Stochastik

FH Aachen – Studienstandort Aachen
21. November 2025

Prof. Dr. Florian Heinrichs

Deskriptive Statistik

Wahrscheinlichkeitstheorie vs. Statistik

Bisher: Wahrscheinlichkeitstheorie
Jetzt: Statistik

Wahrscheinlichkeitstheorie

Modellierung des Zufalls: Wahrscheinlichkeiten, Zufallsvariablen, Verteilungen

- Sei X die Anzahl der Tassen Kaffee, die eine Person pro Tag trinkt
- Was ist $\mathbb{P}(X = k)$?
- Sei Y die Kaffeemenge in ml, die eine Person täglich trinkt
- Was ist $\mathbb{P}(Y \leq x)$?

Statistik

Aussagen anhand von Beobachtungen zufälliger Vorgänge machen

- Eine Person hat in den letzten Tagen $2, 3, 1, 2, 0, 1, 2$ Tassen Kaffee getrunken
- Wie viele Tassen Kaffee wird sie morgen trinken?
- Wie wahrscheinlich ist es, dass die Person morgen 3 oder mehr Tassen trinkt?
- Was ist die zugrundeliegende Verteilung? Erwartungswert? Varianz?

Statistik

- **Vorurteil:** Statistik - das sind Tabellen und Diagramme
- **Fakt:** Deskriptive Statistik beschreibt Daten mit Kennzahlen, Tabellen und Diagrammen (ohne allgemeine Schlussfolgerung)
- **Aber:** Schließende Statistik (Inferenzstatistik, mathematische Statistik) verwendet Stichprobendaten, um allgemeine Aussagen oder Hypothesen über eine größere Population zu machen

Deskriptive Statistik

- Wir sammeln die Anzahl der Tassen Kaffee, die 100 Personen täglich trinken
- Zusammenfassen der Daten: Durchschnitt
- Visualisierung: Säulendiagramm
- Ziel: Darstellung und Zusammenfassung der Daten (Aussage über Kaffeekonsum der 100 Personen)

Statistik

- **Vorurteil:** Statistik - das sind Tabellen und Diagramme
- **Fakt:** Deskriptive Statistik beschreibt Daten mit Kennzahlen, Tabellen und Diagrammen (ohne allgemeine Schlussfolgerung)
- **Aber:** Schließende Statistik (Inferenzstatistik, mathematische Statistik) verwendet Stichprobendaten, um allgemeine Aussagen oder Hypothesen über eine größere Population zu machen

Inferenzstatistik

- Wir sammeln die Anzahl der Tassen Kaffee, die 100 Personen trinken
- Basierend auf den Daten möchten wir eine Aussage über den durchschnittlichen Kaffeekonsum in der gesamten Bevölkerung machen
- Ziel: Rückschlüsse auf die gesamte Population ziehen

Begriffe

Wichtige Begriffe der Statistik

- **Grundgesamtheit/Population**: Vollständige Menge von Individuen oder Objekten, über die eine Aussage gemacht werden soll
  - Beispiel: Alle Personen in Aachen
- **Stichprobe**: Zufällige endliche Teilmenge der Grundgesamtheit $x_1, \ldots, x_n$
  - Beispiel: Zufällige Passanten in der Innenstadt, Studierende in Vorlesung
- **Stichprobenumfang**: Größe der Stichprobe n
  - Beispiel: Anzahl der befragten Personen

Begriffe

Wichtige Begriffe der Statistik

- **Merkmal**: Beobachtbare Eigenschaft der Elemente der Grundgesamtheit
  - Beispiel: Menge an Kaffee, die eine Person täglich trinkt
- **(Merkmals-)Ausprägung**: Konkrete Werte, die ein Merkmal annehmen kann
  - Beispiel: Kaffee (ml) in $[0, \infty)$
  - Beispiel: Kaffee (in Tassen) in $\mathbb{N}_0$
  - Beispiel: Kaffee (in Tassen) in $0, 1, 2, 3, \geq 4$

Weitere Begriffe der Statistik

- Urliste: Direktes Ergebnis der Datenerhebung
- Urwerte: Werte der Urliste

Merkmale

Es gibt unterschiedliche Arten von Merkmalen

Qualitative Merkmale

- Nominal: Einteilung in Kategorien (ohne Ordnung)
  - Beispiel: Farben (rot, blau, grün)
  - Weitere Beispiele?
- Ordinal: Geordnete Kategorien
  - Beispiel: Noten (sehr gut, gut, befriedigend, ausreichend, mangelhaft)
  - Weitere Beispiele?

