

1. Übung: Datenerhebung & Messung

Yichen Han



26. Oktober 2023

1 Organisation

2 Wiederholung

3 Blatt 1

- ① Vor der Übung die Aufgaben selbstständig versuchen zu lösen.
- ② Abgabe der Hausarbeiten (Ende jedes Monats) in Gruppen erwünscht.
- ③ Es wird **keine** Musterlösung inkl. Folien hochgeladen.

- ① Vor der Übung die Aufgaben selbstständig versuchen zu lösen.
- ② Abgabe der Hausarbeiten (Ende jedes Monats) in Gruppen erwünscht.
- ③ Es wird **keine** Musterlösung inkl. Folien hochgeladen.
- ④ Interaktive Lehre mithilfe der Plattform Particify (kostenlos, keine Registrierung erforderlich, anonym)



- ⑤ Fragen & Feedbacks gerne per Email an Yichen.Han@campus.lmu.de
Gerne bin ich auch Ihre Kontaktperson für die Fachschaft.

- ① kleines Quiz für die Vorlesung
- ② kurze Wiederholung der wesentlichen Inhalte der Vorlesung
- ③ Darstellung der Lösungen zu Übungsaufgaben
- ④ ggf. kleine interaktive Aufgaben oder Zeit für Fragen
- ⑤ wenn Zeit knapp ist -> Aufgaben und Wiederholung parallel

- ① Die Übungsblätter bis Ende des 2. Semesters aufbewahren.
- ② Mathematische Kenntnisse auffrischen und ggf. erweitern.
- ③ Das kostenlose Konto von DataCamp gut nutzen.
- ④ Das Studium sinnvoll digitalisieren. (GitHub, L^AT_EX, Stackoverflow, ChatGPT, usw.)

1 Organisation

2 Wiederholung

3 Blatt 1



Quiz 1:



Quiz 1:

- ① Grundbegriffe: Grundgesamtheit, Untersuchungseinheit, Merkmal, Ausprägung, Beobachtung
- ② Skalenniveaus: nominal, ordinal, intervall, verhältnis, absolut
- ③ Merkmalstypen: stetig, diskret, quasi-stetig
- ④ Erhebungsarten (Methode, Datenform, Umfang): Experiment, Befragung, Beobachtung, Vollerhebung, Stichprobe, Querschnittsdaten, Zeitreihe, Längsschnittdaten

Warum sind Querschnittsstudien nicht in der Lage, eine (zeitliche) Tendenz aus den Daten zu verdeutlichen? (Kohorteneffekt)¹:

Szenario:

Eine Querschnittsstudie untersucht die Nutzung von TikTok der verschiedenen Altersgruppen im Jahr 2023. Die Studie findet heraus, dass ältere Menschen tendenziell weniger TikTok bevorzugen.

Diese Schlussfolgerung ist aber nicht sinnvoll.

¹ Thoughtco, Was ist ein Kohorteneffekt? Definition und Beispiele

1 Organisation

2 Wiederholung

3 Blatt 1

Die drei Säulen der Statistik

- **Deskriptive:** Beschreibung von betrachteten Merkmalen, graphische Datenaufbereitung, liefert erster Eindruck der Daten, und hilft bei Datenvierlidierung (erkennt Fehler). Keine Rückschlüsse auf die Grundgesamtheit über Erhebungsdaten möglich.

Die drei Säulen der Statistik

- **Deskriptive:** Beschreibung von betrachteten Merkmalen, graphische Datenaufbereitung, liefert erster Eindruck der Daten, und hilft bei Datenvierlidierung (erkennt Fehler). Keine Rückschlüsse auf die Grundgesamtheit über Erhebungsdaten möglich.
- **Explorative:** Suche nach Strukturen in den Daten (ohne stochastische Methoden), Formulierung von Hypothesen für das den Daten zugrunde liegende stochastische Modell (wichtig für die Induktion).

Die drei Säulen der Statistik

- **Deskriptive:** Beschreibung von betrachteten Merkmalen, graphische Datenaufbereitung, liefert erster Eindruck der Daten, und hilft bei Datenvierlidierung (erkennt Fehler). Keine Rückschlüsse auf die Grundgesamtheit über Erhebungsdaten möglich.
- **Explorative:** Suche nach Strukturen in den Daten (ohne stochastische Methoden), Formulierung von Hypothesen für das den Daten zugrunde liegende stochastische Modell (wichtig für die Induktion).
- **Induktive:** Ziehung von Schlüssen von der Stichprobe auf die Grundgesamtheit; Verwendung entsprechender statistischer Methoden auf Basis eines wahrscheinlichkeitstheoretischen Modells. (Punktschätzung, Konfidenzintervalle, Tests)

Aufgabe 2.1

Bestimmen Sie die Grundgesamtheit (Ω), die statistischen Einheiten (ω_i) und Beobachtungen, die untersuchten Merkmale (X), die theoretisch möglichen Merkmalsausprägungen (S), und weitere erfassbare Merkmale.

- ① Ω : Residenzkonzerte in München im Jahr 2015;
 ω_i : Residenzkonzert i in München im Jahr 2015;
 X : Anzahl der verkauften Karten bei einem Residenzkonzert;
 $S = \mathbb{N}_0$;
weitere: Art und Datum des Konzertes, usw.
- ② Ω : Studierende der LMU;
 ω_i : Studierende(r) i an der LMU;
 X : primäres Verkehrsmittel, das zur Fahrt zur Uni genutzt wird;
 S : Menge aller möglichen Verkehrsmittel;
weitere: Studienfach, Alter, Geschlecht der Studierenden usw.

Aufgabe 2.2

- ③ Ω : Amtlich betriebene Stationen zur Messung von Luftschadstoffen in Bayern;

ω_i : Amtlich betriebene Station i ;

$X = (x, y)$: GPS-Koordinate einer Station;

$S = [47, 51] \times [9, 14]$; ²

weitere: Baujahr der Station, Ballungsraum ja/nein, Entfernung zur nächsten Industrieanlage, usw.

²GPS-Bereich von Bayern, müssen Sie nicht wissen

Aufgabe 3



Quiz 2:

	Natürliche Reihen- folge	Interpretierbare Differenzen	Natürlicher Null- punkt	Natürliche Einheit
Nominalskala	nein	nein	nein	nein
Ordinalskala	ja	nein	nein	nein
Intervallskala	ja	ja	nein	nein
Verhältnisskala	ja	ja	ja	nein
Absolutskala	ja	ja	ja	ja

Tabelle: Charakterisierung der Skalenniveaus

Aufgabe 4

Erläutern Sie geeignete Erhebungsarten (Methode, Datenform, Umfang) für die folgenden Sachverhalte.

- ① Testen eines Düngemittels: *Experiment, Längsschnittanalyse, Stichprobe*
- ② Einschätzung der Fahrtüchtigkeit: *Befragung, Querschnittanalyse, Stichprobe*
- ③ Schätzung der durchschnittlichen Lebensdauer von Leuchtstoffröhren: *Beobachtung, Längsschnittanalyse, Stichprobe*

Aufgabe 5.1

Welches Auswahlverfahren würden Sie wählen?

Szenario 1:

Eine Untersuchung soll Aufschluss über den **durchschnittlichen Quadratmeterpreis** von Mietwohnungen in einer Stadt geben. Sie wissen, dass man die Stadt in **drei „Regionen“**; einteilen kann, in denen die Quadratmeterpreise jeweils ähnlich sind.

Aufgabe 5.1

Welches Auswahlverfahren würden Sie wählen?

Szenario 1:

Eine Untersuchung soll Aufschluss über den **durchschnittlichen Quadratmeterpreis** von Mietwohnungen in einer Stadt geben. Sie wissen, dass man die Stadt in **drei „Regionen“**; einteilen kann, in denen die Quadratmeterpreise jeweils ähnlich sind.

- einfache Zufallsstichprobe? Mietpreise stark von der „Region“ abhängig.

Aufgabe 5.1

Welches Auswahlverfahren würden Sie wählen?

Szenario 1:

Eine Untersuchung soll Aufschluss über den **durchschnittlichen Quadratmeterpreis** von Mietwohnungen in einer Stadt geben. Sie wissen, dass man die Stadt in **drei „Regionen“**; einteilen kann, in denen die Quadratmeterpreise jeweils ähnlich sind.

- einfache Zufallsstichprobe? Mietpreise stark von der „Region“ abhängig.
- → jeweils eine Stichprobe mit dem gleichen Umfang in jeder Region durchführen.

Vorteile: Wissen über jede Region + niedrigere Varianz für das Gesamtmittel (bessere Schätzung)

Geschichtete Stichprobe

Referenz: Fahrmeir et al. Kap.1.4 (S. 23-24)

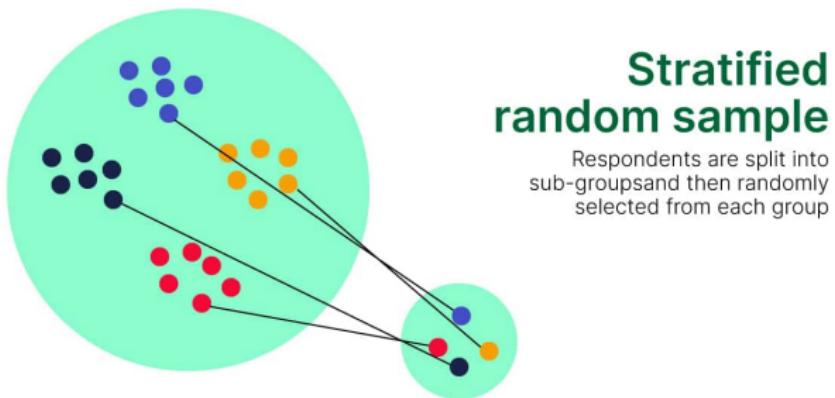


Abbildung: Beispiel der geschichteten Stichprobe

Effizienzgewinn, falls die relevanten Merkmale innerhalb der Schichten homogen und zwischen den Schichten heterogen sind.

Szenario 2:

Zur Bestimmung des Zigarettenkonsums von Hauptschülern in der 8. Klasse soll eine Erhebung mit Hilfe von Fragebögen durchgeführt werden.

- Ist es möglich, in einem Jahrgang eine Stichprobe durchzuführen?
- Ist es möglich, eine Vollerhebung durchzuführen?
- Was muss zufällig ausgewählt werden?

Aufgabe 5.4

(Einstufige) Klumpenstichprobe

Referenz: Fahrmeir et al. Kap.1.4 (S. 23-24)

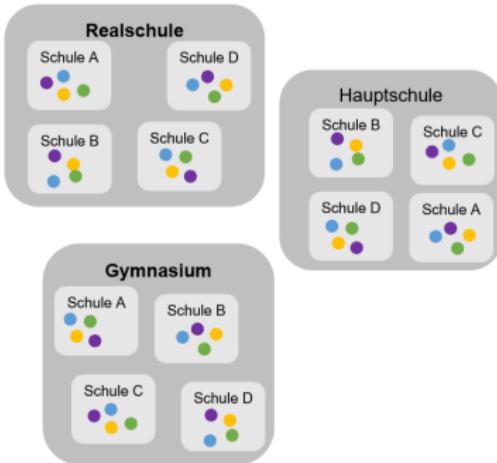


Abbildung: Beispiel der Klumpenstichprobe, Source

- → In ein paar zufällig ausgewählten Hauptschulen jeweils eine Vollerhebung für Schüler:innen der 8. Klasse durchführen.

Grundlagen der Wahrscheinlichkeitsrechnung Blatt 2 für Abgabe:

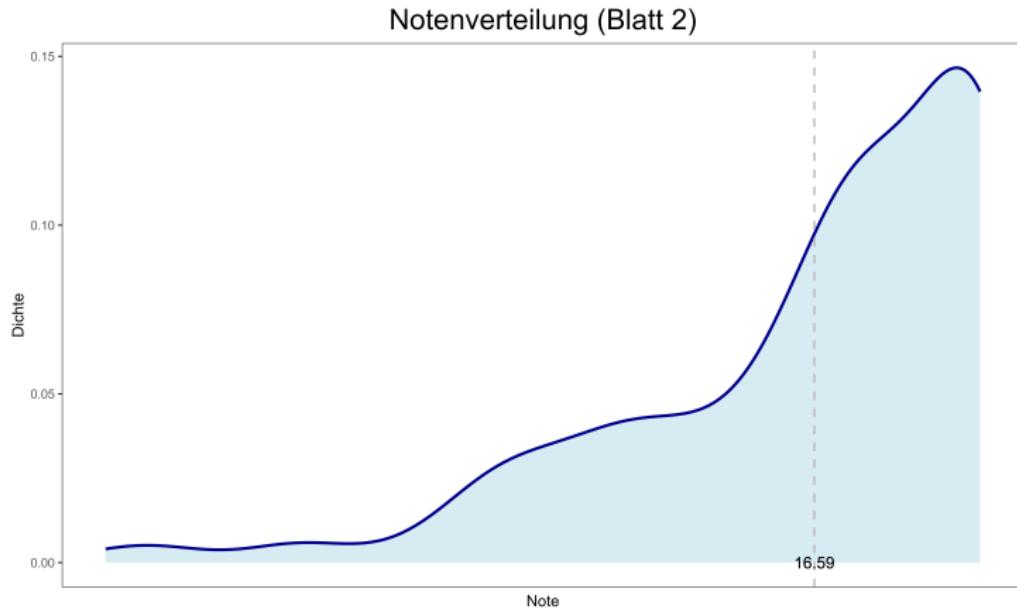
- **Ausschließlich** die erste und zweite Aufgabe werden benotet!
- Die Gruppe erhält die gleiche Bewertung.
- Achten Sie auf die Fristen!

2. Übung: Grundlagen der Wahrscheinlichkeitsrechnung

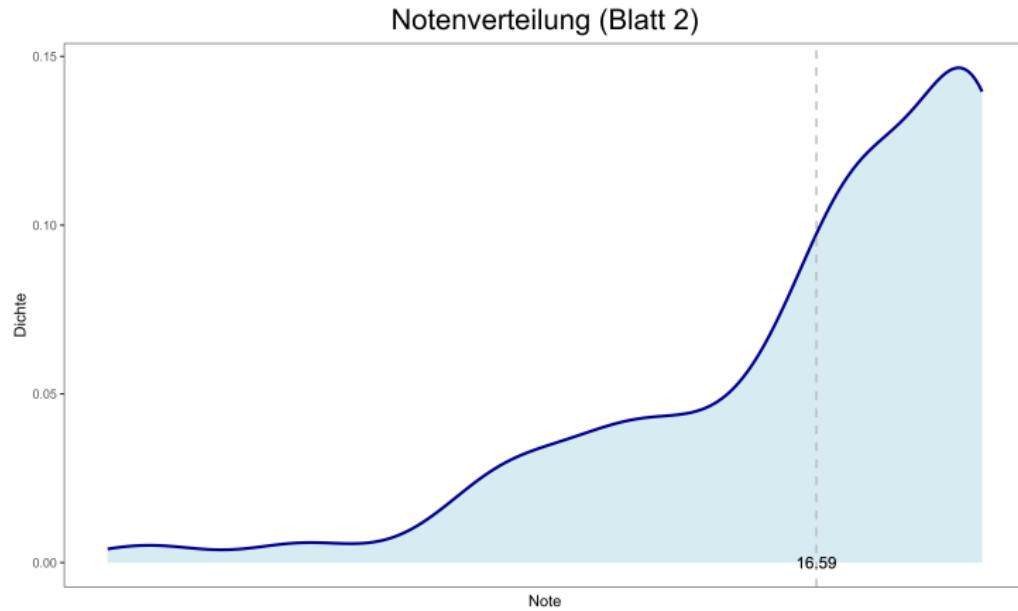
Yichen Han



2. November 2023



falsche Antwort \neq 0 Pkt.
richtige Antwort \neq alle Pkt.



falsche Antwort \neq 0 Pkt.

richtige Antwort \neq alle Pkt.

Ziel: Aufgaben richtig und sinnvoll lösen, ohne übermäßigen Aufwand.

1 Wiederholung

2 Blatt 2

- Welche der folgenden Aussagen wird nicht von den Kolmogorov-Axiomen besagt? (A, B, C)
- Richtig oder falsch? $A \perp B \Leftrightarrow P(A \cup B) = P(A) + P(B)$
- Richtig oder falsch? Disjunkte Ereignisse sind auch unabhängig.
- Richtig oder falsch? Aus "stochastisch unabhängig" folgt "bedingt stochastisch unabhängig". D.h. $A \perp B \Rightarrow A \perp B|C$, mit C ein beliebiges Ereignis in Ω .
- Richtig oder falsch? Mit dem Satz von Bayes kann man mit neu erhobenen Daten überprüfen, wie plausibel eine aus vorherigen Daten generierten Annahme ist.



Quiz 3

Laplace-Wahrscheinlichkeit

$$P(A) = \frac{|A|}{|\Omega|} \quad (1)$$

Kolmogorov-Axiome

$$\begin{aligned} \forall A \subseteq \Omega, \quad & P(A) \geq 0 \\ P(\Omega) &= 1 \\ P(A \cup B) &= P(A) + P(B), \quad A, B \text{ disjunkt} \end{aligned} \quad (2)$$

Bedingte Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (3)$$

Satz von der totalen W.keit

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i) \quad (4)$$

Siebformel

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (5)$$

Stochastische Unabhängigkeit

$$A \perp B \Leftrightarrow P(A \cap B) = P(A)P(B) \quad (6)$$

Bedingte Unabhängigkeit

$$P(A \cap B|C) = P(A|C)P(B|C) \quad (7)$$

Satz von Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (8)$$

Disjunkt vs. Unabhängig:

Disjunktheit (mutual exklusiv)

- $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B)$

Unabhängigkeit

- $P(A \cap B) = P(A)P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A)P(B)$

Frequentists vs. Bayesians (Selbststudium)

Beispiel: Würfelwurf

	Frequentists	Bayesians
Parameter	keine Zufallsvariable	Zufallsvariable
Ziel	Entscheidung	Überzeugung
Antwort	r oder f	weder r noch f

1 Wiederholung

2 Blatt 2

Aufgabe 1 (benotet)

Grundlegende Kombinatorik

5 Tage, 5 Mitglieder

- ➊ 5 Möglichkeiten pro Tag; $5^5 = 3125$

Aufgabe 1 (benotet)

Grundlegende Kombinatorik

5 Tage, 5 Mitglieder

- ① 5 Möglichkeiten pro Tag; $5^5 = 3125$
- ② jeden Tag eine Möglichkeit weniger; $5! = 120$

Aufgabe 1 (benotet)

Grundlegende Kombinatorik

5 Tage, 5 Mitglieder

- ① 5 Möglichkeiten pro Tag; $5^5 = 3125$
- ② jeden Tag eine Möglichkeit weniger; $5! = 120$
- ③ B wählt ohne Reihenfolge 2 Tage aus, dann 3 Personen jeweils 1 Tag;
$$\binom{5}{2} \cdot 3! = 60$$

Aufgabe 1 (benotet)

Grundlegende Kombinatorik

5 Tage, 5 Mitglieder

- ① 5 Möglichkeiten pro Tag; $5^5 = 3125$
- ② jeden Tag eine Möglichkeit weniger; $5! = 120$
- ③ B wählt ohne Reihenfolge 2 Tage aus, dann 3 Personen jeweils 1 Tag;
 $\binom{5}{2} \cdot 3! = 60$
- ④ B hat 6 Möglichkeiten: (1, 3), (1, 4), (1, 5), (2, 4), (2, 5), (3, 5), und
dann 3 Personen jeweils 1 Tag; $6 \cdot 3! = 36$

Alternativ:

4 Möglichkeiten fürs Aufeinanderfolgen (BBXXX, XBBXX, XXBBX, XXXBB), mit 3 Personen jeweils 1 Tag; $\binom{5}{2} \cdot 3! - 4 \cdot 3! = 36$

Aufgabe 2 (benotet)

Kombinatorik und Laplace-W.keit

3 Sonderpreise, 6 Teilnehmende, 2 Autos (Kapazität 4)

① $A := \text{"3 unterschiedliche Gewinner:innen"}$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\binom{6}{3}}{6^3} = \frac{6 \cdot 5 \cdot 4}{216} \approx 0.56$$


Aufgabe 2 (benotet)

Kombinatorik und Laplace-W.keit

3 Sonderpreise, 6 Teilnehmende, 2 Autos (Kapazität 4)

① $A := \text{"3 unterschiedliche Gewinner:innen"}$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\binom{6}{3}}{6^3} = \frac{6 \cdot 5 \cdot 4}{216} \approx 0.56$$

② 2:4 oder 3:3

$$\binom{2}{1} \binom{6}{2} \binom{4}{4} + \binom{6}{3} \binom{3}{3} = 2 \cdot 15 + 20 = 50$$

Autos

Aufgabe 2 (benotet)

Kombinatorik und Laplace-W.keit

3 Sonderpreise, 6 Teilnehmende, 2 Autos (Kapazität 4)

- ① $A := \text{"3 unterschiedliche Gewinner:innen"}$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\binom{6}{3}}{6^3} = \frac{6 \cdot 5 \cdot 4}{216} \approx 0.56$$

- ② 2:4 oder 3:3

$$\binom{2}{1} \binom{6}{2} \binom{4}{4} + \binom{6}{3} \binom{3}{3} = 2 \cdot 15 + 20 = 50$$

- ③ Bedingt? Laplace! $C := \text{"Alle 3 Gewinner:innen im selben Auto"}$, Ω wie in (b).

$$|C| = \binom{3}{2} \binom{3}{3} \binom{1}{1} + \binom{2}{1} \binom{3}{3} \binom{3}{3} + \binom{3}{1} \binom{3}{3} \binom{2}{2} = 8$$

Loser Gw LS Auto *Loser LS*
Loser Gw v. *Gw*

$$\Rightarrow P(C) = \frac{8}{50} = 0.16$$

Aufgabe 3

z.Z.: P ist Wahrscheinlichkeitsmaß
Schritt für Schritt jede Eigenschaft validieren.

