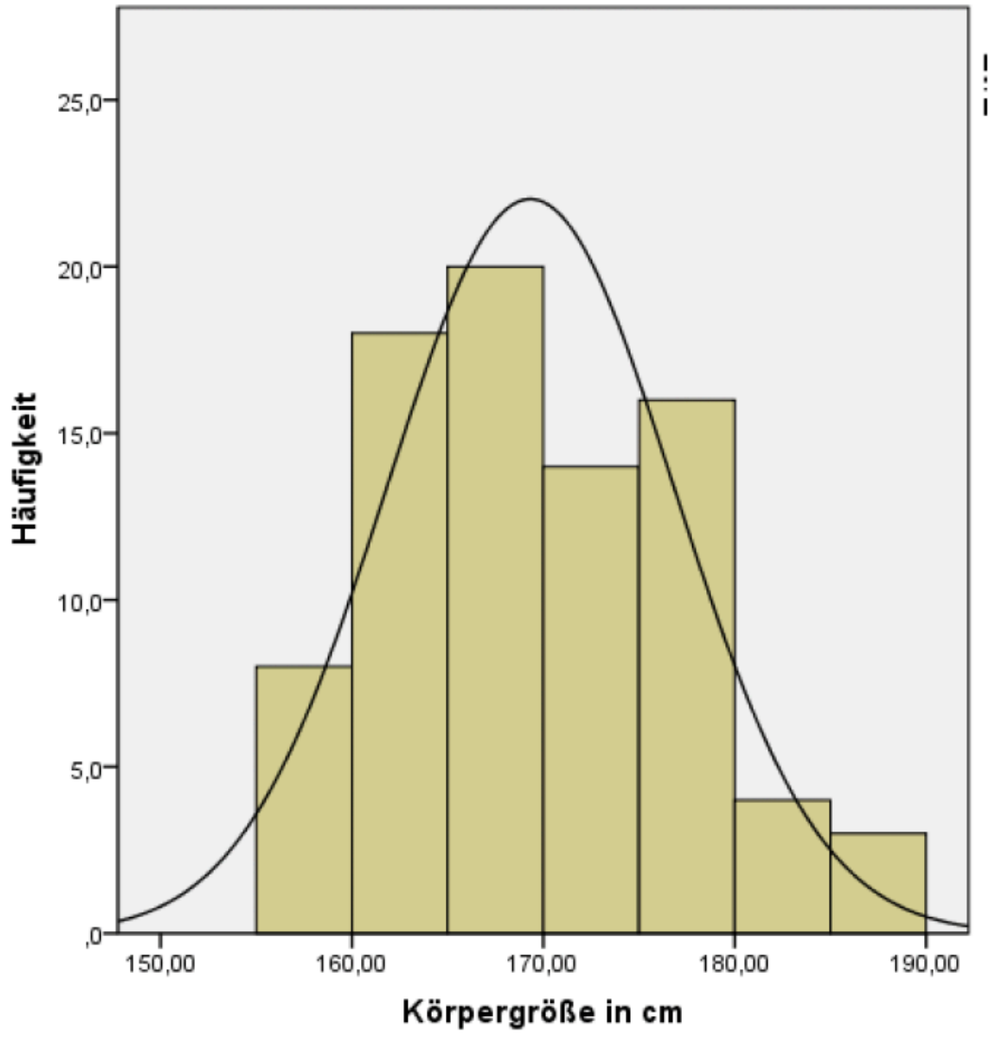
Dadurch entsteht ein *Histogramm*. Die Gesamtfläche aller Rechtecke ist 1. Die einzelnen Rechteckflächen sind proportional zur relativen Häufigkeit des Auftretens der Daten. (Dabei ist $\mathbf{1}$ die Indikatorfunktion.)

Darstellung von Daten



Darstellung von Daten





Diese Histogramme wurden mit dem gleichen Datensatz konstruiert.
Die Balkenbreite kann dem Wunschergebnis angepasst werden.

Darstellung von Daten

Allgemeines Setting:

Wir haben einen Wahrscheinlichkeitsraum (Ω, P). Die Menge Ω ist die Grundpopulation.

Auf diesem Wahrscheinlichkeitsraum haben wir eine Zufallsvariable X, die jedem $\omega \in \Omega$ einen Wert zuweist. Die Verteilung, Erwartungswert, Varianz, …. von X ist oft nicht bekannt.

Bsp.: Ω sind die Studierenden an der UDE, und X weist jedem $\omega \in \Omega$ die Körpergröße in cm zu.

Wir betrachten nun n unabhängige Zufallsvariablen $X_1, \ldots, X_n$, die gleichverteilt sind wie X. Diese entsprechen einer n-maligen Auswertung einzelner Werte von X. Der Vektor $(X_1, \ldots, X_n)$ heißt Stichprobe von X.