Quantitative Merkmale

- Zähldaten: Ergebnis einer Zählung
  - Beispiel: Anzahl Studierende in Veranstaltung
  - Weitere Beispiele?
- Diskret: Mögliche Werte in diskreter Teilmenge von $\mathbb{R}$
  - Beispiel: Körpergröße in m, gerundet auf ganze cm
  - Weitere Beispiele?
- Stetig: Beliebige Werte innerhalb eines Intervalls
  - Beispiel: Körpergröße in m (nicht gerundet)
  - Weitere Beispiele?

Merkmale

> Wir können qualitative in quantitativen Merkmale umwandeln
>   > Wie?
> Dabei geben wir den Daten allerdings mehr Struktur
>   > Welche?
> Qualitative Merkmale und Zähldaten können auch als diskrete Merkmale interpretiert werden

Merkmale

Beispiel 128: Kaffeekonsum

- Qualitative Merkmale
  - Nominal: Lieblingskaffee (Espresso, Filterkaffee, ...)
  - Ordinal: Häufigkeit des Kaffeetrinkens (täglich, manchmal, nie)
- Quantitative Merkmale
  - Zähldaten: Anzahl der Tassen Kaffee, die eine Person täglich trinkt
  - Diskret: Anzahl Teelöffel Zucker in Kaffee (auch halbe möglich)
  - Stetig: Menge des Kaffeekonsums in ml (theoretisch beliebig genau)

Merkmale

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Übung 75

In der Vorlesung “Analysis I” steht die Klausur bevor. In der letzten Woche sollen speziell die Themen wiederholt werden, bei denen die Studierenden im Semester Schwierigkeiten hatten. Um diese Themen zu identifizieren, wird in einer zufällig ausgewählten Übungsgruppe ein Quiz mit Fragen zu wichtigen Themen (Folgen, Reihen, Stetigkeit, Differenzierbarkeit, Integration) und je 4 vorgegebenen Antworten durchgeführt.
Bestimmen Sie die Grundgesamtheit, die Stichprobe, den Stichprobenumfang, die Merkmale und Merkmalsausprägungen der Untersuchung. Handelt es sich um qualitative oder quantitative Merkmale?

Häufigkeit

Definition 50

Sei X ein Merkmal mit Ausprägungen $(a_i)_{i\in I}$ für eine diskrete Indexmenge I. Sei eine Stichprobe mit Beobachtungen $x_1, \ldots, x_n$ von X gegeben. Die *absolute Häufigkeit* h_i von a_i ist definiert durch

$$h_i = \sum_{j=1}^{n} \mathbb{1}_{\{x_j = a_i\}} \quad i \in I.$$

Die *relative Häufigkeit* ist definiert als $r_i = \frac{h_i}{n}$, für $i \in I$.

Bemerkung 20

- Absolute Häufigkeit: Wie oft beobachtet man a_i?
- Für die absolute Häufigkeit gilt $\sum_{i\in I} h_i = n$
- Für die relative Häufigkeit gilt $\sum_{i\in I} r_i = 1$

Häufigkeit

Beispiel 129: Kaffeekonsum

> Eine Person hat in den letzten Tagen $2, 3, 1, 2, 0, 1, 2$ Tassen Kaffee getrunken

> Für die Kategorien $a_1 =$ "0", $a_2 =$ "1", $a_3 =$ "2", $a_4 =$ "3" und $a_5 =$ "≥ 4" erhalten wir die absoluten Häufigkeiten

i	1	2	3	4	5
a_i	"0"	"1"	"2"	"3"	"≥ 4"
h_i	1	2	3	1	0

> Mit $n = 7$ folgt für die relative Häufigkeit

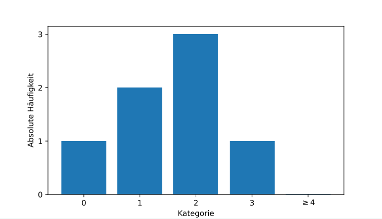
i	1	2	3	4	5
a_i	"0"	"1"	"2"	"3"	"≥ 4"
r_i	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{3}{7}$	$\frac{1}{7}$	0

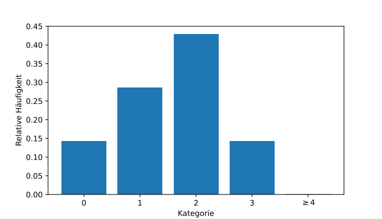
UNIVERSITY OF APPLIED SCIENCES
FH AACHEN

Häufigkeit

Beispiel 129: Kaffeekonsum

- Eine Person hat in den letzten Tagen $2, 3, 1, 2, 0, 1, 2$ Tassen Kaffee getrunken
- Wie können wir die Häufigkeiten (relativ und absolut) visualisieren?
- Mit einem **Säulendiagramm**