Aufgabe 3

z.Z.: P ist Wahrscheinlichkeitsmaß

Schritt für Schritt jede Eigenschaft validieren.

A1) $P(A) = \sum_{i: \omega_i \in A} p_i \geq 0 \quad \forall A \subset \Omega, \text{ da } p_i = P(\{\omega_i\}) \in [0, 1] \quad \checkmark$

A2) $P(\Omega) = \sum_{i: \omega_i \in \Omega} p_i = \sum_{i=1}^n p_i = 1 \quad \checkmark$

A3) betrachte zwei disjunkte Mengen $A, B \subset \Omega$:

$$\begin{aligned} P(A \cup B) &= \sum_{\omega_i \in A \cup B} p_i \\ &= \sum_{\omega_i \in A \vee \omega_i \in B} p_i = \sum_{\omega_i \in A} p_i + \sum_{\omega_i \in B} p_i - \sum_{\omega_i \in A \cap B} p_i \\ &= \sum_A p_i + \sum_B p_i = P(A) + P(B) \quad \checkmark \end{aligned}$$

Aufgabe 3

z.Z.: P ist Wahrscheinlichkeitsmaß

Schritt für Schritt jede Eigenschaft validieren.

A1) $P(A) = \sum_{i: \omega_i \in A} p_i \geq 0 \quad \forall A \subset \Omega, \text{ da } p_i = P(\{\omega_i\}) \in [0, 1] \quad \checkmark$

A2) $P(\Omega) = \sum_{i: \omega_i \in \Omega} p_i = \sum_{i=1}^n p_i = 1 \quad \checkmark$

A3) betrachte zwei disjunkte Mengen $A, B \subset \Omega$:

$$\begin{aligned} P(A \cup B) &= \sum_{i: \omega_i \in (A \cup B)} p_i = \sum_{i: \omega_i \in A \vee \omega_i \in B} p_i \\ &= \sum_{i: \omega_i \in A} p_i + \sum_{i: \omega_i \in B} p_i - \underbrace{\sum_{i: \omega_i \in A \wedge \omega_i \in B} p_i}_{= \sum_{i \in \emptyset} p_i = 0} = P(A) + P(B) \end{aligned}$$

Aufgabe 4



Quiz 4:

Aufgabe 4



Quiz 4:

① $P(A) = P(\bar{B}) \Rightarrow \bar{A} = B$

Falsch. Gegenbeispiel: $\Omega = \{1, 2, 3\}$, $A = \{1, 2\}$, $B = \{1\}$

Aufgabe 4



Quiz 4:

① $P(A) = P(\bar{B}) \Rightarrow \bar{A} = B$

Falsch. Gegenbeispiel: $\Omega = \{1, 2, 3\}$, $A = \{1, 2\}$, $B = \{1\}$

② $P(A) = 0 \Rightarrow P(A \cap B) = 0$

Wahr.

$$0 \leq$$

Lösung 1 (Einschachtelung): $P(A \cap B) \leq P(A) = 0$

Lösung 2 (Bayes): $P(A \cap B) = P(B|A)P(A) = 0$

Aufgabe 4



Quiz 4:

① $P(A) = P(\bar{B}) \Rightarrow \bar{A} = B$

Falsch. Gegenbeispiel: $\Omega = \{1, 2, 3\}$, $A = \{1, 2\}$, $B = \{1\}$

② $P(A) = 0 \Rightarrow P(A \cap B) = 0$

Wahr.

Lösung 1 (Einschachtelung): $P(A \cap B) \leq P(A) = 0$

Lösung 2 (Bayes): $P(A \cap B) = P(B|A)P(A) = 0$

③ Berechne c für $P(\{\omega_i\}) = \frac{c}{n!}$, $\Omega = \mathbb{Z}_0^+$

A1 $\Rightarrow c \geq 0$

A2:

$$\Rightarrow \sum_{k=0}^{\infty} \frac{c}{k!} = c \cdot \sum_{k=0}^{\infty} \frac{1}{k!} = c \cdot e \stackrel{!}{=} 1 \Rightarrow c = e^{-1}$$

Aufgabe 5.1

$$P(A) = 3/4, \ P(B) = 1/3, \text{ z.Z.: } \frac{1}{12} \leq P(A \cap B) \leq \frac{1}{3}.$$

Aufgabe 5.1

$$P(A) = 3/4, \ P(B) = 1/3, \text{ z.Z.: } \frac{1}{12} \leq P(A \cap B) \leq \frac{1}{3}.$$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \leq 1 \\ &= 3/4 + 1/3 - P(A \cap B) \leq 1 \\ &= 13/12 - P(A \cap B) \leq 1 \end{aligned}$$

$$\Rightarrow P(A \cap B) \geq 13/12 - 1 = 1/12 \quad (\text{untere Grenze})$$

$P(A \cap B) \leq \min \{P(A), P(B)\}$.
" = " gdw. $B \subset A \Rightarrow P(B)$

Aufgabe 5.1

$$P(A) = 3/4, \quad P(B) = 1/3, \text{ z.Z.: } \frac{1}{12} \leq P(A \cap B) \leq \frac{1}{3}.$$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \leq 1 \\ &= 3/4 + 1/3 - P(A \cap B) \leq 1 \\ &= 13/12 - P(A \cap B) \leq 1 \\ \implies P(A \cap B) &\geq 13/12 - 1 = 1/12 \quad (\text{untere Grenze}) \end{aligned}$$

$P(A \cap B)$ ist maximal, falls $B \subset A$, d.h. wenn B eintritt ist auch A erfüllt :

$$\implies P(A \cap B) = P(B) = 1/3 \quad (\text{obere Grenze})$$

$$\text{Insgesamt gilt somit: } \frac{1}{12} \leq P(A \cap B) \leq \frac{1}{3}$$

Aufgabe 5.2

$P(A) = 3/4$, $P(B) = 1/3$, Wertebereich von $P(A \cup B)$?

Aufgabe 5.2

$P(A) = 3/4$, $P(B) = 1/3$, Wertebereich von $P(A \cup B)$?

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&\geq 3/4 + 1/3 - 1/3 = 3/4\end{aligned}$$

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&\leq 3/4 + 1/3 - 1/12 = 1\end{aligned}$$

somit: $\frac{3}{4} \leq P(A \cup B) \leq 1$

Zusatzaufgabe: Bayes in Medizin

Angenommen, eine seltene Krankheit betrifft 1 von 1000 Personen in einer Bevölkerung. Ein Test auf diese Krankheit ist zu 99% zuverlässig, d.h., die Wahrscheinlichkeit, dass eine kranke Person ein positives Testergebnis erhält, beträgt 99%, und die Wahrscheinlichkeit, dass eine gesunde Person ein negatives Testergebnis erhält, beträgt ebenfalls 99%.

Wenn eine Person zufällig ausgewählt wird und ein positives Testergebnis erhält, wie hoch ist die Wahrscheinlichkeit, dass diese Person tatsächlich die Krankheit hat?



Abbildung: Quiz 5

Zusatzaufgabe

$K :=$ "Person ist krank."

$T :=$ "Person wird als positiv getestet."

Wir wissen:

$$P(K) = 1 : 1000 = 0.001$$

$$P(T|K) = P(\bar{T}|\bar{K}) = 0.99$$

$$P(T|\bar{K}) = 0.01$$

$$P(\bar{K}) = 0.999$$

Wir suchen: $P(K|T)$. Satz von Bayes: $P(K|T) = \frac{P(T|K) \cdot P(K)}{P(T)}$

Satz v. tot. W.keit:

$$\begin{aligned} P(T) &= P(T|K)P(K) + P(T|\bar{K})P(\bar{K}) \\ &= 0.99 \times 0.001 + 0.01 \times 0.999 = 0.01098 \end{aligned}$$

$$P(K|T) = \frac{0.99 \times 0.001}{0.01098} \approx 9.02\%$$

3. Bedingte Wahrscheinlichkeit und Bayes

Yichen Han



9. November 2023

Entschuldigung fürs Durcheinander mit dem Zeitplan...



Abbildung: Quiz 6

- Die Abstraktion von Wahrscheinlichkeitsausdrücken aus Textbeschreibungen unter realistischen Szenarien.
- Die Anwendung von den gelernten Formeln (bedingte W.keit, S.v.tot.W.keit, Bayes), um aus einer gegebenen bedingten Wahrscheinlichkeit $P(A|B)$ u.a. die umgekehrte bedingte Wahrscheinlichkeit $P(B|A)$ zu berechnen.

Bedingte Wahrscheinlichkeit

$$P(A \cap B) = P(A|B)P(B) \quad (1)$$

Satz von der totalen W.keit

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i) \quad (2)$$

Stochastische Unabhängigkeit

$$A \perp B \Leftrightarrow P(A \cap B) = P(A)P(B) \quad (3)$$

Satz von Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (4)$$

Was ihr bisher über Wahrscheinlichkeit gelernt habt, ist aber nicht
unangreifbar...

¹Credit: Prof. Dr. Volker Schmid, Prof. Dr. David Rügamer

Was ihr bisher über Wahrscheinlichkeit gelernt habt, ist aber nicht unangreifbar...

Überlegt euch die folgenden Fragen: ¹

- Kann eine Wahrscheinlichkeit immer auf der Potenzmenge $\mathcal{P} := \{A | A \in \Omega\}$ definiert werden? Was passiert, wenn \mathcal{P} sehr groß ist? (S. Satz von Vitali)
- Laplace-Wahrscheinlichkeit: $P(A) = \frac{|A|}{|\Omega|}$ Was passiert bei $|A| = |\Omega| = \infty$? Was soll man dann machen?

¹Credit: Prof. Dr. Volker Schmid, Prof. Dr. David Rügamer

Was ihr bisher über Wahrscheinlichkeit gelernt habt, ist aber nicht unangreifbar...

Überlegt euch die folgenden Fragen:¹

- Kann eine Wahrscheinlichkeit immer auf der Potenzmenge $\mathcal{P} := \{A | A \in \Omega\}$ definiert werden? Was passiert, wenn \mathcal{P} sehr groß ist? (S. Satz von Vitali)
- Laplace-Wahrscheinlichkeit: $P(A) = \frac{|A|}{|\Omega|}$ Was passiert bei $|A| = |\Omega| = \infty$? Was soll man dann machen?

Zwar könnetet ihr schon intuitive Ideen dazu haben, aber eine mathematische Erläuterung ist derzeit für euch noch unmöglich.
(Maßtheorie)

Studiert weiter mit bestehender Zweifel. Alles wird in Stat2 geklärt.

¹Credit: Prof. Dr. Volker Schmid, Prof. Dr. David Rügamer

1 Blatt 3

2 Wiederholung

Aufgabe 1

Gegeben seien zwei Ereignisse A und B mit $P(A) = P(B) = 1/2$ und $P(B|A) = 1/2$. Sind A und B unabhängig? Wie groß ist $P(A \cup B)$?

Aufgabe 1

Gegeben seien zwei Ereignisse A und B mit $P(A) = P(B) = 1/2$ und $P(B|A) = 1/2$. Sind A und B unabhängig? Wie groß ist $P(A \cup B)$?

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{1}{2}$$

$$\Rightarrow P(B \cap A) = \frac{1}{4} = P(A) \cdot P(B) \Leftrightarrow A \perp B$$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} \\ &= \frac{3}{4} \end{aligned}$$

Aufgabe 2.1

Die Eingänge eines Ladens sind mit einer Alarmanlage gegen Diebstahl gesichert. Wenn Diebe die Anlage passieren, wird mit W'keit 0.995 Alarm ausgelöst. Bei unbescholtenen Kund:innen beträgt die W'keit 0.006. Erfahrungswerte zeigen, daß auf 1000 Kund:innen zwei Dieb:innen kommen.

- a Mit welcher W'keit alarmiert die Anlage zu Recht? Mit welcher Wahrscheinlichkeit werden harmlose Kund:innen erschreckt?

Aufgabe 2.1

Die Eingänge eines Ladens sind mit einer Alarmanlage gegen Diebstahl gesichert. Wenn Diebe die Anlage passieren, wird mit W'keit 0.995 Alarm ausgelöst. Bei unbescholtenen Kund:innen beträgt die W'keit 0.006. Erfahrungswerte zeigen, daß auf 1000 Kund:innen zwei Dieb:innen kommen.

- a Mit welcher W'keit alarmiert die Anlage zu Recht? Mit welcher Wahrscheinlichkeit werden harmlose Kund:innen erschreckt?

$D :=$ "Echter Diebstahl."

$A :=$ "Die Anlage alarmiert."

Wir wissen:

$$P(A|D) = 0.995 \quad P(A|\bar{D}) = 0.006 \quad P(D) = 0.002$$

Wir suchen: $P(D|A) = \frac{P(A|D) \cdot P(D)}{P(A)}$ und $P(\bar{D}|A) = 1 - P(D|A)$.

S. Tafel.

Aufgabe 2.2

$D := \text{"Echter Diebstahl."}$

$A := \text{"Die Anlage alarmiert."}$

- b Wie groß müssten die W'keiten für korrekten und falschen Alarm seien, damit zumindest die Hälfte der Kompromittierten tatsächlich etwas geklaut haben?

Wir wissen:

$$P(D) = 0.002$$

$$P(\overline{D}) = 0.998$$

Wir lösen:

$$P(D|A) \stackrel{!}{=} 0.5$$

S. Tafel.

Aufgabe 2.2

$D := \text{"Echter Diebstahl."}$

$A := \text{"Die Anlage alarmiert."}$

- b Wie groß müssten die W'keiten für korrekten und falschen Alarm seien, damit zumindest die Hälfte der Kompromittierten tatsächlich etwas geklaut haben?

Wir wissen:

$$P(D) = 0.002$$

$$P(\overline{D}) = 0.998$$

Wir lösen:

$$P(D|A) \stackrel{!}{=} 0.5$$

S. Tafel.

$$P(A|D) = 499P(A|\overline{D})$$

Aufgabe 3.1

Betrachten Sie die Ereignisse

- A: Anschnallpflicht befolgt
- K: Schwere Kopfverletzung nach Unfall

Aufgabe 3.1

Betrachten Sie die Ereignisse

- A : Anschnallpflicht befolgt
- K : Schwere Kopfverletzung nach Unfall

$$P(\bar{A}) = 0.15$$

$$P(K|A) = 0.08$$

$$P(\bar{K}|\bar{A}) = 0.62$$

$$P(A) = 0.85$$

$$P(\bar{K}|A) = 0.92$$

$$P(K|\bar{A}) = 0.38$$

- a Interpretieren Sie das Ereignis $\bar{A} \cap K$ und berechnen Sie $P(\bar{A} \cap K)$.

Aufgabe 3.1

Betrachten Sie die Ereignisse

- A : Anschnallpflicht befolgt
- K : Schwere Kopfverletzung nach Unfall

$$P(\bar{A}) = 0.15$$

$$P(K|A) = 0.08$$

$$P(\bar{K}|\bar{A}) = 0.62$$

$$P(A) = 0.85$$

$$P(\bar{K}|A) = 0.92$$

$$P(K|\bar{A}) = 0.38$$

a Interpretieren Sie das Ereignis $\bar{A} \cap K$ und berechnen Sie $P(\bar{A} \cap K)$.

$\bar{A} \cap K$: Der Fahrer ist nicht angeschnallt und hat schwere Kopfverletzung.

$$P(\bar{A} \cap K) = P(K|\bar{A}) \cdot P(\bar{A}) = 0.057$$

Aufgabe 3.2

$$P(\bar{A}) = 0.15$$

$$P(K|A) = 0.08$$

$$P(\bar{K}|\bar{A}) = 0.62$$

$$P(A) = 0.85$$

$$P(\bar{K}|A) = 0.92$$

$$P(K|\bar{A}) = 0.38$$

b Sind die Ereignisse \bar{A} und K stochastisch unabhängig?

Aufgabe 3.2

$$P(\bar{A}) = 0.15$$

$$P(K|A) = 0.08$$

$$P(\bar{K}|\bar{A}) = 0.62$$

$$P(A) = 0.85$$

$$P(\bar{K}|A) = 0.92$$

$$P(K|\bar{A}) = 0.38$$

- b Sind die Ereignisse \bar{A} und K stochastisch unabhängig?

$$P(K) = P(K|A)P(A) + P(K|\bar{A})P(\bar{A}) = 0.125$$

$$P(\bar{A})P(K) = 0.15 \times 0.125 \neq P(\bar{A} \cap K) \Rightarrow \text{Nein}$$

- c Wie groß ist die Wahrscheinlichkeit, dass jemand bei einem Unfall nicht angegurtet war falls eine schwere Kopfverletzung diagnostiziert wurde?

Aufgabe 3.2

$$P(\bar{A}) = 0.15$$

$$P(K|A) = 0.08$$

$$P(\bar{K}|\bar{A}) = 0.62$$

$$P(A) = 0.85$$

$$P(\bar{K}|A) = 0.92$$

$$P(K|\bar{A}) = 0.38$$

- b Sind die Ereignisse \bar{A} und K stochastisch unabhängig?

$$P(K) = P(K|A)P(A) + P(K|\bar{A})P(\bar{A}) = 0.125$$

$$P(\bar{A})P(K) = 0.15 \times 0.125 \neq P(\bar{A} \cap K) \Rightarrow \text{Nein}$$

- c Wie groß ist die Wahrscheinlichkeit, dass jemand bei einem Unfall nicht angegurtet war falls eine schwere Kopfverletzung diagnostiziert wurde?

$$P(\bar{A}|K) = \frac{P(\bar{A} \cap K)}{P(K)} = 0.456$$

Alternativ: Bayes

Aufgabe 4.1

Die Software stuft eine tatsächlich bedrohliche Kommunikation mit einer sehr hohen Wahrscheinlichkeit von 99,5% richtig ein. Die Wahrscheinlichkeit dafür, dass eine harmlose E-Mail fälschlicherweise als potentielle Bedrohung klassifiziert wird, liegt dagegen nur bei 0,5%. In Deutschland gibt es 71.000.000 Internetnutzer:innen. Nachfolgend gehen wir davon aus, dass

- jede:r Nutzer:in täglich 10 unverschlüsselte E-Mails verschiickt,
- 10.000 Nutzer:innen das Internet für die Vorbereitung illegaler und terroristischer Aktivitäten nutzen
- und jede vierte E-Mail aus diesem Personenkreis direkte Kommunikation über eine bedrohliche Aktivität enthält.

Wie groß ist die Wahrscheinlichkeit dafür, dass an einem beliebigen Tag eine durch die Software als potentielle Bedrohung eingestufte E-Mail tatsächlich eine reale Bedrohung aufweist?

Aufgabe 4.2

$$P(B|V) = \frac{P(B \cap V)}{P(V)}$$

$$P(B) = \frac{|B|}{|\Omega|} = \frac{10.000 \cdot 10 \cdot 0.25}{71.000.000 \cdot 10} \approx 0.000035$$

$$P(V) = P(V|B) \cdot P(B) + P(V|\bar{B}) \cdot P(\bar{B}) \approx 0.005$$

$$P(B \cap V) = P(V|B) \cdot P(B) = 0.000033$$

$$P(B|V) = \frac{0.000033}{0.005} = 0.0066 = 0.66\%$$

1 Blatt 3

2 Wiederholung

Zusatzaufgabe 1: Abstraktion

Eine Fabrik produziert elektronische Bauteile mit drei Maschinen (A, B und C). Maschine A produziert 30% der Bauteile, Maschine B 50% und Maschine C die restlichen 20%. Die Wahrscheinlichkeit eines Produktionsfehlers liegt bei Maschine A bei 1%, bei Maschine B bei 2% und bei Maschine C bei 5%.