Die Ergebnisse, die wir bei einer Auswertung der Stichprobe erhalten, sind n reelle Zahlen $(x_1, \ldots, x_n) = (X_1(\omega_1), \ldots, X_n(\omega_n))$. Diese Zahlenreihe nennen wir konkrete Stichprobe von X oder Realisierung von $(X_1, \ldots, X_n)$.

Darstellung von Daten

Definition 5.1. a) Die Zahl

$$\overline{x} := \overline{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$$

heißt *(Stichproben-)* oder *arithmetisches Mittel* bzw. kurz *Mittelwert* der Daten $x_1, \ldots, x_n$.

b) Die Zahl

$$s^2 := s_x^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

heißt *(Stichproben-)* oder *empirische Varianz* der Daten $x_1, \ldots, x_n$.
Die Zahl $s_x = \sqrt{s_x^2}$ heißt *(Stichproben-)* oder *empirische Standardabweichung* von $x_1, \ldots, x_n$.

Kennen wir $\overline{x}$ und $x_1, \ldots, x_{n-1}$, so kennen wir auch x_n. Daher teilen wir bei der Varianz durch $n-1$ „Freiheitsgrade“.

Darstellung von Daten

Definition 5.1. a) Die Zahl

$$\overline{x} := \overline{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$$

heißt *(Stichproben-)* oder *arithmetisches Mittel* bzw. kurz *Mittelwert* der Daten $x_1, \ldots, x_n$.

b) Die Zahl

$$s^2 := s_x^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

heißt *(Stichproben-)* oder *empirische Varianz* der Daten $x_1, \ldots, x_n$.
Die Zahl $s_x = \sqrt{s_x^2}$ heißt *(Stichproben-)* oder *empirische Standardabweichung* von $x_1, \ldots, x_n$.

Betrachten wir die Zufallsvariablen $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ und $S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$, so

ist $E(\overline{X}) = E(X)$ und $E(S^2) = Var(X)$.

Die Zahlen $\overline{x}$ und s^2 dienen also als „Schätzung“ für $E(X)$ und $Var(X)$.

Darstellung von Daten

Definition 5.3. a) Sortieren wir die Daten $x_1, \ldots, x_n$ der Größe nach, d.h. bilden wir $x_{(1)} = \min\limits_{1 \le i \le n} x_i$ bis $x_{(n)} = \max\limits_{1 \le i \le n} x_i$, so nennen wir

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

die *geordnete Stichprobe* zu $x_1, \ldots, x_n$. Die Differenz $x_{(n)} - x_{(1)}$ heißt *Stichprobenspannweite.*

b) Der *(empirische) Median* oder *Zentralwert* der Stichprobe $x_1, \ldots, x_n$ ist definiert durch

$$x_{1/2} := \begin{cases} x_{(\frac{n+1}{2})} & \text{für ungerades } n \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{für gerades } n \end{cases} .$$

Der Median von $|x_1 - x_{1/2}|, |x_2 - x_{1/2}|, \ldots, |x_n - x_{1/2}|$ heißt *Median-Abweichung* von $x_1, \ldots, x_n$.

Darstellung von Daten

c) Ist $0 < p < 1$, so heißt die Zahl

$$x_p := \begin{cases} x_{([np+1])} & \text{falls } n \cdot p \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & \text{falls } n \cdot p \in \mathbb{N} \end{cases} .$$

empirisches p-Quantil von $x_1, \ldots, x_n$. (Dabei ist $[y]$ die größte ganze Zahl kleiner oder gleich $y \in \mathbb{R}$.)

d) Die Quantile $x_{3/4}$ und $x_{1/4}$ heißen *oberes* bzw. *unteres Quartil*. Die Differenz $x_{3/4} - x_{1/4}$ heißt *Quartilsabstand* der $x_1, \ldots, x_n$.

Darstellung von Daten

Definition 5.7. Der *Box-Plot* wird häufig beim Vergleich verschiedener Stichproben verwandt. Er benutzt Quantile zur graphischen Darstellung von Lage und Streuung der Daten. Außerdem werden potentielle Ausreißer hervorgehoben.
Zur Anfertigung des Box-Plot wird ein senkrechtes oder waagerechtes Rechteck (eine *Kiste*) gezeichnet, die vom unteren bis zum oberen Quartil geht und beim Median unterteilt wird. Die Breite des Rechtecks wird meist nach ästhetischen Gesichtspunkten gewählt. Nach oben und unten bzw. links und rechts wird die Kiste durch zwei Stäbe verlängert, wobei der Endpunkt des nach oben aufgesetzten Stabes kleiner ist als das obere Quartil plus das 1,5-fache des Quartilsabstandes, also kleiner als