Häufigkeit

```python
import numpy as np
import matplotlib.pyplot as plt

if __name__ == '__main__':
    categories = ["0", "1", "2", "3", "$\geq 4$"]
    abs_freq = [1, 2, 3, 1, 0]

    plt.bar(categories, abs_freq)
    plt.xlabel('Kategorie')
    plt.ylabel('Absolute Haeufigkeit')
    plt.show()

    rel_freq = np.array(y) / 7
    plt.bar(categories, rel_freq)
    plt.xlabel('Kategorie')
    plt.ylabel('Relative Haeufigkeit')
    plt.show()
```

Merkmale

Übung 76

Die Klausur zur "Analysis I" wurde geschrieben und die Ergebnisse der ersten 20 Studierenden liegen vor:

$$1.7,\ 2.7,\ 2.7,\ 3.3,\ 2.0,\ 4.0,\ 5.0,\ 4.0,\ 1.0,\ 3.0,\ 3.0,\ 2.3,\ 3.0,\ 4.0,\ 2.3,\ 2.7,\ 5.0,\ 1.0,\ 1.3,\ 1.3$$

1. Visualisieren Sie die Stichprobe mit einem geeigneten Säulendiagramm.
2. Sind die Ausprägungen qualitativ oder quantitativ?

Histogramm

Beispiel 130: Kaffeekonsum

- Kaffee sollte in der French Press etwa 3.5 bis 4 Minuten ziehen
- In einem Büro wurden die folgenden Ziehzeiten gemessen:

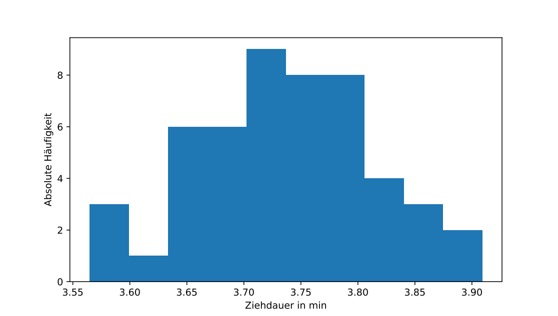
  3.85, 3.66, 3.57, 3.87, 3.66, 3.68, 3.80, 3.82, 3.76, 3.80, 3.60, 3.66, 3.78, 3.56, 3.71, 3.70, 3.67, 3.58, 3.74, 3.71, 3.79, 3.70, 3.76, 3.73, 3.75, 3.71, 3.79, 3.78, 3.81, 3.73, 3.69, 3.67, 3.81, 3.82, 3.89, 3.75, 3.72, 3.77, 3.65, 3.86, 3.72, 3.79, 3.79, 3.77, 3.74, 3.66, 3.91, 3.67, 3.73, 3.72
- Wie können wir die Daten visualisieren?
- Mit einem **Histogramm**
  - Einteilung der Werte in Klassen
  - $a_1 = \min_{i=1}^n x_i$, $a_d = \max_{i=1}^n x_i$
  - Zerlegung des Intervalls $[a_1, a_d]$ in (halboffene) Intervalle $[a_1, a_2), [a_2, a_3), \ldots, [a_{d-2}, a_{d-1}), [a_{d-1}, a_d]$

Histogramm

Beispiel 130: Kaffeekonsum

Wir können die Daten mit einem Histogramm visualisieren

3.85, 3.66, 3.57, 3.87, 3.66, 3.68, 3.80, 3.82, 3.76, 3.80, 3.60, 3.66, 3.78, 3.56, 3.71, 3.70, 3.67, 3.58, 3.74, 3.71, 3.79, 3.70, 3.76, 3.73, 3.75, 3.71, 3.79, 3.78, 3.81, 3.73, 3.69, 3.67, 3.81, 3.82, 3.89, 3.75, 3.72, 3.77, 3.65, 3.86, 3.72, 3.79, 3.79, 3.77, 3.74, 3.66, 3.91, 3.67, 3.73, 3.72



Häufigkeit

```
import numpy as np
import matplotlib.pyplot as plt

if __name__ == '__main__':
    x = (np.random.randn(50) * 5 + 225) / 60
    plt.hist(x)
    plt.xlabel('Ziehdauer in min')
    plt.ylabel('Absolute Haeufigkeit')
    plt.show()
```

Histogramm

- Wie viele Klassen?
  - Generell: ca. $\sqrt{n}$
  - Standard in `matplotlib`: 10
- Wie breit sollten die Klassen sein?
  - Oft: gleiche Breite sinnvoll
- Wie hoch sollten die Rechtecke sein, wenn die Klassen nicht gleich breit sind?
  - **Achtung:** Die Häufigkeit sollte* proportional zur **Fläche** des Rechtecks sein
  - Für die Höhe d_j des Rechtsecks über $[a_j, a_{j-1})$ soll* gelten