Ein zufällig ausgewähltes Bauteil weist einen Fehler auf.

Welche der folgenden Aussagen sind korrekt?



Abbildung: Quiz 7 (Aufgabe 1)

1. Gesamtwahrscheinlichkeit eines Fehlers $P(F)$:

$$P(F) = P(F|A)P(A) + P(F|B)P(B) + P(F|C)P(C) = 0.023$$

2. Bedingte Wahrscheinlichkeiten:

- Für Maschine A:

$$P(A|F) = \frac{P(F|A)P(A)}{P(F)} = 0.1304$$

- Für Maschine B:

$$P(B|F) = \frac{P(F|B)P(B)}{P(F)} = 0.4348$$

- Für Maschine C:

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)} = 0.4348$$

In einer klinischen Studie werden Daten zur Diagnose einer seltenen Krankheit X gesammelt. Die Krankheit tritt bei 1 von 10.000 Personen in der Bevölkerung auf. Es gibt einen diagnostischen Test, dessen Sensitivität (Wahrscheinlichkeit, dass der Test positiv ist, wenn die Krankheit vorliegt) bei 99% liegt und dessen Spezifität (Wahrscheinlichkeit, dass der Test negativ ist, wenn die Krankheit nicht vorliegt) bei 95% liegt.

- ① Bestimmen Sie für das Merkmal "Das Ergebnis des diagnostischen Tests (positiv/negativ)" das Skalenniveau.
- ② Welche Erhebungsart ist für diese Studie geeignet?
- ③ Ein Patient erhält ein positives Testergebnis. Berechnen Sie unter Verwendung der Odds-Notation und des Satzes von Bayes die Odds, dass der Patient tatsächlich die Krankheit X hat.

Zusatzaufgabe 2: Kap.1 mit Bayes in Odds

Gegeben:

$$P(K) = 0.0001 \quad P(T|K) = 0.99 \quad P(\bar{T}|\bar{K}) = 0.95$$

$$\text{Prior Odds } = \gamma(K) = \frac{0.0001}{0.9999} = 1 : 9999$$

$$\text{Likelihood Ratio } = \frac{P(T|K)}{P(\bar{T}|\bar{K})} = \frac{0.99}{1 - 0.95} = 99 : 5$$

$$\begin{aligned}\text{Posterior Odds } &= \gamma(K|T) = \text{Prior Odds} \cdot \text{Likelihood Ratio} \\ &= 1 : 9999 \times 99 : 5 = 1 : 505\end{aligned}$$

Die Odds, dass ein Patient mit einem positiven Testergebnis tatsächlich die Krankheit X hat, sind 1 zu 505.

4. Zusammenhangsmaße für diskrete Merkmale

Yichen Han



16. November 2023

1 Wiederholung (Experiment)

2 Blatt 4

Die heutige Wiederholung wird durch ein interaktives multivariates Experiment ersetzt.

Ihr werdet euer Verständnis für die in der Vorlesung behandelten Konzepte vertiefen. Erlebt den Prozess einer deskriptiven Datenanalyse hautnah – von der Datenerhebung bis zur Auswertung.

Die Daten für das Experiment stammen direkt von euch und euren Interaktionen.

Dieses Experiment ist nur als Lehrmittel konzipiert. Es soll keine umfassende wissenschaftliche Forschungsarbeit repräsentieren.

Gaming-Verhalten und Entscheidungsfindung

Ziel des Experiments

- Untersuchung des Zusammenhangs zwischen Spielpräferenzen und Entscheidungsfindungsprozessen.
- Identifikation von Mustern im Spielverhalten und deren Korrelation mit realen Entscheidungsstilen.

Teilnahme

- Daten werden durch eine Online-Umfrage anonym erhoben.
- Keine Vorkenntnisse über Gaming oder Psychologie erforderlich.
- Möglichkeit, Einblicke in die eigene Entscheidungsfindung zu gewinnen.



Abbildung: Füllen Sie diese Umfrage aus

Methoden

- χ^2 -Koeffizient, Kontingenzkoeffizient, Odds Ratio
- Kontingenztafel, Mosaikplots

- f und h gut unterscheiden...
- Stärke des Zusammenhangs sinnvoll vergleichen?
⇒ Immer korrigierter Kontingenzkoeffizient:
Ein Vergleich zwischen $[0, 1]$ ist viel machbarer als in $[0, n \cdot (\min\{m, k\} - 1)]$!
- Einschränkung der Odds Ratio: nur für binäre Ausprägungen anwendbar.
- Interpretation der Zusammenhangsmaße besonders beachten.

1 Wiederholung (Experiment)

2 Blatt 4

Achtung: ausführliche Rechenschritte der komplizierten Zusammenhangsmaße werden gespart!

Aufgabe 1.1

	nein	w	m	hij	LMU
nein	10	14	24	24	24
ja	12	8	20	20	20
hij	22	24	44	44	44

44 Teilnehmer: 27.27%: weiblich UND machen ein Auslandssemester,
63.64% (männlich): kein Auslandssemester machen, insgesamt 45.45%
wollen ein Auslandssemester machen.

- a Erstellen Sie die zugehörige Kontingenztabelle der absoluten Häufigkeiten.

Merkmale: Geschlecht (G) und Bereitschaft für Auslandssemester (A)
Ausprägungen: $\{w, m\}$ und $\{ja, nein\}$.

$$f(ja \cap m) = f(ja) - f(ja \cap w) = 0,1818$$
$$f(nein) = 1 - f(ja) = 0,5455$$
$$f(m) = \frac{f(ja \cap m)}{f(ja \cup m)} = 0,5 \Rightarrow f(w) = 0,5$$
$$f(nein \cap w) = \frac{f(w)}{f(ja \cup m)} = 0,2273 - f(ja \cap w)$$

Aufgabe 1.1

44 Teilnehmer: 27.27%: weiblich UND machen ein Auslandssemester, 63.64% (männlich): kein Auslandssemester machen, insgesamt 45.45% wollen ein Auslandssemester machen.

- a Erstellen Sie die zugehörige Kontingenztabelle der absoluten Häufigkeiten.

Merkmale: Geschlecht (G) und Bereitschaft für Auslandssemester (A)
Ausprägungen: $\{w, m\}$ und $\{ja, nein\}$.

A	G		$h_{i\bullet}$
	w	m	
ja	12	8	20
nein	10	14	24
$h_{\bullet j}$	22	22	44

Aufgabe 1.2

- b) Zsmh. Geschlecht \sim Auslandsaufenthalt? Verwenden Sie hierfür eine geeignete Maßzahl mit dem Wertebereich $[0, 1]$ und interpretieren Sie das Ergebnis.

Aufgabe 1.2

- b) Zsmh. Geschlecht \sim Auslandsaufenthalt? Verwenden Sie hierfür eine geeignete Maßzahl mit dem Wertebereich $[0, 1]$ und interpretieren Sie das Ergebnis.

$$K^* = \frac{\sqrt{\frac{\chi^2}{\chi^2+n}}}{\sqrt{\frac{M-1}{M}}} \in [0, 1], \quad M = \min\{k, l\} = 2$$

Aufgabe 1.2

- b Zsmh. Geschlecht \sim Auslandsaufenthalt? Verwenden Sie hierfür eine geeignete Maßzahl mit dem Wertebereich $[0, 1]$ und interpretieren Sie das Ergebnis.

$$K^* = \frac{\sqrt{\frac{\chi^2}{\chi^2+n}}}{\sqrt{\frac{M-1}{M}}} \in [0, 1], \quad M = \min\{k, l\} = 2$$

chisq.test()

$$\begin{aligned}\chi^2 &= \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \\ &= \frac{44(10 \times 8 - 14 \times 12)^2}{(10+14)(10+12)(14+8)(12+8)} \approx 1.467 \\ K^* &= \frac{\sqrt{1.467/(1.467 + 44)}}{\sqrt{1/2}} \approx 0.254 \Rightarrow \text{schwacher Zsmh.}\end{aligned}$$

Aufgabe 1.3

- c Vergleichen Sie die Chancen für einen geplanten Auslandsaufenthalt zwischen den Geschlechtern mit einer geeigneten Maßzahl.

Aufgabe 1.3

- c Vergleichen Sie die Chancen für einen geplanten Auslandsaufenthalt zwischen den Geschlechtern mit einer geeigneten Maßzahl.

$$\gamma(ja, nein | w, m) = \frac{h_{11} h_{22}}{h_{21} h_{12}} = \frac{12 \cdot 14}{10 \cdot 8} = 2.1$$

Interpretation: Die Chance (Odds) weiblicher Befragten auf einen Auslandsaufenthalt ist etwas mehr als doppelt so hoch wie die Chance der männlichen Befragten.

- d Wie groß ist der Anteil unter den weiblichen Befragten, ein Auslandssemester absolvieren zu wollen?

Aufgabe 1.3

- c Vergleichen Sie die Chancen für einen geplanten Auslandsaufenthalt zwischen den Geschlechtern mit einer geeigneten Maßzahl.

$$\gamma(ja, nein|w, m) = \frac{h_{11}h_{22}}{h_{21}h_{12}} = \frac{12 \cdot 14}{10 \cdot 8} = 2.1$$

Interpretation: Die Chance (Odds) weiblicher Befragten auf einen Auslandsaufenthalt ist etwas mehr als doppelt so hoch wie die Chance der männlichen Befragten.

- d Wie groß ist der Anteil unter den weiblichen Befragten, ein Auslandssemester absolvieren zu wollen?

$$f(ja|w) = \frac{f(ja \cap w)}{f(w)} = \frac{0.2727}{0.5} \approx 0.545$$

Aufgabe 2

$$\textcircled{3} \quad f_{13} = f_{1\bullet} \cdot f_{\bullet 3} = 0.3 \times 0.6 = 0.18 \Rightarrow f_{12} = 0.2$$

$$\textcircled{4} \quad f_{\bullet 1} = \frac{0.3}{0.6} = 0.5$$

	b_1	b_2	b_3	\sum
a_1	0.3	f_{12}	f_{13}	$f_{1\bullet}$
a_2	f_{21}	f_{22}	0.12	0.4
\sum	$f_{\bullet 1}$	$f_{\bullet 2}$	$f_{\bullet 3}$	$f_{\bullet\bullet}$

1

Hinweis: $f_{ab} = f_{a\bullet} \cdot f_{\bullet b}$

$$\textcircled{1} \quad f_{12} + f_{13} = 0.3$$

$$f_{21} + f_{22} = 0.28$$

$$\textcircled{2} \quad f_{\bullet 3} = \frac{f_{23}}{f_{2\bullet}} = 0.3 \Rightarrow f_{21} = 0.68$$

$$\textcircled{5} \quad f_{\bullet 2} = \frac{f_{12}}{f_{1\bullet}} = 0.2$$

Aufgabe 2

	b_1	b_2	b_3	\sum
a_1	0.3	f_{12}	f_{13}	$f_{1\bullet}$
a_2	f_{21}	f_{22}	0.12	0.4
\sum	$f_{\bullet 1}$	$f_{\bullet 2}$	$f_{\bullet 3}$	$f_{\bullet \bullet}$

Hinweis: $f_{ab} = f_{a\cdot} \cdot f_{\cdot b}$

	b_1	b_2	b_3	\sum
a_1	0.3	0.12	0.18	0.6
a_2	0.2	0.08	0.12	0.4
\sum	0.5	0.2	0.3	1

Aufgabe 3.1

Y

5

		Anzahl der reparierten Autos							
		3	5	6	8	10	11	12	15
Anzahl der Beschäftigten	2	3	2	0	0	0	0	0	0
	3	1	2	2	0	0	0	0	0
	5	1	0	4	4	1	0	0	0
	8	0	1	4	5	3	5	2	0
	10	0	0	0	1	1	0	3	5

- a Bestimmen Sie die Randhäufigkeiten von X und Y . Geben Sie außerdem die bedingten relativen Häufigkeiten von Y unter der Bedingung $X = 8$ an.

Aufgabe 3.1

		Anzahl der reparierten Autos								$h_{i\bullet}$
		3	5	6	8	10	11	12	15	
Anzahl der Beschäftigten	2	3	2	0	0	0	0	0	0	5
	3	1	2	2	0	0	0	0	0	5
	5	1	0	4	4	1	0	0	0	10
	8	0	1	4	5	3	5	2	0	20
	10	0	0	0	1	1	0	3	5	10
	$h_{\bullet j}$	5	5	10	10	5	5	5	5	50

- a Bestimmen Sie die Randhäufigkeiten von X und Y . Geben Sie außerdem die bedingten relativen Häufigkeiten von Y unter der Bedingung $X = 8$ an.

b_j	3	5	6	8	10	11	12	15
$f_Y(b_j X = 8)$	0	$\frac{1}{20} = 0.05$	0.2	0.25	0.15	0.25	0.1	0

Aufgabe 3.2

		Anzahl der reparierten Autos								$h_{i\bullet}$
		3	5	6	8	10	11	12	15	
Anzahl der Beschäftigten	2	3	2	0	0	0	0	0	0	5
	3	1	2	2	0	0	0	0	0	5
	5	1	0	4	4	1	0	0	0	10
8		0	1	4	5	3	5	2	0	20
10		0	0	0	1	1	0	3	5	10
$h_{\bullet j}$		5	5	10	10	5	5	5	5	50

- b) Geben Sie unter den Betrieben mit 8 Beschäftigten den Anteil derjenigen an, die 10 Autos repariert haben.

$$f_Y(Y=10|X=8) = \frac{3}{20} = 15\%$$

Aufgabe 3.2

		Anzahl der reparierten Autos								$h_{i\bullet}$
		3	5	6	8	10	11	12	15	
Anzahl der Beschäftigten	2	3	2	0	0	0	0	0	0	5
	3	1	2	2	0	0	0	0	0	5
	5	1	0	4	4	1	0	0	0	10
8		0	1	4	5	3	5	2	0	20
10		0	0	0	1	1	0	3	5	10
$h_{\bullet j}$		5	5	10	10	5	5	5	5	50

- c Geben Sie den Anteil der Betriebe an, die genau 8 Beschäftigte haben und höchstens 10 Autos repariert haben.

$$f(Y \leq 10, X = 8) = \frac{0 + 1 + 4 + 5 + 3}{50} = 26\%$$

Aufgabe 3.3

		Anzahl der reparierten Autos								$h_{i\bullet}$
		3	5	6	8	10	11	12	15	
Anzahl der Beschäftigten	2	3	2	0	0	0	0	0	0	5
	3	1	2	2	0	0	0	0	0	5
	5	1	0	4	4	1	0	0	0	10
	8	0	1	4	5	3	5	2	0	20
	10	0	0	0	1	1	0	3	5	10
h_{\bullet}		5	5	10	10	5	5	5	5	50

- d Geben Sie den Anteil der Betriebe an, die höchstens zehn Autos repariert haben.

$$f(Y \leq 10) = 1 - f(Y > 10) = 1 - \frac{5 + 2 + 3 + 5}{50} = 70\%$$

Aufgabe 3.4

- e Berechnen Sie die bedingten arithmetischen Mittel von Y unter der Bedingung $X = a_i$ für alle $i = 1, \dots, 5$.

Aufgabe 3.4

- e Berechnen Sie die bedingten arithmetischen Mittel von Y unter der Bedingung $X = a_i$ für alle $i = 1, \dots, 5$.

$$\bar{y}_{X=a_i} = \frac{1}{\sum_{j=1}^8 w_j} \sum_{j=1}^8 b_j w_j$$

$f_Y(b_j | a_i)$
 $= f_Y(Y=b_j | X=a_i)$

$$= \frac{\sum_{j=1}^8 b_j \cdot f_Y(b_j | a_i)}{\sum_{j=1}^8 f_Y(b_j | a_i)}$$

$$Y=3 = b_1$$
$$3 \cdot \frac{3}{5}$$

$X \Rightarrow$

a_i	2	3	5	8	10
$\bar{y}_{X=a_i}$	3.8	5	6.9	8.9	12.9

Aufgabe 3.5

f Sind X und Y empirisch unabhängig? Begründen Sie Ihre Entscheidung.

Aufgabe 3.5

f Sind X und Y empirisch unabhängig? Begründen Sie Ihre Entscheidung.

z.Z. $\forall i, j, h_{ij} = \tilde{h}_{ij} = \frac{h_i \cdot h_j}{n} \leftarrow \text{[redacted]} h_i \cdot h_j$

Aufgabe 3.5

f Sind X und Y empirisch unabhängig? Begründen Sie Ihre Entscheidung.

z.Z. $\forall i, j, h_{ij} = \tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n} \Leftrightarrow f_{ij} = f_{i\cdot} f_{\cdot j}$.

Gegenbeispiel: $\tilde{h}_{11} = \frac{5 \times 5}{50} = 0.5 \neq 3 = h_{11}$.

Daraus folgt, Merkmale X und Y sind nicht empirisch unabhängig.

5. Zufallsvariablen und Verteilungsfunktionen

Yichen Han



23. November 2023

1 Wiederholung

2 Blatt 5



Abbildung: Quiz 8

Stichwörter

Zufallsvariable, Träger, Verteilungsfunktion, Treppenfunktion

Indikatorfunktion, Dichte

Empirische Häufigkeitsverteilung, empirische Verteilungsfunktion (ECDF)

Eigenschaften einer Verteilungsfunktion:

- ① $\lim_{x \rightarrow -\infty} F(x) = 0$
- ② $\lim_{x \rightarrow +\infty} F(x) = 1$
- ③ $F(x)$ monoton steigend

Eigenschaften der Dichte:

- ① $\int_{\mathbb{R}} f(x) dx = 1$
- ② $\forall x \in T_X, f(x) \geq 0$

Eigenschaften einer ECDF:

- ① $\forall x < a_1, F_n(x) = 0$
- ② $\forall x \geq a_k, F_n(x) = 1$
- ③ $F_n(x)$ monoton wachsende Treppenfunktion
- ④ rechtsseitig stetig

1 Wiederholung

2 Blatt 5

Aufgabe 1.1

r, g

Auf einer Hauptstraße regeln Ampeln an vier Kreuzungen unabhängig voneinander den Verkehr. Jede von ihnen gestattet oder verbietet die Weiterfahrt mit einer Wahrscheinlichkeit von 0.5. Wir wollen die Anzahl X der Verkehrsampeln, an denen eine gegebene Person ohne Halt vorbeifahren kann, beschreiben.

- (a) Geben Sie die stochastische Grundgesamtheit Ω des hier zugrundeliegenden Laplace-Wahrscheinlichkeitsraums an.

$$\underline{\Omega := \{r, g\}^4} = \{(r, g, r, g), (r, r, g, g), \dots\}$$

Aufgabe 1.1

Auf einer Hauptstraße regeln Ampeln an vier Kreuzungen unabhängig voneinander den Verkehr. Jede von ihnen gestattet oder verbietet die Weiterfahrt mit einer Wahrscheinlichkeit von 0.5. Wir wollen die Anzahl X der Verkehrsampeln, an denen eine gegebene Person ohne Halt vorbeifahren kann, beschreiben.

- (a) Geben Sie die stochastische Grundgesamtheit Ω des hier zugrundeliegenden Laplace-Wahrscheinlichkeitsraums an.