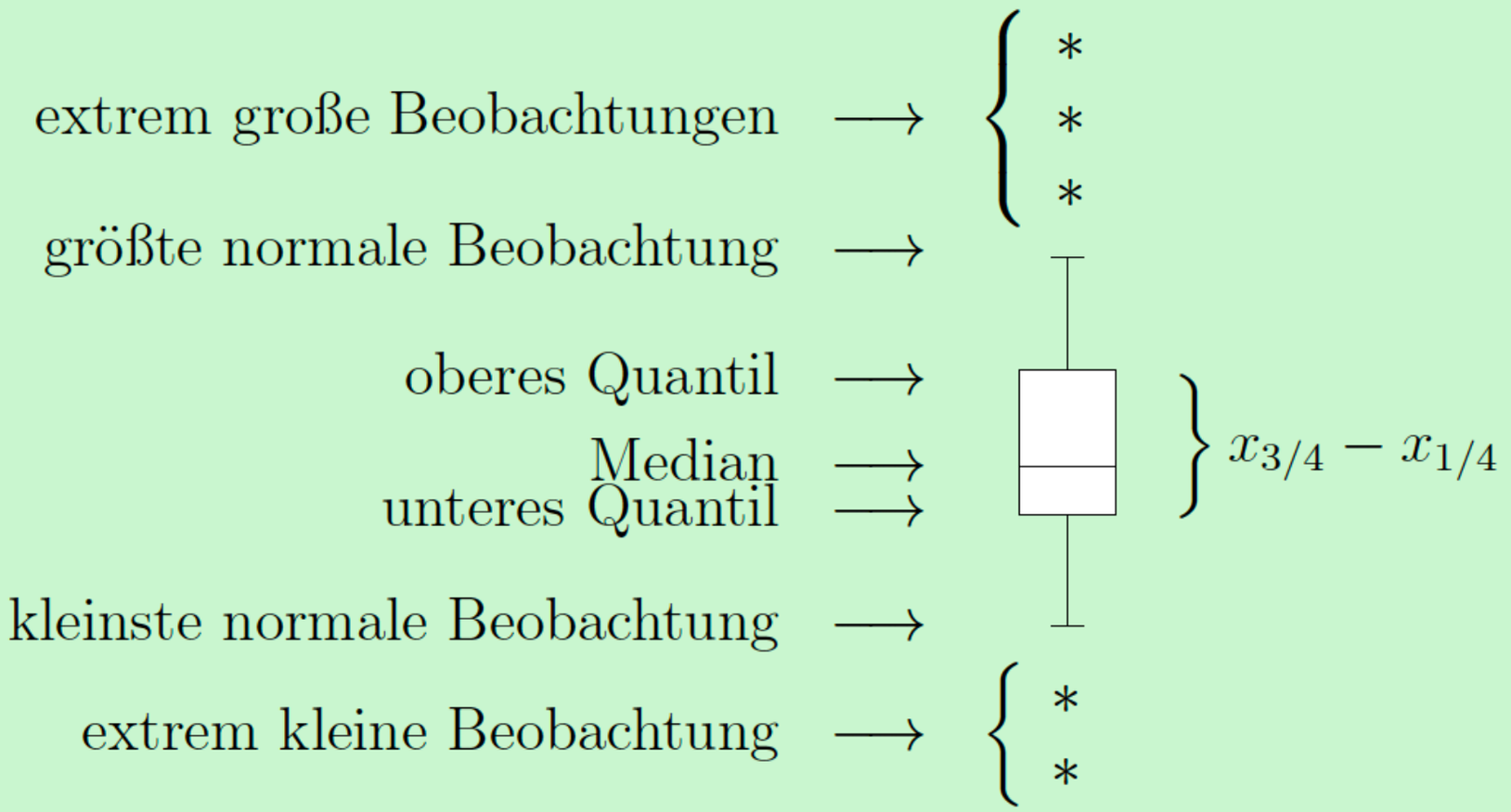
$$x_{3/4} + 1,5 \cdot (x_{3/4} - x_{1/4})$$

ist, die sog. *größte normale Beobachtung*. Der Endpunkt des nach unten aufgesetzten Stabes ist größer als

$$x_{1/4} - 1,5 \cdot (x_{3/4} - x_{1/4}) \,,$$

die sog. *kleinste normale Beobachtung*. Extrem große Beobachtungen sind Daten, die oberhalb von $x_{3/4}+1,5\cdot(x_{3/4}-x_{1/4})$ liegen, extrem kleine Beobachtungen sind Daten, die unterhalb von $x_{1/4} - 1,5 \cdot (x_{3/4} - x_{1/4})$ liegen. Die sog. Ausreißer nach oben und unten werden durch einen Stern oder Punkt gekennzeichnet.
Wegen des Rechtecks in der Mitte des Diagramms spricht man auch von einem *Kisten-Diagramm*.