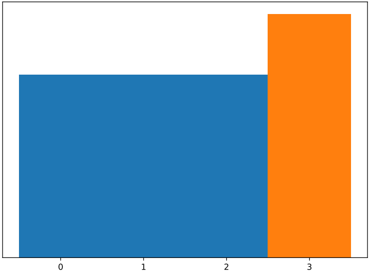
$$d_j \cdot (a_{j+1} - a_j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[a_j, a_{j-1})}(x_i)$$

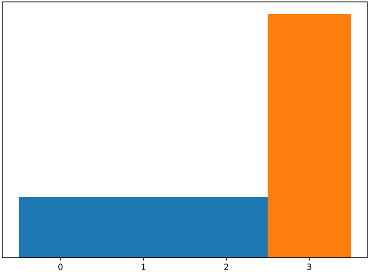
  - Andernfalls suggeriert das Diagramm andere Verhältnisse!

*in Prüfungen: **muss**

Histogramm

- 7 Personen wurden befragt wie viele Tassen Kaffee sie täglich trinken
- Die Antworten sind $3, 3, 0, 3, 1, 2, 3$ Tassen
- Welche Visualisierung ist geeigneter zur Darstellung der Daten, wenn wir zwei Kategorien ("≤ 2", "≥ 3") haben?





Merkmale

Übung 77

Die Klausur zur "Analysis I" wurde geschrieben und die Punktezahlen der ersten 20 Studierenden liegen vor. Es konnten insgesamt 100 Punkte erreicht werden.

$$92,\ 87,\ 100,\ 63,\ 84,\ 45,\ 100,\ 75,\ 93,\ 80,\ 87,\ 67,\ 83,\ 64,\ 41,\ 86,\ 93,\ 41,\ 68,\ 77$$

1. Visualisieren Sie die Stichprobe mit einem Histogramm. Unterteilen Sie dafür den Bereich der erreichbaren Punkte in geeignete Intervalle.
2. Sind die Ausprägungen qualitativ oder quantitativ?

Kontingenztafel

> Wie können wir Stichproben aus zwei diskreten Merkmalen darstellen?
> Mit einer **Kontingenztafel**

2. Merkmal / 1. Merkmal	b_1	b_2	$\dots$	b_ℓ	Randhäufigkeit 1. Merkmal
a_1	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
a_2	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
a_k	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Randhäufigkeit 2. Merkmal	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Kontingenztafel

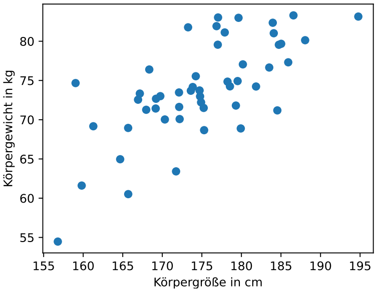
Beispiel 131

100 Personen wurden befragt welchen und wie viel Kaffee sie trinken.

Kaffeeart / Tassen pro Tag	Espresso	Cappuccino	Americano	Randhäufigkeit (Tassen pro Tag)
1	1	19	18	38
2	21	4	12	37
≥ 3	1	20	4	25
Randhäufigkeit (Kaffeeart)	23	43	34	100

Streudiagramm

- Wie können wir Stichproben mit zwei stetigen Merkmalen darstellen?
- Mit einem **Streudiagramm**
- Sei eine Stichprobe mit den beiden stetigen Merkmalen Körpergröße und -gewicht gegeben



Literatur I

- Bundesministerium für Gesundheit (2024).
  Infektionsradar.
  https://infektionsradar.gesund.bund.de/de/covid/inzidenz.
  Abgerufen: 2024-10-15.

- Dehling, H. and Haupt, B. (2006).
  Einführung in die Wahrscheinlichkeitstheorie und Statistik.
  Springer-Verlag.

Literatur II

Dinnes, J., Sharma, P., Berhane, S., van Wyk, S., Nyaaba, N., Domen, J., Taylor, M., Cunningham, J., Davenport, C., Dittrich, S., Emperador, D., Hooft, L., Leeflang, M., McInnes, M., Spijker, R., Verbakel, J., Takwoingi, Y., Taylor-Phillips, S., Van den Bruel, A., and Deeks, J. (2022).
Rapid, point-of-care antigen tests for diagnosis of sars-cov-2 infection.
Cochrane Database of Systematic Reviews, (7).

Henze, N. et al. (1997).
Stochastik für Einsteiger, volume 4.
Springer.