Wir nehmen an, die Ampeln haben zwei Ausprägungen: "r" für rot, "g" für grün.

$$\begin{aligned}\Omega &= \{r, g\}^4 = \{(r, r, r, r), (r, r, r, g), \dots\} \\ |\Omega| &= |\{r, g\}|^4 = 2^4 = 16 \\ &= |\{r, g\}^4|\end{aligned}$$

Aufgabe 1.2

(b) Definieren Sie X und bestimmen Sie die Wahrscheinlichkeitsverteilung

$$X: \Omega \rightarrow T_x := \{0, 1, 2, 3, 4\}$$

$$X(w) = \begin{cases} 0 & w \in \{(r, x, y, z) : \\ & x, y, z \in \{0, 1\}\} \\ 1 & \\ 2 & \\ 3 & \\ 4 & \end{cases}$$

Aufgabe 1.2

(b) Definieren Sie X und bestimmen Sie die Wahrscheinlichkeitsverteilung.

$$\text{ZV } X : \Omega \rightarrow T_X = \{0, 1, 2, 3, 4\}$$

$$X(\omega) = \begin{cases} 0 & \omega \in \{(r, \underline{x}, y, z) | x, y, z \in \{r, g\}\} \\ 1 & \omega \in \{(g, r, x, y) | x, y \in \{r, g\}\} \\ 2 & \omega \in \{(g, g, r, g), (g, g, r, r)\} \\ 3 & \omega \in \{(g, g, g, r)\} \\ 4 & \omega \in \{(g, g, g, g)\} \end{cases}$$

$$P(X=0) = \frac{2}{76} = 0.5$$

Aufgabe 1.2

(b) Definieren Sie X und bestimmen Sie die Wahrscheinlichkeitsverteilung
ZV $X : \Omega \rightarrow T_X = \{0, 1, 2, 3, 4\}$

$$X(\omega) = \begin{cases} 0 & \omega \in \{(r, x, y, z) | x, y, z \in \{r, g\}\} \\ 1 & \omega \in \{(g, r, x, y) | x, y \in \{r, g\}\} \\ 2 & \omega \in \{(g, g, r, g), (g, g, r, r)\} \\ 3 & \omega \in \{(g, g, g, r)\} \\ 4 & \omega \in \{(g, g, g, g)\} \end{cases}$$

$$\Rightarrow f_X(x) = \begin{cases} 8/16 = 0.5 & x = 0 \\ 4/16 = 0.25 & x = 1 \\ 2/16 = 0.125 & x = 2 \\ 1/16 = 0.0625 & x = 3 \\ 1/16 = 0.0625 & x = 4 \\ 0 & sonst \end{cases}$$

Aufgabe 1.3

(c) Bestimmen Sie die Verteilungsfunktion dieser Zufallsvariable.

Aufgabe 1.3

(c) Bestimmen Sie die Verteilungsfunktion dieser Zufallsvariable.

$$F(x) = P(X \leq x) \stackrel{\text{diskret}}{=} \sum_{i: x_i \leq x} f_X(x_i).$$

Aufgabe 1.3

- (c) Bestimmen Sie die Verteilungsfunktion dieser Zufallsvariable.

$$F(x) = P(X \leq x) \stackrel{\text{diskret}}{=} \sum_{i: x_i \leq x} f_X(x_i).$$

Damit gilt: $F(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0.5 & \text{für } 0 \leq x < 1 \\ 0.75 & \text{für } 1 \leq x < 2 \\ 0.875 & \text{für } 2 \leq x < 3 \\ 0.9375 & \text{für } 3 \leq x < 4 \\ 1 & \text{für } x \geq 4. \end{cases}$

- (d) Bestimmen Sie die Wahrscheinlichkeit dafür, dass Verkehrsteilnehmer:innen ohne Anzuhalten mindestens bis zur dritten Ampel fahren dürfen.

$$1 - P(X < 3) = 1 - P(X \leq 2)$$

Aufgabe 1.3

- (c) Bestimmen Sie die Verteilungsfunktion dieser Zufallsvariable.

$$F(x) = P(X \leq x) \stackrel{\text{diskret}}{=} \sum_{i: x_i \leq x} f_X(x_i).$$

Damit gilt: $F(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0.5 & \text{für } 0 \leq x < 1 \\ 0.75 & \text{für } 1 \leq x < 2 \\ 0.875 & \text{für } 2 \leq x < 3 \\ 0.9375 & \text{für } 3 \leq x < 4 \\ 1 & \text{für } x \geq 4. \end{cases}$

- (d) Bestimmen Sie die Wahrscheinlichkeit dafür, dass Verkehrsteilnehmer:innen ohne Anzuhalten mindestens bis zur dritten Ampel fahren dürfen.

$$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - F_X(2) = 0.125$$

Aufgabe 2

Welche der nachfolgenden Funktionen sind Verteilungsfunktionen stetiger Zufallsvariablen? Bestimmen Sie zu allen gültigen Verteilungsfunktionen jeweils die zugehörige Dichtefunktion.

(a) $F(x) = 2x^3 \quad \textcircled{3}$

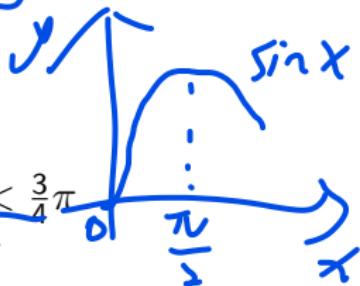


$f(x) = (2x^{-3}) I_{[1, \infty]}$ $F(x) = \begin{cases} 0 & x < 1 \\ 1 - x^{-2} & x \geq 1 \end{cases} \quad \textcircled{1} \quad \textcircled{2}$

(b)

~~$F(x) = \begin{cases} 0 & x < 0 \\ \log(1 + \sin(x)) & 0 \leq x < \frac{\pi}{2} \\ 1 & x \geq \frac{\pi}{2} \end{cases}$~~

$$\begin{aligned} &x < 0 \\ &0 \leq x < \frac{3}{4}\pi \\ &x \geq \frac{3}{4}\pi \end{aligned}$$



~~(c) $F(x) = \begin{cases} 0 & x < 0 \\ e^{-x} & x \geq 0 \end{cases} \quad \textcircled{3}$~~

~~(1)~~

$= e^{-x} I_{[0, \infty]} \quad F(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp(-x) & x \geq 0 \end{cases} \quad \textcircled{2}$

Aufgabe 3.1

Es sei eine Zufallsvariable mit Dichte

$$f(x) = (cx - 6x^2) I(x \in [0, 1]).$$

- (a) Bestimmen Sie die Konstante c so, dass $f(x)$ eine gültige Wahrscheinlichkeitsdichte darstellt.

$$\int_0^1 f(x) dx = 1$$
$$\forall x \in [0, 1] f(x) \geq 0$$

Aufgabe 3.1

Es sei eine Zufallsvariable mit Dichte

$$f(x) = (cx - 6x^2) I(x \in [0, 1]).$$

- (a) Bestimmen Sie die Konstante c so, dass $f(x)$ eine gültige Wahrscheinlichkeitsdichte darstellt.

Zu zeigen: $\int_{\mathbb{R}} f(x) dx = 1$, und $f(x) \geq 0$

Aufgabe 3.1

Es sei eine Zufallsvariable mit Dichte

$$f(x) = (cx - 6x^2) I(x \in [0, 1]).$$

- (a) Bestimmen Sie die Konstante c so, dass $f(x)$ eine gültige Wahrscheinlichkeitsdichte darstellt.

Zu zeigen: $\int_{\mathbb{R}} f(x) dx = 1$, und $f(x) \geq 0$

$$\int_{\mathbb{R}} (cx - 6x^2) I_{[0,1]}(x) dx$$

$$\begin{aligned} \int_0^1 cx - 6x^2 dx &= c \int_0^1 x dx - 6 \int_0^1 x^2 dx \\ &= \frac{1}{2}c[x^2]_0^1 - 6 \left[\frac{1}{3}x^3 \right]_0^1 \\ &= \frac{1}{2}c - 2 \stackrel{!}{=} 1 \\ \Rightarrow \textcolor{blue}{c} &= 6 \end{aligned}$$

$$f(x) = 6x - 6x^2 = 6x(1-x) \geq 0 \text{ für } x \in [0, 1] \quad \checkmark$$

Aufgabe 3.2

(b) Ermitteln Sie die zugehörige Verteilungsfunktion $F(x)$.

Aufgabe 3.2

(b) Ermitteln Sie die zugehörige Verteilungsfunktion $F(x)$.

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(t) dt \\ &= \int_{-\infty}^x (6t - 6t^2) I_{[0,1]}(x) dt \\ &= \int_0^x (6t - 6t^2) dt I_{[0,1]}(x) \\ &= \left(\left[\frac{6}{2}t^2 - \frac{6}{3}t^3 \right]_0^x \right) I_{[0,1]}(x) \\ &= (3x^2 - 2x^3) I_{[0,1]}(x) + I_{[1,\infty)}(x) \end{aligned}$$

aus $\begin{cases} 0, & x < 0 \\ 3x^2 - 2x^3, & x \in [0, 1] \\ 1, & x > 1 \end{cases}$

Aufgabe 3.3

(c) Mit welcher W'keit werden in einer Woche

- i) mehr als 0.8 Hektoliter verbraucht?
- ii) genau 0.5 Hektoliter verbraucht?
- iii) zwischen 0.5 und 0.8 Hektoliter verbraucht?

$$1 - \bar{F}(0.8)$$

$$P(X > 0.8) = 1 - P(X \leq 0.8)$$

$$P(X = 0.5) = 0$$

$$P(0.5 \leq X \leq 0.8)$$

$$= F(0.8) - F(0.5)$$

Aufgabe 3.3

(c) Mit welcher W'keit werden in einer Woche

- i) mehr als 0.8 Hektoliter verbraucht?
- ii) genau 0.5 Hektoliter verbraucht?
- iii) zwischen 0.5 und 0.8 Hektoliter verbraucht?

i)

$$\begin{aligned}P(X > 0.8) &= 1 - F(0.8) \\&= 1 - 0.896 = 0.104\end{aligned}$$

ii)

$$P(X = 0.5) = P(0.5 \leq x \leq 0.5) = F(0.5) - F(0.5) = 0$$

iii)

$$\begin{aligned}P(0.5 \leq X \leq 0.8) &= F(0.8) - F(0.5) \\&= 0.896 - 0.5 = 0.396\end{aligned}$$

Aufgabe 3.4

- (d) Wie hoch ist der Bierverbrauch pro Woche, der mit einer Wahrscheinlichkeit von 0.5 überschritten wird?

Aufgabe 3.4

- (d) Wie hoch ist der Bierverbrauch pro Woche, der mit einer Wahrscheinlichkeit von 0.5 überschritten wird?

$$F(x) = 3x^2 - 2x^3 \stackrel{!}{=} 0.5$$

Numerische (mit dem Rechner) Lösung!

Aufgabe 3.4

- (d) Wie hoch ist der Bierverbrauch pro Woche, der mit einer Wahrscheinlichkeit von 0.5 überschritten wird?

$$F(x) = 3x^2 - 2x^3 \stackrel{!}{=} 0.5$$

Numerische (mit dem Rechner) Lösung!

```
library(rootSolve)
# Define the function representing the equation
equation <- function(x) {
  return(3 * x^2 - 2 * x^3 - 0.5)
}
# Find roots of the equation
solutions <- uniroot.all(equation, c(-1, 2))
print(solutions)
```

Output:

```
[1] 0.5000000 -0.3658408 1.3658408
```

Aufgabe 4.1

Der Besitzer des Kinos *Cinemania* macht sich Gedanken über die Wirtschaftlichkeit seines Hauses. An 100 Tagen zählt er daher die Anzahl der Besucher.

a_j	41	42	43	44	45	46	47	48	49	50	51
h_j	1	9	13	13	20	15	10	7	5	4	3

- (a) Wie heißt das untersuchte Merkmal und wie ist es skaliert?
- (b) Berechnen Sie die relativen und kumulierten relativen Häufigkeiten.

Aufgabe 4.1

Der Besitzer des Kinos *Cinemania* macht sich Gedanken über die Wirtschaftlichkeit seines Hauses. An 100 Tagen zählt er daher die Anzahl der Besucher.

a_j	41	42	43	44	45	46	47	48	49	50	51
h_j	1	9	13	13	20	15	10	7	5	4	3

- (a) Wie heißt das untersuchte Merkmal und wie ist es skaliert?
- (b) Berechnen Sie die relativen und kumulierten relativen Häufigkeiten.

X: Besucherzahl, absolutskaliert.

Aufgabe 4.1

Der Besitzer des Kinos *Cinemania* macht sich Gedanken über die Wirtschaftlichkeit seines Hauses. An 100 Tagen zählt er daher die Anzahl der Besucher.

a_j	41	42	43	44	45	46	47	48	49	50	51
h_j	1	9	13	13	20	15	10	7	5	4	3

- (a) Wie heißt das untersuchte Merkmal und wie ist es skaliert?
(b) Berechnen Sie die relativen und kumulierten relativen Häufigkeiten.

X: Besucherzahl, absolutskaliert.

- Relative Häufigkeit für a_j : $f_j = f(a_j) = \frac{h_j}{n}$ mit $n = 100$
- Kumulierte relative Häufigkeit für a_j : $F_j = \sum_{l=1}^j f_l$

Aufgabe 4.1

Der Besitzer des Kinos *Cinemania* macht sich Gedanken über die Wirtschaftlichkeit seines Hauses. An 100 Tagen zählt er daher die Anzahl der Besucher.

a_j	41	42	43	44	45	46	47	48	49	50	51
h_j	1	9	13	13	20	15	10	7	5	4	3

- (a) Wie heißt das untersuchte Merkmal und wie ist es skaliert?
(b) Berechnen Sie die relativen und kumulierten relativen Häufigkeiten.

X: Besucherzahl, absolutskaliert.

- Relative Häufigkeit für a_j : $f_j = f(a_j) = \frac{h_j}{n}$ mit $n = 100$
- Kumulierte relative Häufigkeit für a_j : $F_j = \sum_{l=1}^j f_l$

51

a_j	41	42	43	44	45	46	47	48	49	50	51
h_j	1	9	13	13	20	15	10	7	5	4	3
f_j	0.01	0.09	0.13	0.13	0.20	0.15	0.10	0.07	0.05	0.04	0.03
F_j	0.01	0.10	0.23	0.36	0.56	0.71	0.81	0.88	0.93	0.97	1

Aufgabe 4.2

(c) Zeichnen Sie die empirische Verteilungsfunktion.

Aufgabe 4.2

42 42 42 .. 9 mal

(c) Zeichnen Sie die empirische Verteilungsfunktion.

```
a_j <- c(41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51)
```

```
h_j <- c(1, 9, 13, 13, 20, 15, 10, 7, 5, 4, 3)
```

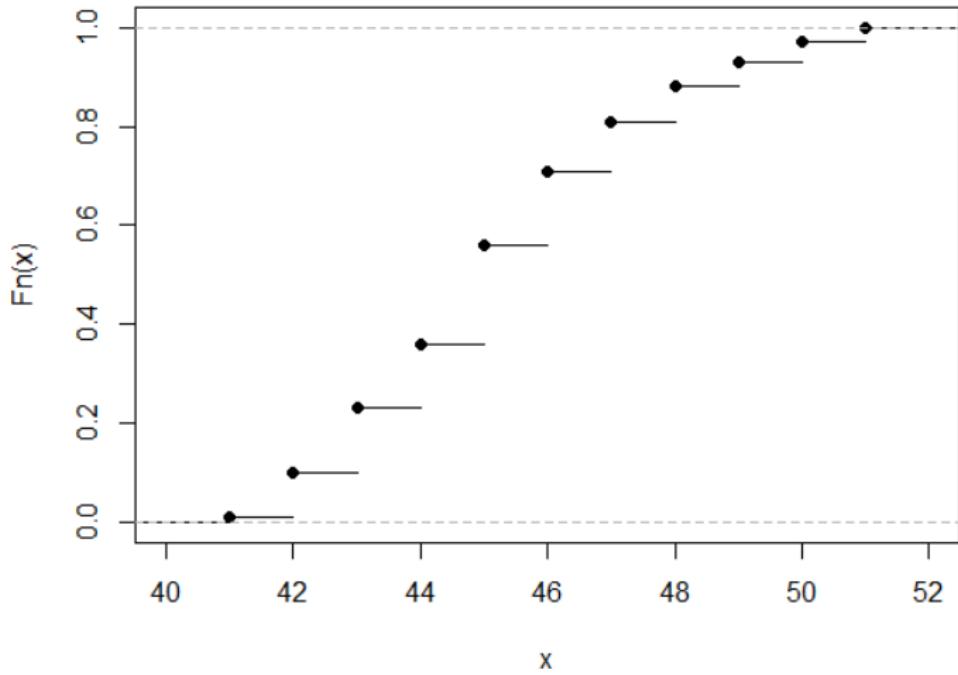
```
data <- data.frame(a_j, h_j)
```

```
sample <- rep(data$a_j, times = data$h_j)
```

```
plot(ecdf(sample),  
      main = "Empirische Verteilungsfunktion der F(x)")
```

Aufgabe 4.2

Empirische Verteilungsfunktion der $F(x)$



Aufgabe 4.3

- (d) Es wird ein Einheitspreis von 6 Euro erhoben. Wie groß ist der Anteil der Tage, an denen der Kinobesitzer weniger als 270 Euro einnimmt?

$$\begin{array}{r} \cancel{270} \\ - 6 \\ \hline F(44) \end{array} = 45$$

Aufgabe 4.3

- (d) Es wird ein Einheitspreis von 6 Euro erhoben. Wie groß ist der Anteil der Tage, an denen der Kinobesitzer weniger als 270 Euro einnimmt?

$$N = \frac{260}{6} = 45$$

⇒ max. 44 Besucher

$$F(44) = \frac{1 + 9 + 13 + 13}{100} = 36\%$$

- (e) Wie groß ist der Anteil der Tage, an denen mindestens 45, aber maximal 50 Besucher kommen?

$$P(45 \leq X \leq 50) - f(45)$$

$$P(45 \leq X \leq 50)$$

Aufgabe 4.3

- (d) Es wird ein Einheitspreis von 6 Euro erhoben. Wie groß ist der Anteil der Tage, an denen der Kinobesitzer weniger als 270 Euro einnimmt?

$$N = \frac{260}{6} = 45$$

⇒ max. 44 Besucher

$$F(44) = \frac{1 + 9 + 13 + 13}{100} = 36\%$$

- (e) Wie groß ist der Anteil der Tage, an denen mindestens 45, aber maximal 50 Besucher kommen?

$$P(45 \leq x \leq 50) = F(50) - F(45) = 0.97 - 0.36 = 61\%$$

Nächste Woche: **Statistische Grafiken**

– ohne Zweifel das wichtigste Thema des Semesters!

Vorbereitung auf das Blatt 6 notwendig! vsl. keine Zeit für Wiederholung.

6. Statistische Grafiken

Übung für Deskriptive Statistik

AUTHOR

Yichen Han

PUBLISHED

November 30, 2023

Code Austausch

Es gibt noch die Zeit, eure Codes für Aufgabe 2 hochzuladen, damit sie präsentiert werden!

Seht bitte die Ankündigung auf Moodle.

Lösung Aufgabe 2 Blatt 5 in Latex

(a) $F(x) = 0 \forall x \leq 1$

$$x^{-2} \xrightarrow{x \rightarrow \infty} 0 \implies F(x) \xrightarrow{x \rightarrow \infty} 1$$

$$F'(x) = 0 \forall x < 1 \text{ und } F'(x) = 0 - (-2)x^{-3} = 2x^{-3} > 0 \forall x \geq 1$$

\implies gültige Verteilungsfunktion.

$$f(x) = \begin{cases} 2x^{-3} & x \geq 1 \\ 0 & \text{sonst} \end{cases}$$

(b) sin ist keine monotone Funktion auf $0 \leq x < \frac{3}{4}$ \implies keine Verteilungsfunktion.

(c) $F(x) = 0 \forall x < 0$

$$\exp(-x) \xrightarrow{x \rightarrow \infty} 0 \implies F(x) \xrightarrow{x \rightarrow \infty} 1$$

$$F'(x) = 0 \forall x < 0 \text{ und } F'(x) = 0 - (-1)\exp(-x) = \exp(-x) > 0 \forall x \geq 0$$

\implies gültige Verteilungsfunktion.

$$f(x) = \begin{cases} \exp(-x) & x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Lösung in Latex

Schlüsselfähigkeiten

- Verteilungsfunktion und Dichte: fortgeschritten (Aufgabe 1)
- Format der statistischen Grafiken kennenzulernen und interpretieren (Aufgabe 4)
- Format der statistischen Grafiken kritisch analysieren und das Design verbessern (Aufgabe 3)

- Statistische Grafiken: selbständige Anwendung unter realen Szenarien mit R-Programmierung (Aufgabe 2)

Zusammenfassung

- Grammar of Graphics
 - Geometrien: Linie, Balken, Punkte, Fläche, etc.
 - Ästhetische Zuordnungen: Farbe, Größe, Form, Position, etc.
 - Skalen: lineare Skala, log-Skala, etc.
 - Koordinatensysteme: kartesisches Koordinatensystem, Polarkoordinatensystem, etc.
 - Facetten: Aufteilung der Grafik in mehrere Panels
 - etc. (sehr fortgeschritten)
- Farbskalen
 - Qualitative Farbskala: Farben ohne Ordnung
 - Sequentielle Farbskala: Farben mit Ordnung
 - Divergierende Farbskala: Farben mit Ordnung und Mittelwert / Neutralpunkt
- Hierarchie der Wahrnehmung
 - Position
 - Abstand & Länge
 - Steigung & Winkel
 - Fläche
 - Volumen
 - Farben
- Gelernte Grafiken
 - Mosaikplot
 - Histogramm
 - Streudiagramm
 - Liniendiagramm
 - Balkendiagramm / Säulendiagramm / Stapeldiagramm
 - Kreisdiagramm (immer vermeiden)
 - etc.
- Kerndichteschätzer (KDE)
 - Motivation: Histogramm ist abhängig von der Anzahl der Bins bzw. Kantenlänge der Bins.