Darstellung von Daten



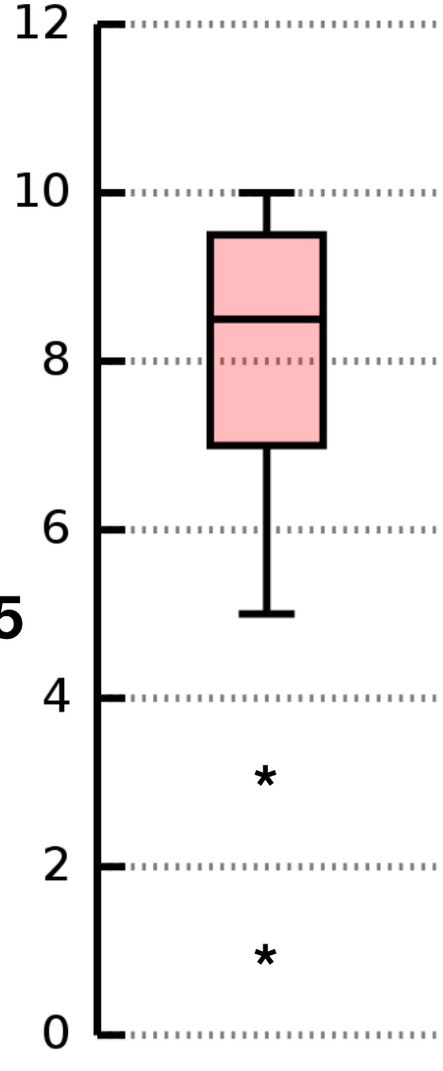
Darstellung von Daten

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i (unsortiert)	9	6	7	7	3	9	10	1	8	7	9	9	8	10	5	10	10	9	10	8
$x_{(i)}$ (sortiert)	1	3	5	6	7	7	7	8	8	8	9	9	9	9	9	10	10	10	10	10

Unteres Quartil: 7 **Median: 8,5** **Oberes Quartil: 9,5**

Größte normale Beobachtung: 10

Kleinste normale Beobachtung: 5



Quartilsabstand: 2,5

×1,5=3,75

7-3,75=3,25. Alles, was kleiner ist, gilt als Ausreißer.

Schätzprobleme

Allgemeines Setting:

Wir haben einen Wahrscheinlichkeitsraum (Ω, P). Die Menge Ω ist die Grundpopulation.

Auf diesem Wahrscheinlichkeitsraum haben wir eine Zufallsvariable X, die jedem $\omega \in \Omega$ einen Wert zuweist. Die Verteilung, Erwartungswert, Varianz, …. von X ist oft nicht bekannt.

Der/die fehlende/n Parameter soll anhand einer konkreten Stichprobe geschätzt werden. (Natürlich möglichst gut!).

Beispiel S1: Wie groß ist die Wahrscheinlichkeit, dass eine gegebene Münze „Kopf“ zeigt?

Wir werfen die Münze 100-mal und erhalten dabei 60-mal „Kopf“.

Schätzprobleme

Beispiel S1: Wie groß ist die Wahrscheinlichkeit, dass eine gegebene Münze „Kopf“ zeigt?
Wir werfen die Münze 100-mal und erhalten dabei 60-mal „Kopf“.

In jedem Fall handelt es sich beim Münzwurf um ein Bernoulli-Experiment.

Betrachte die Zufallsvariable $X = \begin{cases} 1 & \text{"Kopf"} \\ 0 & \text{"Zahl"} \end{cases}$ **, mit** $P(X = 1) = p$ **unbekannt.**

Mit obigen Angaben vermuten wir natürlich $p = \frac{60}{100} = \frac{3}{5}$**, da dies die (gemessene) relative Häufigkeit des Ereignisses „Kopf“ ist.**

Alternativ: Betrachte 100 **Kopien** $X_1, \ldots, X_{100}$ **von** X**, mit**

$$X_i = 1 \iff \text{"Kopf" im } i\text{-ten Wurf}$$

Dann ist $Y = \sum_{i=1}^{100} X_i$ **binomialverteilt, mit** $P(Y = k) = \binom{100}{k} p^k (1-p)^{100-k}$

Finde nun $p \in [0,1]$ **so, dass** $P(Y = 60)$ **maximal wird.**

Das ist natürlich das erwartete $p = \frac{60}{100}$**.**