- Lösung: Ersetze Histogramm durch glatte Funktion $\hat{f}(x)$.
- Kerndichteschätzer: $K(\cdot)$ für Kernfunktion, h für Bandbreite.

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)$$

- Weitere Vorteile: Berücksichtigung der Entfernung der beobachteten Werte mit abnehmender Gewichtung.
- Abhängigkeit von h : zu klein (\Rightarrow) zu viele Details, zu groß (\Rightarrow) zu wenig Details.
(Goodness of fit / Anpassungsgüte)

Aufgabe 1

$f(x) = a \frac{x^3 - x^2 + x - 1}{x-1} \mathbb{I}_{[2,5]}(x)$ Hinweis: Vereinfachen Sie zuerst den Ausdruck der Dichte.
 $\begin{aligned} x^3 - x^2 + x - 1 &= (x-1)(x^2+1) \Rightarrow f(x) = a(x^2+1)\mathbb{I}_{[2,5]}(x) \\ \end{aligned}$

(a) Bestimmen Sie den Wert von a .

$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \int_{-\infty}^2 0 dx + \int_2^5 a(x^2+1) dx = a \left[\frac{x^3}{3} + x \right]_2^5 = a \left(\frac{125}{3} + 5 \right) = a \frac{140}{3} \\ a &= \frac{1}{42} \text{ (Es gilt: } x^2+1 \geq 1 \text{ für alle } x \in \mathbb{R}, \text{ also } f(x) \geq 0 \text{ für alle } x \in \mathbb{R} \text{ mit } a = \frac{1}{42}) \end{aligned}$

(b) Bestimmen Sie die Verteilungsfunktion $F(x)$.

Warum "fortgeschritten"?

Das gängige Vorgehen: Dichte integrieren und mit passenden Indikatorfunktionen modifizieren.

Hier: $G(x) := \int_{-\infty}^x f(x) dx, G(2) \neq 0$

Deswegen brauchen wir eine Konstante c , sodass $(G(2) + c = 0)$.

Wir lösen die Gleichung, und erhalten $c = -\frac{1}{9}$.

$F(x) = \left(\frac{1}{42} \left(\frac{x^3}{3} + x \right) - \frac{1}{9} \right) \mathbb{I}_{[2,5]}(x) + \mathbb{I}_{(-\infty, 2)}(x)$

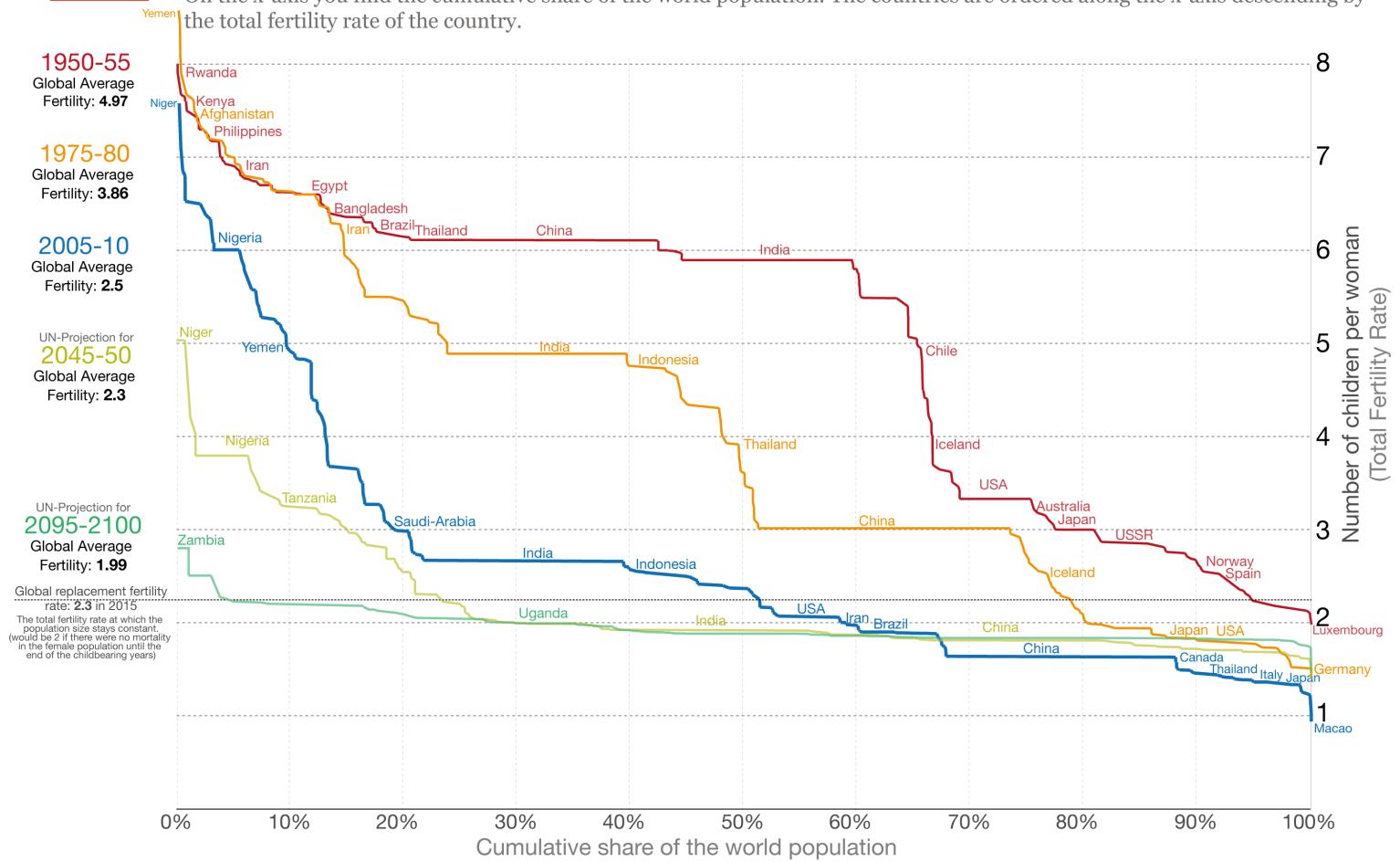
(c) Berechnen Sie $P(3 < X < 4)$.

$P(3 < X < 4) = F(4) - F(3) = 0.3175$

Aufgabe 4

World population by level of fertility over time (1950-2100)

On the x-axis you find the cumulative share of the world population. The countries are ordered along the x-axis descending by the total fertility rate of the country.



Data source: United Nations Population Division (2012 revision).

The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic.

Licensed under CC-BY-SA by the author Max Roser.

Weltbevölkerung und Fruchtbarkeitsniveau über die Zeit

(a) Welche Untersuchungseinheiten aus welcher Grundgesamtheit werden in der Grafik dargestellt?

- GG: Länder der Erde
- UE: einzelne Länder

(b) Was für eine Erhebungsart und Datenstruktur liegen hier vor?

Erhebungsart:

- Vollerhebung,
- Längsschnittdaten (mehrere Beobachtungen mit mehreren Merkmalen pro UE.)

(c) Welches Skalenniveau haben Gesamtfruchtbarkeitsrate und Bevölkerungsanteil jeweils?

- Gesamtfruchtbarkeit: (Anzahl Kinder), absolutskaliert
- Bevölkerungsanteil: verhältnisskaliert (Prozent)

(d) Sind die auf der linken Seite der Grafik angegeben Zeiträume die Ausprägungen eines ordinal-, nominal- oder intervallskalierten Merkmals? Begründen Sie Ihre Antwort kurz.

ordinalskaliert: in Form von Intervallen und hat Ordnung.

(Alternativ: intervallskaliert, da die Intervalle gleich groß sind und die Differenz ist einigermaßen interpretierbar. –

nicht zu empfehlen)

(e) Was für eine Art von Farbskala wurde in der Grafik verwendet? Welche Art von Farbskala wäre hier eventuell besser geeignet und warum?

- qualitative Farbskala
- besser: sequentielle Farbskala, da die Farben (Zeitraum) eine Ordnung haben.

(f) Welche “Geometrie” wird hier zur Darstellung benutzt? Geben Sie für alle in der Grafik gezeigten Merkmale die verwendeten ästhetischen Zuordnungen an.

- Geometrien: Linie bzw. Treppenfunktion

Ästhetische Zuordnungen:

- Zeitraum: Farbe
- Bevölkerungsanteil: x-Koordinate
- Gesamtfruchtbarkeitsrate (TFR): y-Koordinate

(g) Welche ästhetischen Eigenschaften welcher Geometrien würden Sie für welche Merkmale verwenden, um in einer wohlüberlegten statistischen Grafik auf Basis dieser Daten die zeitlichen Entwicklungen der Gesamtfruchtbarkeitsraten zwischen ausgewählten Ländern einfach vergleichbar zu machen?

Bonus: Welche ggplot2-Befehle produzieren eine solche Grafik?

- Geometrien: Linie

Ästhetische Zuordnungen:

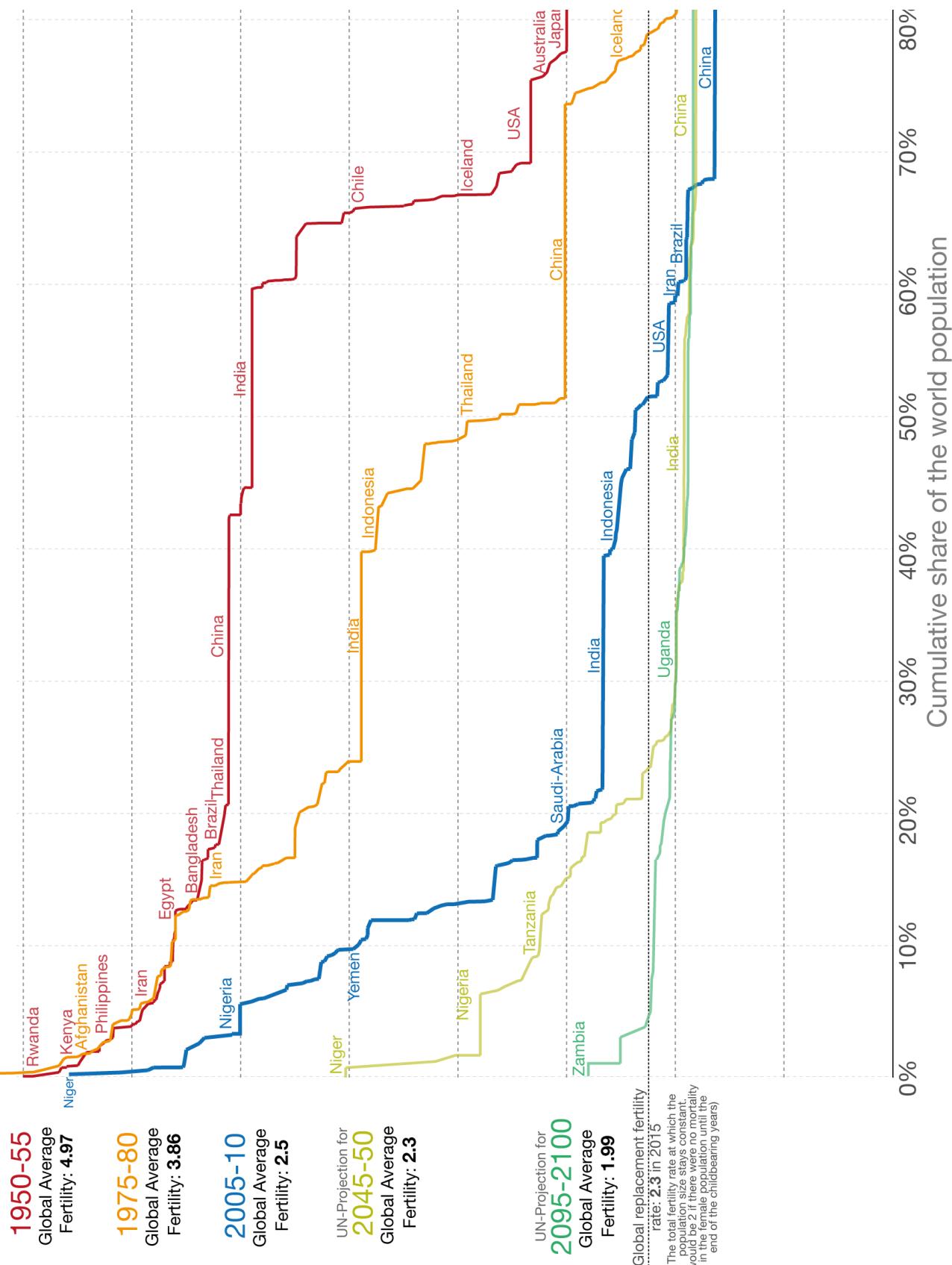
- Zeitraum: x-Koordinate
- Gesamtfruchtbarkeitsrate (TFR): y-Koordinate
- Land: Farbe

```
ggplot(data = fertility_gini, aes(x = year, y = tfr, color = country)) +  
  geom_line()
```

(h) Betrachten Sie die in der Grafik in rot eingezeichnete Linie. Stellen Sie sich vor, wir vertauschen die horizontalen und vertikalen Achsen der Grafik durch eine Rotation um 90° den Uhrzeigersinn. Wäre die dadurch entstehende Funktion äquivalent zur empirischen Verteilungsfunktion der Gesamtfruchtbarkeitsraten aller Länder der Erde im angegebenen Zeitraum? Begründen Sie Ihre Antwort.



World population by level of fertility over time (On the x-axis you find the cumulative share of the world population. The countries are ordered a. the total fertility rate of the country.

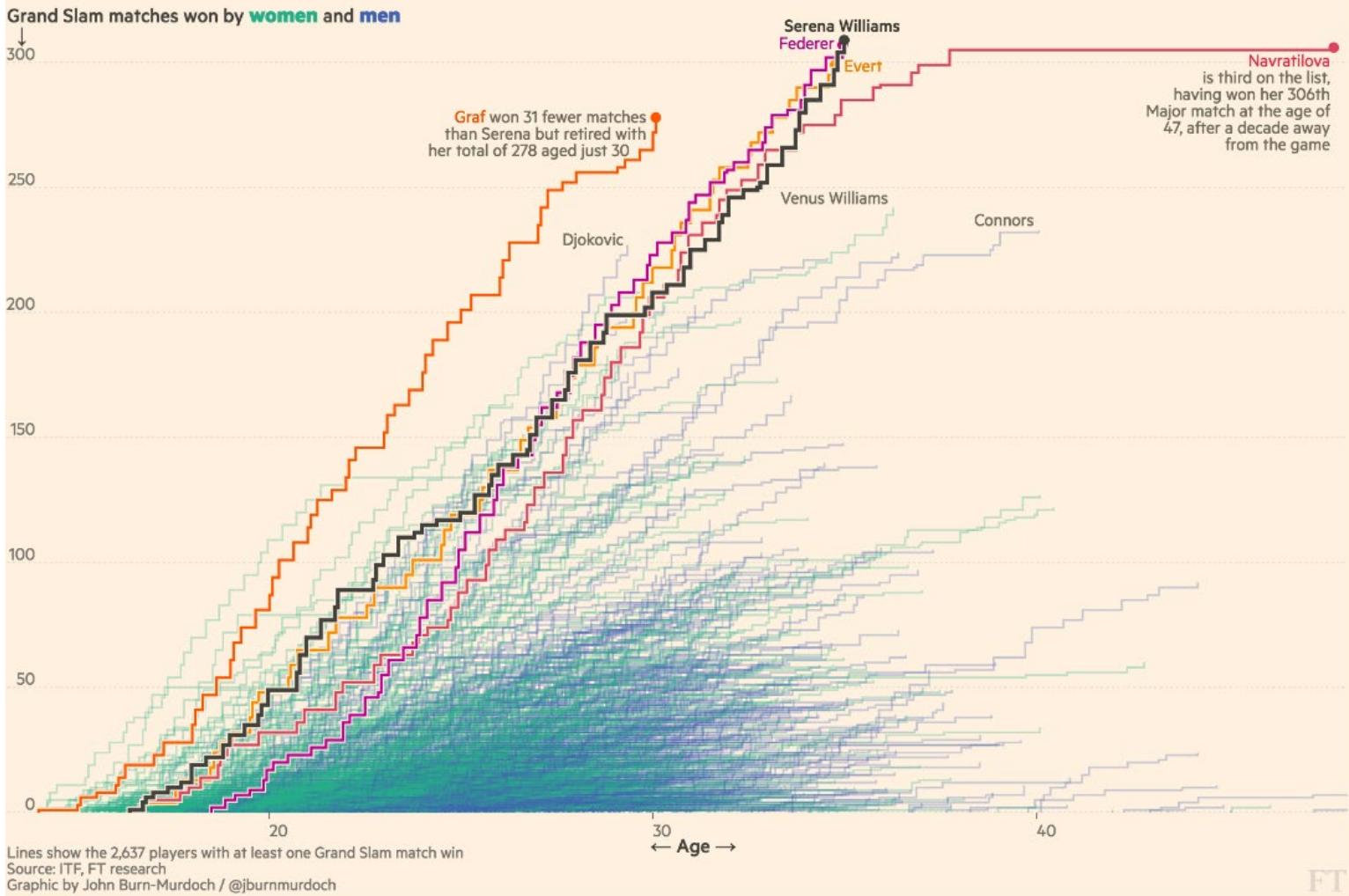


Umgedrehte Grafik

Nein, die Funktion ist keine ECDF:

- x-Achse: die Zahlen gehen rückwärts, von 8 bis 1.
- y-Achse: steht nicht für Anteil der UE, sondern für den Anteil der Weltbevölkerung.

Aufgabe 3



Grand Slam Matches

FT

(a) Beschreiben Sie die Datengrundlage dieser Grafik. Was sind die Untersuchungseinheiten, Beobachtungen, Merkmale? Welche Erhebungsart liegt hier vor?

UE: Tennisprofis

Merkmale:

- Alter
- Geschlecht
- Anzahl der Grand Slam Matches
- optional: Name

Erhebungsart:

- Vollerhebung,
- Längsschnittdaten (mehrere Beobachtungen mit mehreren Merkmalen pro UE.)

(b) Analysieren Sie diese Grafik formal mit den Begriffen der in der Vorlesung eingeführten “Grammar of Graphics” - Welche ästhetischen Mappings werden hier wie verwendet? Welche Geometrien, Skalen, etc. werden hier wie verwendet?

- Geometrien: Linien

Ästhetische Mappings:

- Alter: x-Koordinate / horizontal
- Anzahl der Grand Slam Matches: y-Koordinate / vertikal
- Geschlecht (+ Name): Farbe

Skalen:

- kartesisches Koordinatensystem
- y-Achse: 0 bis ca. 300
- x-Achse: 0 bis ca. 50

(c) Inwiefern weicht die Grafik von der in der Vorlesung definierten einfachen Grammatik ab?

- Die Grafik hat keine Legende. Die Farben werden nur implizit erklärt (als Bestandteil der Überschrift).
- Die Grafik hat zwei Farbskalen.
- Die Text-Annotationen sind direkt in der Grafik enthalten.

(d) Sie interessieren sich besonders dafür, wie bzw. ob sich die Karriereverläufe weiblicher und männlicher Tennisprofis unterscheiden. Welche Elemente der Darstellung würden Sie wie verändern, entfernen oder hinzufügen um eine neue Grafik zu erzeugen, die diese Frage besser beantworten kann?

Problem: Overplotting

Was bedeutet das?

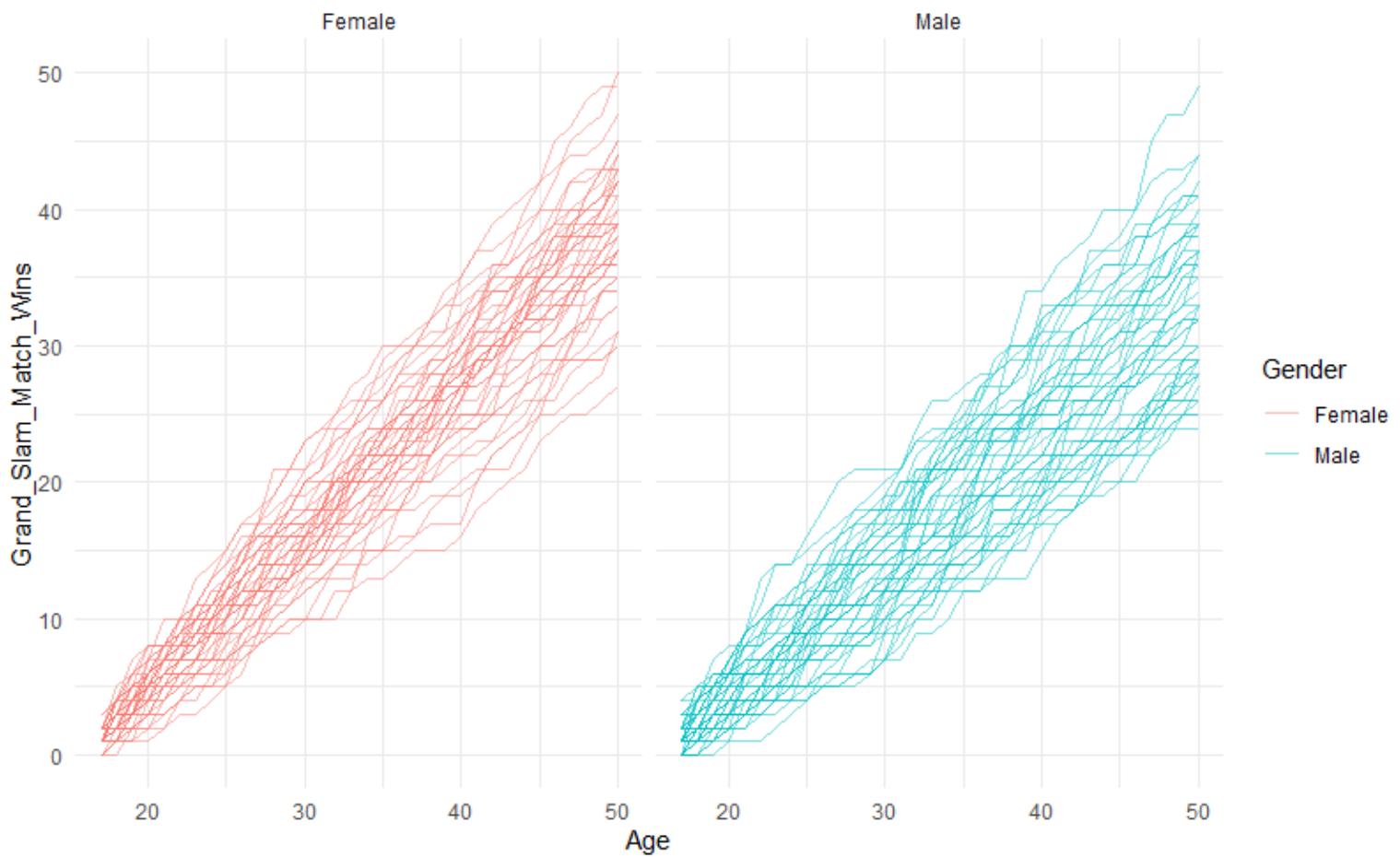
Definition: Es tritt auf, wenn zu viele Datenpunkte auf einem kleinen Bereich eines Diagramms gezeichnet werden, sodass einzelne Punkte nicht mehr klar unterscheidbar sind und die wahren Datenstrukturen oder Muster dadurch verdeckt werden.

Lösung 1

- Facettierung nach Geschlecht
- Keine Hervorhebung der einzelnen Spieler
- (optional) Transparanz der einzelnen Linien anpassen und Linien für Mittelwert hinzufügen.

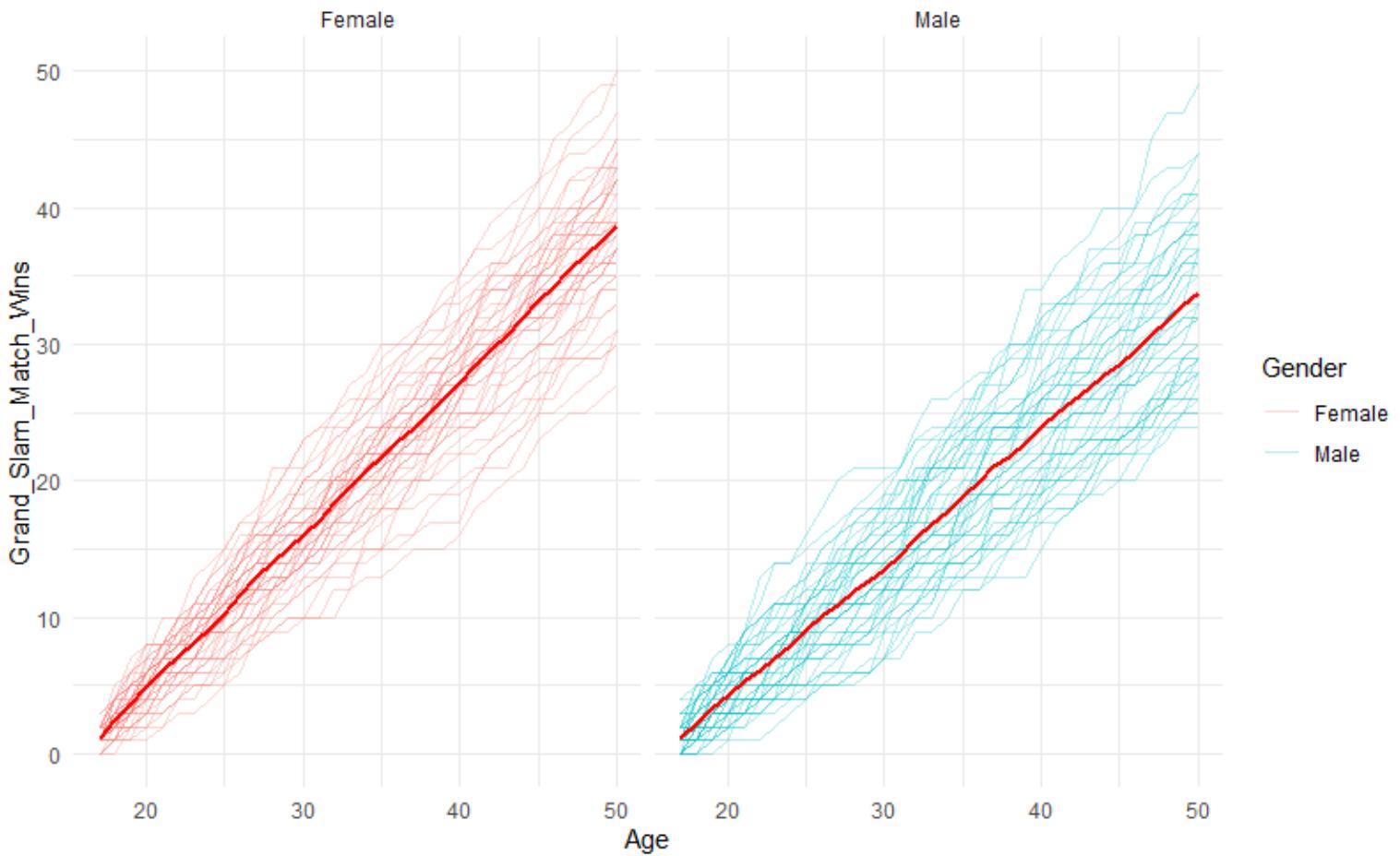
Verteilung der Grand Slam Match Wins nach Alter und Geschlecht

Mit zufällig simulierten Daten



Verteilung der Grand Slam Match Wins mit Mittelwertlinien

Mit zufällig simulierten Daten

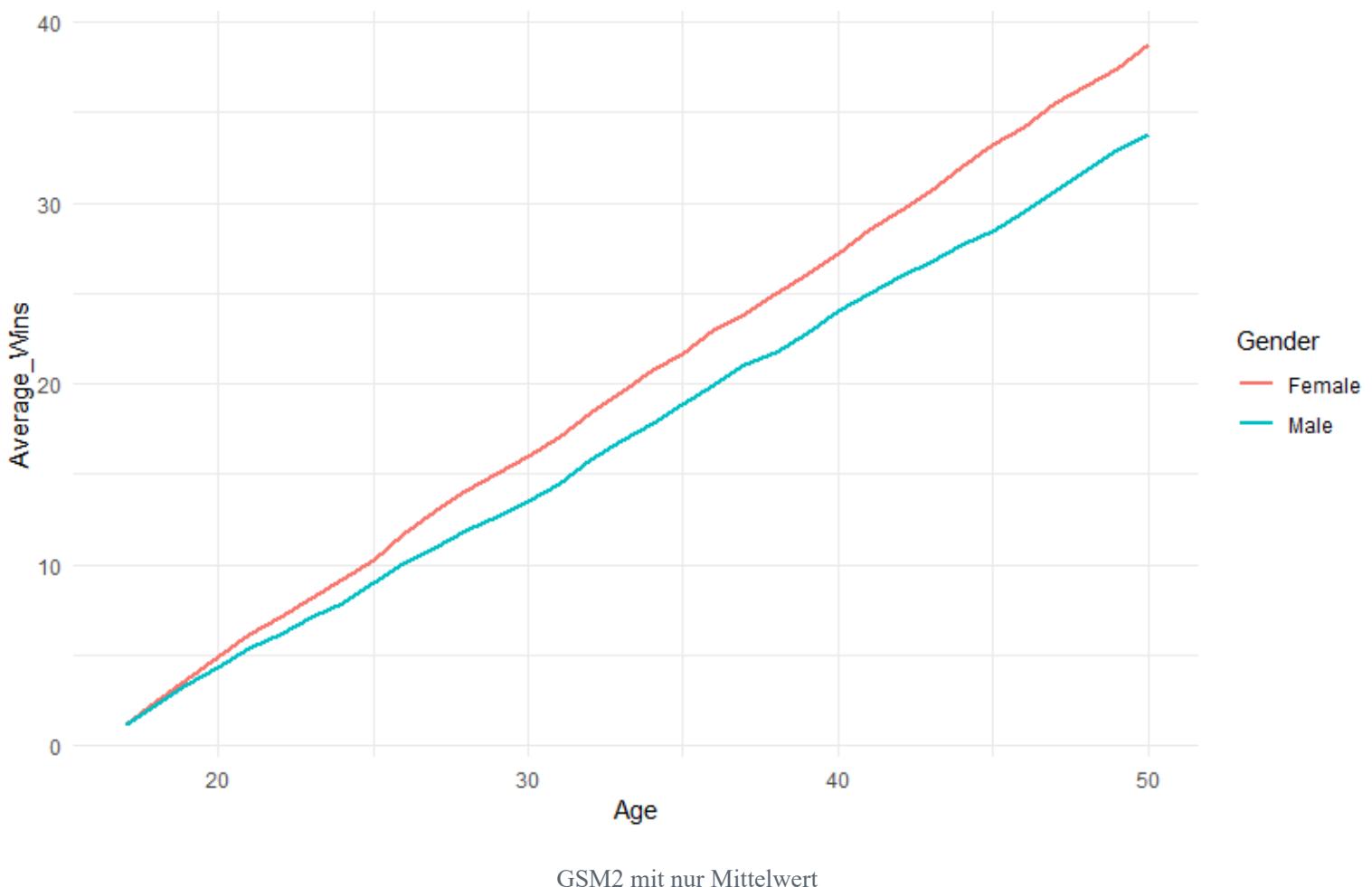


Lösung 2 (nicht zu empfehlen)

- Eine Farbskala in einer Grafik, bezeichnet durch zwei Mittelwertlinien.

Mittelwert der Grand Slam Match Wins nach Alter und Geschlecht

Mit zufällig simulierten Daten,
Verlust der Information über einzelne Spieler und mögliche Verfälschung



GSM2 mit nur Mittelwert

Habt ihr sonst noch Lösungsvorschläge?

Aufgabe 2

Variable	Labels
sex	Geschlecht 1 = männlich 2 = weiblich
sleep_hours_work	Anzahl der Stunden im Schlafen (Woche)
sleep_hours_wkend	Anzahl der Stunden im Schlafen (Wochenende)
snore_freq	Häufigkeit vom Schnarchen mit 4 Kategorien: 0 = nie 1 = selten

2 = gelegentlich

3 = häufig

snort_freq

Häufigkeit von Apnoe

mit 4 Kategorien:

0 = nie

1 = selten

2 = gelegentlich

3 = häufig

sleep_freq_day

Häufigkeit der schlafirgen Gefühle im Tag

mit 5 Kategorien:

0 = nie

1 = selten

2 = manchmal

3 = oft

4 = fast immer

Vorbereitung

```
library(ggplot2)
library(dplyr)
library(knitr)

# Daten einlesen
sleep <- read.csv("nhanes_sleep1718.csv")

# Wertelabels zuweisen
wertelabels <- list(
  snore_freq = c("nie", "selten", "gelegentlich", "häufig"),
  snort_freq = c("nie", "selten", "gelegentlich", "häufig"),
  sleepy_freq_day = c("nie", "selten", "manchmal", "oft", "fast immer")
)

# Gehe jede Variable durch und weise die Wertelabels zu
for (variable in names(wertelabels)) {
  # Beginne bei 0 für die Levels, da die Zählung der Codes bei 0 beginnt
  # Benutze "ordered" für ordinale Variablen
  sleep[[variable]] <- ordered(sleep[[variable]],
                                levels = 0:(length(wertelabels[[variable]]) - 1),
                                labels = wertelabels[[variable]])
}

sleep$sex <- factor(sleep$sex,
                     levels = c(1, 2),
                     labels = c("männlich", "weiblich"))

kable(head(sleep))
```

ID sex	sleep_hours_work	sleep_hours_wkend	snore_freq	snort_freq	sleepy_freq_day
93705 weiblich	8.0	8.0	gelegentlich	nie	nie
93706 männlich	10.5	11.5	selten	nie	selten

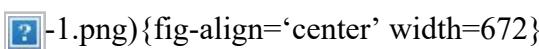
93708	weiblich	8.0	8.0	nie	nie	manchmal
93709	weiblich	NA	6.5	selten	nie	selten
93711	männlich	NA	NA	gelegentlich	selten	oft
93712	männlich	7.5	NA	selten	selten	manchmal

a) Häufigkeitstabelle und ein Säulendiagramm

```
table_sleepy <- prop.table(table(sleep$sleepy_freq_day))
table_sleepy |> round(2) |> kable() # pipeline operator: |>
```

Var1	Freq
nie	0.17
selten	0.24
manchmal	0.33
oft	0.17
fast immer	0.09

```
barplot(table_sleepy,
       main="Relative Häufigkeiten von 'sleepy_freq_day'",
       xlab="sleepy_freq_day", ylab="Relative Häufigkeit")
```



b) Analyse mit Geschlecht

```
# Stichprobe nach Geschlecht unterteilen und Größe ermitteln
maenner <- subset(sleep, sex == "männlich")
frauen <- subset(sleep, sex == "weiblich")
cat("Anzahl der männlichen Befragten:", nrow(maenner), "\n")
```

Anzahl der männlichen Befragten: 2992

```
cat("Anzahl der weiblichen Befragten:", nrow(frauen), "\n")
```

Anzahl der weiblichen Befragten: 3169

```
# oder:
kable(table(sleep$sex))
```

Var1	Freq
männlich	2992
weiblich	3169

```
# Säulendiagramm für die relativen Häufigkeiten von sleepy_freq_day für weibliche Befragte
frauen_sleepy <- prop.table(table(frauen$sleepy_freq_day))
# base-R plot
barplot(frauen_sleepy,
```

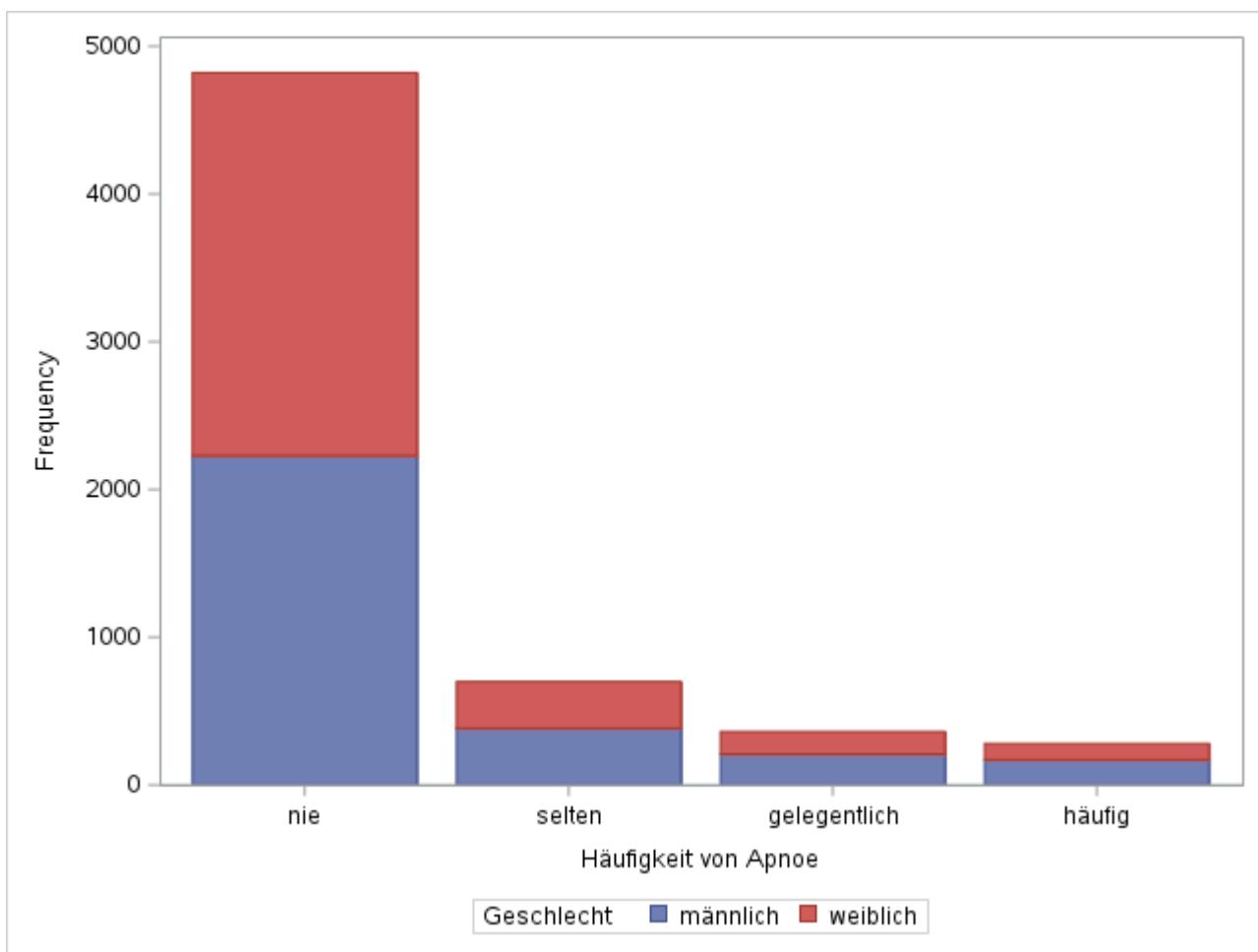
```
main="Relative Häufigkeiten von 'sleepy_freq_day' bei Frauen",
xlab="sleepy_freq_day", ylab="relative Häufigkeit")
```

QUESTION-1.png){fig-align='center' width=672}

```
# mit ggplot2: (weight = 1/n damit rel. Häufigkeiten geplottet werden)
ggplot(frauen) +
  geom_bar(aes(x = sleepy_freq_day, weight = 1/nrow(frauen)))
```

QUESTION-2.png){fig-align='center' width=672}

c) Kritik an Grafik



Beispieldiagramm

Problemstellung:

Schwierigkeit beim Vergleich von Apnoe-Häufigkeiten zwischen Männern und Frauen **aufgrund unterschiedlicher Anzahl an Personen, die nie bzw. öfter als nie an Apnoe leiden.**

Beispielsweise sind die Angaben unter der Kategorie "häufig" auf dem Diagramm nahezu unvergleichbar klein.

Stapeldiagramm mit absoluten Häufigkeiten (Beispiel):

- Ungenau für den Vergleich, da unterschiedliche Personenzahlen die Interpretation erschweren.
- Kategorie "häufig" kaum vergleichbar wegen geringer sichtbarer Unterschiede.

Säulendiagramm mit nebeneinander angeordneten Säulen:

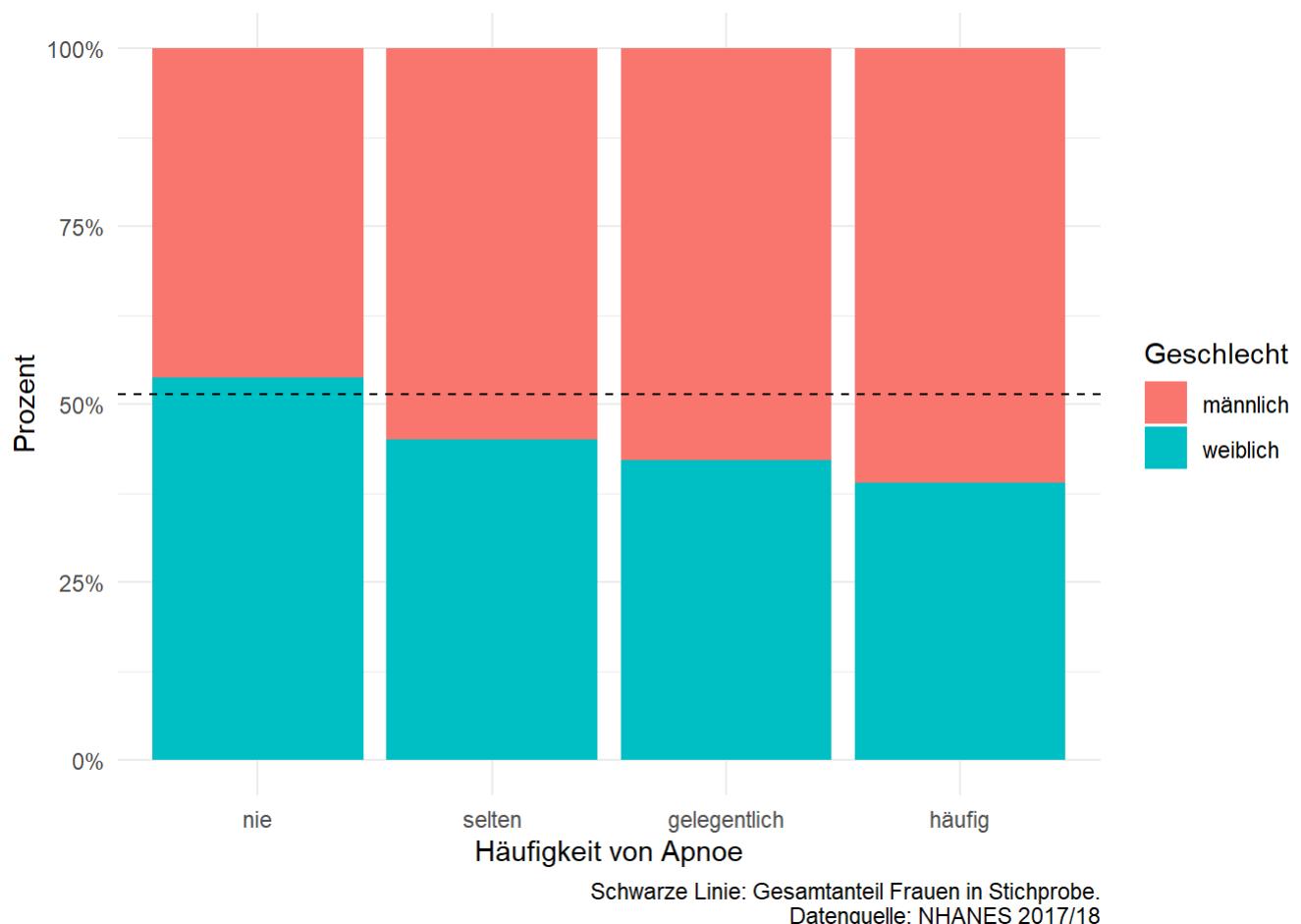
- Bessere Darstellung der absoluten Häufigkeiten durch direkten Vergleich an derselben Achse.
- Ermöglicht den Vergleich der Anzahl von Männern und Frauen für jede Apnoe-Häufigkeit.

Empfehlung: 100%-Stapeldiagramm:

- Ideal für die Fragestellung, da es die prozentualen Unterschiede in jeder Kategorie unabhängig von der Gesamtzahl der Beobachtungen zeigt.
- Erleichtert den direkten Vergleich der Geschlechter innerhalb jeder Kategorie.

d) Optimale Grafik

```
# Version 1: bedingten Hfgk von Geschlecht|Schnarchen
ggplot(sleep) +
  # gestapeltes Balkendiagramm der bedingten Hfgk von Geschlecht|Schnarchen
  geom_bar(aes(x = snort_freq, fill = sex), position = "fill") +
  # Referenzlinie: insgesamter Anteil der Frauen:
  geom_hline(yintercept = mean(sleep$sex == "weiblich"), linetype = 2) +
  # schöne Labels & captions:
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Häufigkeit von Apnoe", y = "Prozent", fill = "Geschlecht",
       caption = "Schwarze Linie: Gesamtanteil Frauen in Stichprobe.
                  Datenquelle: NHANES 2017/18") +
  theme_minimal()
```

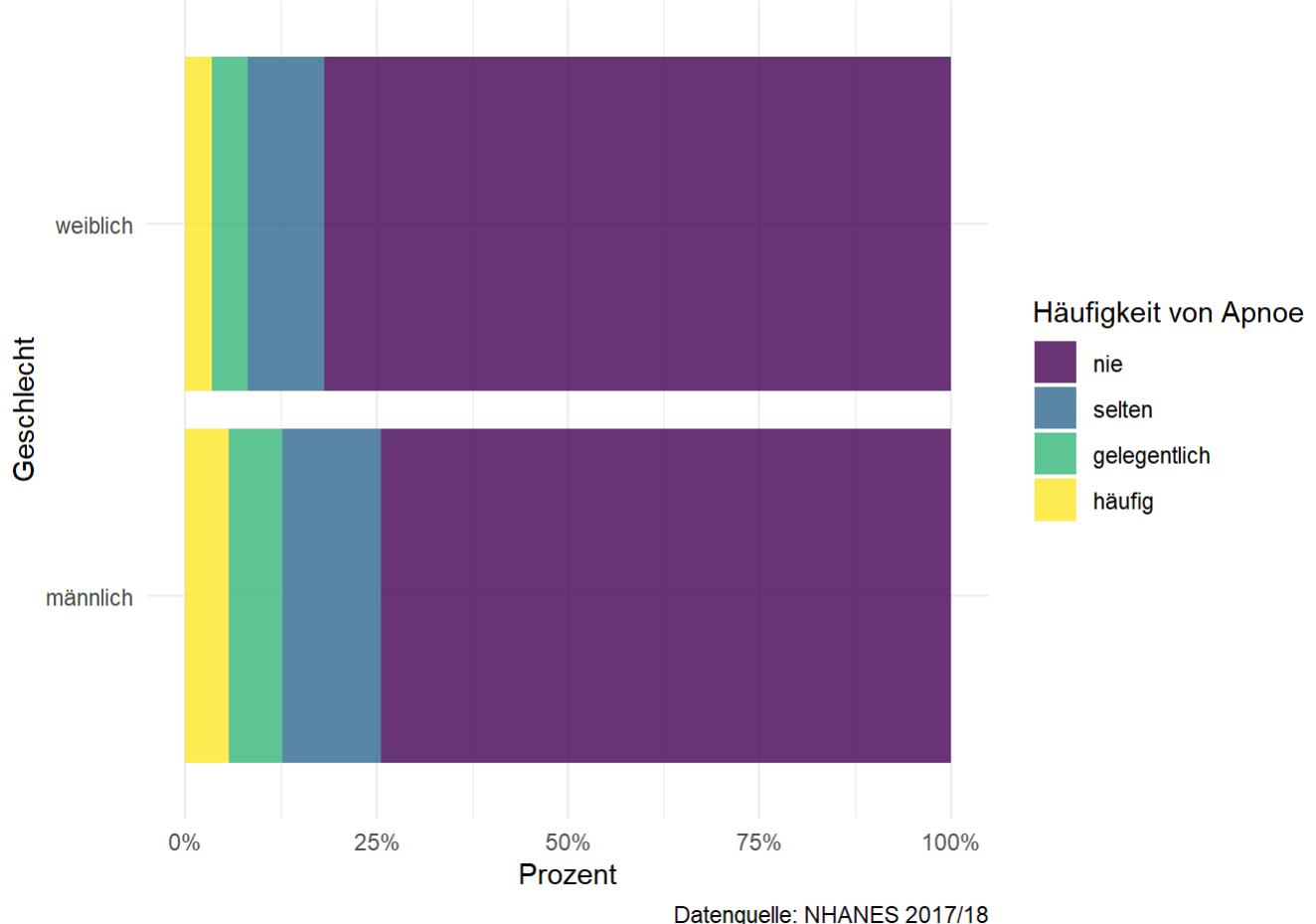


```
# Version 2 (besser): bedingten Hfgk von Schnarchen|Geschlecht
```

```

ggplot(sleep) +
  # gestapeltes Balkendiagramm der bedingten Hfgk von Schnarchen|Geschlecht
  geom_bar(aes(fill = snort_freq, x = sex), alpha = .8, position = "fill") +
  # snort_freq ist ordinal, also *sequentielle* farbskala:
  scale_fill_viridis_d() +
  # horizontale balken mit horizontalen textlabels sind besser lesbar:
  coord_flip() +
  # schöne Labels & captions:
  scale_y_continuous(labels = scales::percent) +
  labs(fill = "Häufigkeit von Apnoe", y = "Prozent", x = "Geschlecht",
       caption = "Datenquelle: NHANES 2017/18") +
  theme_minimal()

```



Männliche Befragte weisen höhere Anteile an “selten”, “gelegentlich” und “häufig” Antworten auf, während weibliche Befragte mehr “nie” Antworten berichten. Dies deutet darauf hin, dass unter den Befragten männliche Personen tendenziell häufiger über Apnoe-Episoden berichten als weibliche Personen.

7. benotete Hausaufgabe II

A3d, A4cde
Beweis (Kriterien)
Wahl der Maßzahlen

Korrektur Ende
nächster Woche,
hoffentlich...

Yichen Han



7. Dezember 2023

Aufgabe 1

Eine Studie von Eriksson, Wu et al. (2012) zeigt, dass die Präferenz oder Aversion gegenüber dem Geschmack von Koriander, den einige als seifenartig beschreiben, auf genetische Faktoren zurückzuführen ist. Basierend auf den modifizierten Daten aus der Studie werden in dieser Aufgabe weiterführende Analysen durchgeführt. Wir betrachten für die ganze Aufgabe denselben Datensatz und nehmen an, dass es keine fehlenden Daten gibt.

- Zusammenhangsmaße für diskrete Merkmale
- Kontingenztafeln und Häufigkeiten

Aufgabe 1.1

	Geschmack		Summe
	wie Seife	nicht wie Seife	
Männlich	865	6437	7302
Weiblich	1129	6173	7302
Summe	1994	12610	14604

Tabelle: Geschmack von Koriander nach Geschlecht

- (a) (5 P) Analyse der empirischen Abhängigkeit von Geschlecht und Wahrnehmung des Geschmacks von Koriander. Benutzen Sie hier ein Maß, das **nur die Stärke, aber nicht die Richtung des Zusammenhangs quantifiziert**.

Aufgabe 1.1

	Geschmack		Summe
	wie Seife	nicht wie Seife	
Männlich	865	6437	7302
Weiblich	1129	6173	7302
Summe	1994	12610	14604

Tabelle: Geschmack von Koriander nach Geschlecht

- (a) (5 P) Analyse der empirischen Abhängigkeit von Geschlecht und Wahrnehmung des Geschmacks von Koriander. Benutzen Sie hier ein Maß, das **nur die Stärke, aber nicht die Richtung des Zusammenhangs quantifiziert**.

Beste Wahl: korrigierter Kontingenzkoeffizient ($K^* \in [0, 1]$).

Alternativ: χ^2 -Koeffizient $\in [0, n(\min(k, m) - 1)]$

Falsch: Odds Ratio (Chancenvergleich & mit Richtung)

Aufgabe 1.2

$$\chi^2 = n \sum_i \sum_j \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \quad (1)$$

⇒ die Tafel in relative Häufigkeiten umwandeln und Werte einsetzen.

Aufgabe 1.2

$$\chi^2 = n \sum_i \sum_j \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \quad (1)$$

⇒ die Tafel in relative Häufigkeiten umwandeln und Werte einsetzen.
Oder für 2x2-Tafeln:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \quad (2)$$

$$\chi^2 = \frac{14604(865 \times 6173 - 1129 \times 6437)^2}{7302 \times 1994 \times 12610 \times 7302} \approx 40.48 \text{ (40, 40.5)},$$

mit $0 < \chi^2 \ll n(\min(k, m) - 1) = 14604 \Rightarrow$ der Koeffizient ist klein.

Aufgabe 1.2

$$\chi^2 = n \sum_i \sum_j \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \quad (1)$$

⇒ die Tafel in relative Häufigkeiten umwandeln und Werte einsetzen.
Oder für 2x2-Tafeln:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \quad (2)$$

$$\chi^2 = \frac{14604(865 \times 6173 - 1129 \times 6437)^2}{7302 \times 1994 \times 12610 \times 7302} \approx 40.48 \text{ (40, 40.5)},$$

mit $0 < \chi^2 \ll n(\min(k, m) - 1) = 14604 \Rightarrow$ der Koeffizient ist klein.

Weiter:

$$K^* = \frac{\sqrt{\frac{\chi^2}{n+\chi^2}}}{\sqrt{(M-1)/M}} \approx 0.074 \in [0, 1] \Rightarrow \text{der Koeffizient ist klein.}$$

Aufgabe 1.3

	Geschmack		Summe
	wie Seife	nicht wie Seife	
Alter (unter 50)	7.55%		55.55%
Alter (über 50)		38.35%	
Summe	13.65%		

Tabelle: Geschmack von Koriander nach Alter (unvollständig)

- (b) Wandeln Sie Tabelle 2 in absoluten Häufigkeiten um und ergänzen Sie die fehlenden Einträge in der Tabelle. Berechnen Sie anschließend das Odds Ratio und interpretieren Sie das Ergebnis.

Aufgabe 1.3

	Geschmack		Summe
	wie Seife	nicht wie Seife	
Alter (unter 50)	1103		8112
Alter (über 50)		5601	
Summe	1994		14604

Tabelle: Geschmack von Koriander nach Alter (unvollständig)

- (b) Wandeln Sie Tabelle 2 in absoluten Häufigkeiten um und ergänzen Sie die fehlenden Einträge in der Tabelle. Berechnen Sie anschließend das Odds Ratio und interpretieren Sie das Ergebnis.

Derselbe Datensatz mit dem gleichen N!

Diskrepanz bei der Abrundung (um 1-2) führt nicht zum Punktabzug.

Aufgabe 1.4

	Geschmack		Summe
	wie Seife	nicht wie Seife	
Alter (unter 50)	1103	7009	8112
Alter (über 50)	891	5601	6492
Summe	1994	12610	14604

Aufgabe 1.4

	Geschmack		Summe
	wie Seife	nicht wie Seife	
Alter (unter 50)	1103	7009	8112
Alter (über 50)	891	5601	6492
Summe	1994	12610	14604

$$\begin{aligned}\gamma(\text{Seife, nicht Seife} | < 50, > 50) &= \frac{h_{11} h_{22}}{h_{21} h_{12}} \\ &= \frac{1103 \times 5601}{891 \times 7009} \\ &\approx 0.99\end{aligned}$$

Die Odds für in beiden Subpopulationen fast gleich, Alter und Koreanderempfinden sind fast empirisch unabhängig.

Aufgabe 1.5

Abstammung	nicht wie Seife	wie Seife
Afroamerikaner	545	55
Aschkenasen	634	104
Ostasiaten	424	39
Europäer	13213	1973
Latinos	820	78
Nordeuropäer	11794	1736
Südasiaten	322	13
Südeuropäer	458	71

Tabelle: Geschmack von Koriander nach Abstammung

- (c) (5 P) Berechnen Sie die bedingte relative Häufigkeit, dass Nordeuropäer einen seifigen Geschmack von Koriander wahrnehmen. Welche Populationen haben eine geringere Chance als Europäer, den Geschmack von Koriander als seifenartig wahrzunehmen? (Odds Ratio)

Aufgabe 1.6

$$f(S|NE) = \frac{1736}{11794+1736} \approx 12.83\%.$$

Aufgabe 1.6

$$f(S|NE) = \frac{1736}{11794+1736} \approx 12.83\%.$$

z.B. $\gamma(Seife, nicht\ Seife \mid Afro, EU) = \frac{55 \times 13213}{1973 \times 545} \approx 0.676$

Aufgabe 1.6

$$f(S|NE) = \frac{1736}{11794+1736} \approx 12.83\%.$$

z.B. $\gamma(Seife, nicht\ Seife \mid Afro, EU) = \frac{55 \times 13213}{1973 \times 545} \approx 0.676$

Afroamerikaner, Latinos, Ostasiaten und Südasiaten haben alle eine signifikant geringere Chance, einen seifigen Geschmack von Koriander zu erkennen, verglichen mit Europäern (Odds Ratios von jeweils 0.676, 0.637, 0.615 und 0.270).

Mögliche Vorgehen:

- ① alle Odds Ratio berechnen und interpretieren.
- ② Schlussfolgerung erstmal aus der Tabelle ziehen (mit sinnvoller Begründung) und dann mit Odds Ratio beweisen.

Aufgabe 2

Eine stetige Zufallsvariable X besitze

$$F(x) = a \left(\frac{x^3 + x^2 - x - 1}{x + 1} - 8 \right) \mathbb{I}_{[3,6]}(x) + \mathbb{I}_{]6,\infty]}(x)$$

als Verteilungsfunktion.

- ① (1 P) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?
- ② (4 P) Bestimmen Sie a so, dass $F(x)$ eine gültige Verteilungsfunktion sein kann und überprüfen Sie, dass $F(x)$ mit diesem Wert eine gültige Verteilungsfunktion ist.
- ③ (3 P) Geben Sie die Dichte $f(x)$ von X an.
- ④ (3 P) Berechnen Sie $P(4 < X < 5)$.

Aufgabe 2.1

- (a) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?

Aufgabe 2.1

- (a) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?

Weil der Träger der ZV offensichtlich nur das Intervall $[3, 6]$ ist und damit sich der Ausdruck so kürzen lässt dass keine Division durch 0 auftritt.

Aufgabe 2.1

- (a) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?

Weil der Träger der ZV offensichtlich nur das Intervall $[3, 6]$ ist und damit sich der Ausdruck so kürzen lässt dass keine Division durch 0 auftritt.

- (b) Bestimmen Sie a so, dass $F(x)$ eine gültige Verteilungsfunktion ist.

Aufgabe 2.1

- (a) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?

Weil der Träger der ZV offensichtlich nur das Intervall $[3, 6]$ ist und damit sich der Ausdruck so kürzen lässt dass keine Division durch 0 auftritt.

- (b) Bestimmen Sie a so, dass $F(x)$ eine gültige Verteilungsfunktion ist.

$$\text{Wegen } (x^3 + x^2 - x - 1) = (x^2 - 1)(x + 1)$$
$$\text{ist } F(x) = a((x^2 - 1) - 8) \mathbb{I}_{[3,6]}(x) + \mathbb{I}_{]6,\infty]}(x)$$

Aufgabe 2.1

- (a) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?

Weil der Träger der ZV offensichtlich nur das Intervall $[3, 6]$ ist und damit sich der Ausdruck so kürzen lässt dass keine Division durch 0 auftritt.

- (b) Bestimmen Sie a so, dass $F(x)$ eine gültige Verteilungsfunktion ist.

$$\text{Wegen } (x^3 + x^2 - x - 1) = (x^2 - 1)(x + 1) \\ \text{ist } F(x) = a((x^2 - 1) - 8) \mathbb{I}_{[3,6]}(x) + \mathbb{I}_{]6,\infty]}(x)$$

Daraus folgt,

$$F(6) = a(6^2 - 9) = 27a \stackrel{!}{=} 1 \Rightarrow a = \frac{1}{27}$$

Aufgabe 2.1

- (a) Warum kann man die Definitionslücke der oben angegebenen Funktion für $x = -1$ hier ignorieren?

Weil der Träger der ZV offensichtlich nur das Intervall $[3, 6]$ ist und damit sich der Ausdruck so kürzen lässt dass keine Division durch 0 auftritt.

- (b) Bestimmen Sie a so, dass $F(x)$ eine gültige Verteilungsfunktion ist.

$$\text{Wegen } (x^3 + x^2 - x - 1) = (x^2 - 1)(x + 1) \\ \text{ist } F(x) = a((x^2 - 1) - 8) \mathbb{I}_{[3,6]}(x) + \mathbb{I}_{]6,\infty]}(x)$$

Daraus folgt,

$$F(6) = a(6^2 - 9) = 27a \stackrel{!}{=} 1 \Rightarrow a = \frac{1}{27}$$

Monotonie (quadratische Funktion auf positivem Träger) und Grenzwerte ($F(x) = 0 \forall x < 3$; $F(x) = 1 \forall x > 6$) liegen als hinreichende Eigenschaften vor. □

Aufgabe 2.2

(c) Geben Sie die Dichte $f(x)$ von X an.

Aufgabe 2.2

(c) Geben Sie die Dichte $f(x)$ von X an.

$$\begin{aligned}f(x) &= F'(x)\mathbb{I}_{[3,6]}(x) \\&= \frac{2}{27}x\mathbb{I}_{[3,6]}(x)\end{aligned}$$

Aufgabe 2.2

(c) Geben Sie die Dichte $f(x)$ von X an.

$$\begin{aligned}f(x) &= F'(x)\mathbb{I}_{[3,6]}(x) \\&= \frac{2}{27}x\mathbb{I}_{[3,6]}(x)\end{aligned}$$

(d) Berechnen Sie $P(4 < X < 5)$.

Aufgabe 2.2

(c) Geben Sie die Dichte $f(x)$ von X an.

$$\begin{aligned}f(x) &= F'(x)\mathbb{I}_{[3,6]}(x) \\&= \frac{2}{27}x\mathbb{I}_{[3,6]}(x)\end{aligned}$$

(d) Berechnen Sie $P(4 < X < 5)$.

$$P(4 < X < 5) \stackrel{\text{stetig}}{=} P(4 \leq X \leq 5) = F(5) - F(4) = 1/3$$

Aufgabe 3

Sei X_a eine Familie von stetigen Zufallsvariablen mit Dichtefunktionen

$$f_{X_a}(x) = c(1 + \sin(x))\mathbb{I}_{[-a,a]}(x)$$

- ① (4 P) Zeigen Sie, dass $f_{X_a}(x)$ für $0 < a < \infty$ mit $c = \frac{1}{2a}$ eine Wahrscheinlichkeitsdichte ist.
- ② (2 P) Skizzieren Sie $f_{X_a}(x)$ (schematisch) für $a = 2\pi$ auf dem Intervall $[-(a + 0.1), a + 0.1]$. Ist $P(X_a \in (-1.5\pi, 0))$ größer, kleiner oder gleich $P(X_a \in (0, 1.5\pi))$?
- ③ (3 P) Bestimmen Sie die Verteilungsfunktion der Zufallsvariable X_a .
- ④ (4 P) Für welche Werte von a ist der Median von X_a exakt Null?
- ⑤ (2 P) Bestimmen Sie $P(X_a = 0)$.

Aufgabe 3.1

- (a) Zeigen Sie, dass $f_{X_a}(x)$ für $0 < a < \infty$ mit $c = \frac{1}{2a}$ eine Wahrscheinlichkeitsdichte ist.

Aufgabe 3.1

(a) Zeigen Sie, dass $f_{X_a}(x)$ für $0 < a < \infty$ mit $c = \frac{1}{2a}$ eine Wahrscheinlichkeitsdichte ist.

$$(1) \forall x, \sin(x) \geq -1 \Rightarrow \forall x, f(x) \geq 0$$

$$(2) \int_{-a}^a f(u) du \stackrel{!}{=} 1$$

$$\Rightarrow c \int_{-a}^a (1 + \sin(x)) du \stackrel{!}{=} 1$$

$$\Rightarrow c[x - \cos(x)]_{-a}^a \stackrel{!}{=} 1$$

$$\Rightarrow \underline{c(a - \cos(a) + a + \cos(-a))} = c \cdot 2a \stackrel{!}{=} 1$$

$$\Rightarrow c = \frac{1}{2a} \text{ (gegeben)}$$

□

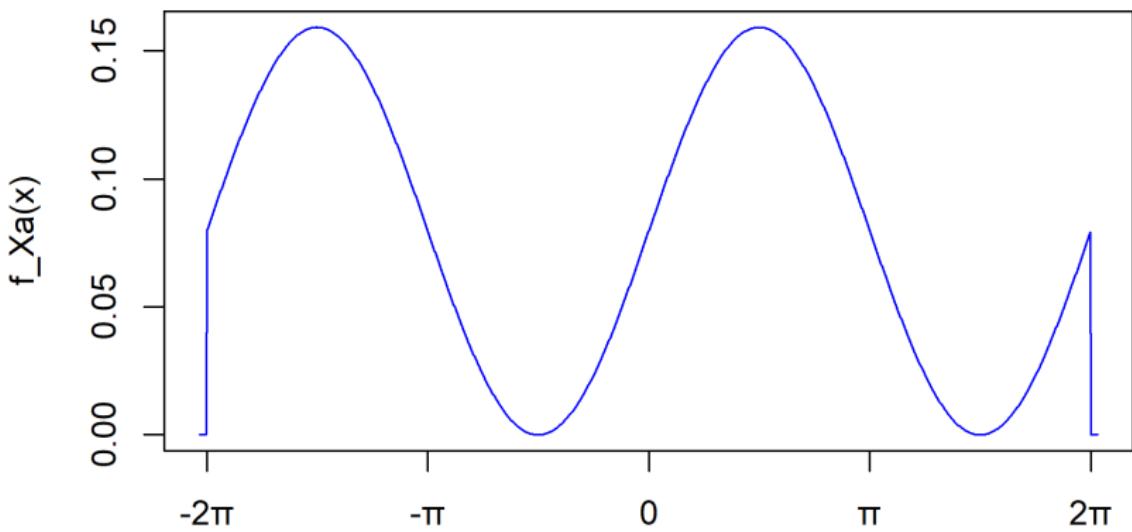
Aufgabe 3.2

- (b) Skizzieren Sie $f_{X_a}(x)$ (schematisch) für $a = 2\pi$ auf dem Intervall $[-(a + 0.1), a + 0.1]$. [...]

Aufgabe 3.2

- (b) Skizzieren Sie $f_{X_a}(x)$ (schematisch) für $a = 2\pi$ auf dem Intervall $[-(a + 0.1), a + 0.1]$. [...]

Plot of $f_{Xa}(x)$ for $a = 2\pi$



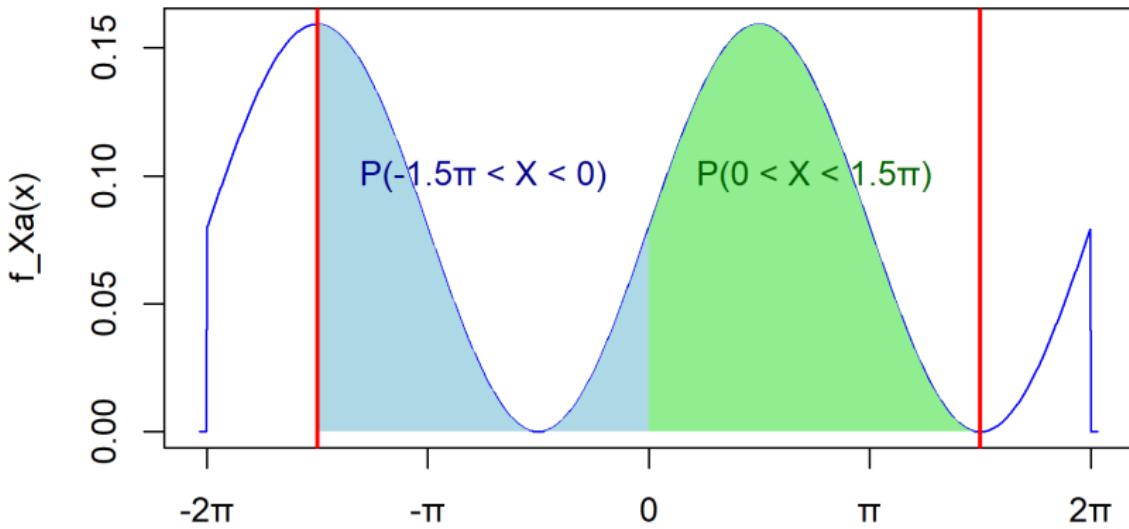
Aufgabe 3.3

- (b) [...] Ist $P(X_a \in (-1.5\pi, 0))$ größer, kleiner oder gleich $P(X_a \in (0, 1.5\pi))$?

Aufgabe 3.3

- (b) [...] Ist $P(X_a \in (-1.5\pi, 0))$ größer, kleiner oder gleich $P(X_a \in (0, 1.5\pi))$?

Plot of $f_{Xa}(x)$ for $a = 2\pi$



Aufgabe 3.4

(c) Bestimmen Sie die Verteilungsfunktion der Zufallsvariable X_a .

Aufgabe 3.4

(c) Bestimmen Sie die Verteilungsfunktion der Zufallsvariable X_a .

$$\begin{aligned}F_{X_a}(x) &= \int_{-a}^x f(u)du \cdot \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x) \\&= \frac{x - \cos(x) + a + \cos(a)}{2a} \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x)\end{aligned}$$

Aufgabe 3.4

(c) Bestimmen Sie die Verteilungsfunktion der Zufallsvariable X_a .

$$\begin{aligned}F_{X_a}(x) &= \int_{-a}^x f(u)du \cdot \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x) \\&= \frac{x - \cos(x) + a + \cos(a)}{2a} \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x)\end{aligned}$$

(d) Für welche Werte von a ist der Median von X_a exakt Null?

Aufgabe 3.4

(c) Bestimmen Sie die Verteilungsfunktion der Zufallsvariable X_a .

$$\begin{aligned}F_{X_a}(x) &= \int_{-a}^x f(u) du \cdot \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x) \\&= \frac{x - \cos(x) + a + \cos(a)}{2a} \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x)\end{aligned}$$

(d) Für welche Werte von a ist der Median von X_a exakt Null?

$$\begin{aligned}F_{X_a}(0) &\stackrel{!}{=} 0.5 \Rightarrow \cos(a) \stackrel{!}{=} 1 \\&\Rightarrow a \in \{2n\pi, n \in \mathbb{N}\}\end{aligned}$$

Aufgabe 3.4

(c) Bestimmen Sie die Verteilungsfunktion der Zufallsvariable X_a .

$$\begin{aligned}F_{X_a}(x) &= \int_{-a}^x f(u) du \cdot \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x) \\&= \frac{x - \cos(x) + a + \cos(a)}{2a} \mathbb{I}_{[-a,a]}(x) + \mathbb{I}_{]a,\infty[}(x)\end{aligned}$$

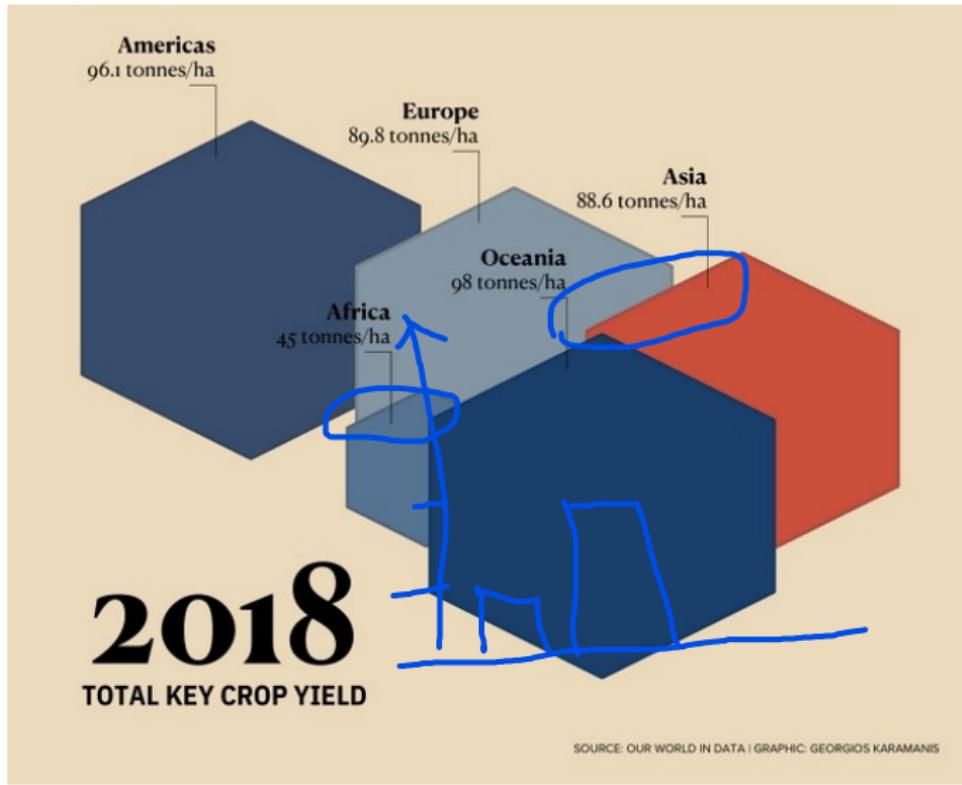
(d) Für welche Werte von a ist der Median von X_a exakt Null?

$$\begin{aligned}F_{X_a}(0) &\stackrel{!}{=} 0.5 \Rightarrow \cos(a) \stackrel{!}{=} 1 \\&\Rightarrow a \in \{2n\pi, n \in \mathbb{N}\}\end{aligned}$$

(e) Bestimmen Sie $P(X_a = 0)$.

$$P(X_a = 0) \stackrel{X \text{ stetig}}{=} 0$$

Aufgabe 4



Aufgabe 4.1

- (a) Geben Sie für alle in der Grafik gezeigten Merkmale das Skalenniveau und die verwendeten Zuordnungen auf ästhetische Eigenschaften der gezeichneten Sechsecke an.

Aufgabe 4.1

- (a) Geben Sie für alle in der Grafik gezeigten Merkmale das Skalenniveau und die verwendeten Zuordnungen auf ästhetische Eigenschaften der gezeichneten Sechsecke an.
- Kontinent: nominalskaliert, Ästhetik: Farbe der Hexagons
 - Ertrag: verhältnisskaliert, Ästhetik: Kantenlänge der Hexagons

Aufgabe 4.1

(a) Geben Sie für alle in der Grafik gezeigten Merkmale das Skalenniveau und die verwendeten Zuordnungen auf ästhetische Eigenschaften der gezeichneten Sechsecke an.

- Kontinent: nominalskaliert, Ästhetik: Farbe der Hexagons
- Ertrag: verhältnisskaliert, Ästhetik: Kantenlänge der Hexagons

(b) Inwiefern verstößt die hier verwendete Farbpalette die in der Vorlesung besprochenen Kriterien für Farbskalen in statistischen Grafiken? Was für eine Art von Farbskala sollte stattdessen verwendet werden?

Aufgabe 4.1

(a) Geben Sie für alle in der Grafik gezeigten Merkmale das Skalenniveau und die verwendeten Zuordnungen auf ästhetische Eigenschaften der gezeichneten Sechsecke an.

- Kontinent: nominalskaliert, Ästhetik: Farbe der Hexagons
- Ertrag: verhältnisskaliert, Ästhetik: Kantenlänge der Hexagons

(b) Inwiefern verletzt die hier verwendete Farbpalette die in der Vorlesung besprochenen Kriterien für Farbskalen in statistischen Grafiken? Was für eine Art von Farbskala sollte stattdessen verwendet werden?

- Problem: kaum unterscheidbare Blautöne und ein hervorstechender Rotton.
Die Farbskala ist damit weder divergierend noch sequentiell noch qualitativ.
- Besser: eine wahrnehmungseinheitliche, gut unterscheidbare, qualitative (sequentielle) Farbskala.

Aufgabe 4.2

- (c) Statistische Grafiken sollen die Datenlage möglichst **unverfälscht** darstellen.
- (d) Statistische Grafiken sollen die Datenlage möglichst **kompakt** darstellen, also: minimal viel verwendete Tinte für maximal viel vermittelte Information.
- (e) Statistische Grafiken sollen die Datenlage möglichst **übersichtlich** darstellen, um den Konsument:innen der Grafik schnelles und präzises Ablesen relevanter quantitativer Informationen zu ermöglichen.

Inwiefern verfehlt die obige Darstellung diese Ziele?

Aufgabe 4.3

- **Unverfälschtheit:** Der visuelle Eindruck entsteht über die Flächen der Hexagone, doch der Ertrag wird durch die Kantenlänge codiert. Die Transformation von 1D zu 2D kann die Unterschiede in den Daten über- oder unterbetonen, was eine verzerrte bzw. verfälschte Wahrnehmung zur Folge haben kann.

Aufgabe 4.3

- **Unverfälschtheit:** Der visuelle Eindruck entsteht über die Flächen der Hexagone, doch der Ertrag wird durch die Kantenlänge codiert. Die Transformation von 1D zu 2D kann die Unterschiede in den Daten über- oder unterbetonen, was eine verzerrte bzw. verfälschte Wahrnehmung zur Folge haben kann.
- **Kompaktheit:** Das metrische Merkmal wird durch die Flächen von eingefärbten Hexagonen repräsentiert. Die Grafik nutzt zudem eine 3D-Anordnung mit Überlappungen für die Hexagone. Weder die Färbung noch die räumliche Anordnung oder die Darstellung durch Flächen sind notwendig. Eine 2D-Darstellung mit 5 Säulen, die die Kontinente auf der x-Achse kategorisieren, wäre ausreichend.

Aufgabe 4.3

- **Unverfälschtheit:** Der visuelle Eindruck entsteht über die Flächen der Hexagone, doch der Ertrag wird durch die Kantenlänge codiert. Die Transformation von 1D zu 2D kann die Unterschiede in den Daten über- oder unterbetonen, was eine verzerrte bzw. verfälschte Wahrnehmung zur Folge haben kann.
- **Kompaktheit:** Das metrische Merkmal wird durch die Flächen von eingefärbten Hexagonen repräsentiert. Die Grafik nutzt zudem eine 3D-Anordnung mit Überlappungen für die Hexagone. Weder die Färbung noch die räumliche Anordnung oder die Darstellung durch Flächen sind notwendig. Eine 2D-Darstellung mit 5 Säulen, die die Kontinente auf der x-Achse kategorisieren, wäre ausreichend.
- **Übersichtlichkeit:** Der genaue Wert vom Ertrag ist zwischen den Kontinenten aufgrund der willkürlich angeordneten Polygone mit teilweiser Überlappung und fehlender Reihung schwer bis unmöglich zu vergleichen.

Zusammenfassung: Was hat die Grafik falsch gemacht?

- ① Arbiträre 3D-Anordnung der Geometrien
- ② Verfälschung durch Flächen
- ③ Überflüssige Farbskala
- ④ Fehlende Skalierung des metrischen Merkmals
- ⑤ Überlappung der Polygonen
- ⑥ Komplexität ohne Mehrwert

Zusammenfassung: Was hat die Grafik falsch gemacht?

- ① Arbiträre 3D-Anordnung der Geometrien
 - ② Verfälschung durch Flächen
 - ③ Überflüssige Farbskala
 - ④ Fehlende Skalierung des metrischen Merkmals
 - ⑤ Überlappung der Polygonen
 - ⑥ Komplexität ohne Mehrwert
- (f) Definieren Sie eine alternative grafische Darstellung für die in der obenstehenden Grafik gezeigten Daten, welche diese unverfälscht, kompakt und übersichtlich visualisiert.

Zusammenfassung: Was hat die Grafik falsch gemacht?

- ① Arbiträre 3D-Anordnung der Geometrien
 - ② Verfälschung durch Flächen
 - ③ Überflüssige Farbskala
 - ④ Fehlende Skalierung des metrischen Merkmals
 - ⑤ Überlappung der Polygonen
 - ⑥ Komplexität ohne Mehrwert
- (f) Definieren Sie eine alternative grafische Darstellung für die in der obenstehenden Grafik gezeigten Daten, welche diese unverfälscht, kompakt und übersichtlich visualisiert.

“Eine 2D-Darstellung mit 5 Säulen, die die Kontinente auf der x-Achse kategorisieren.”

Geometrien: Säulen bzw. Balken

Ästhetische Zuordnungen: x-Achse (Kontinent), y-Achse (Ertrag)

Aufgabe 4.5 (baseR vs. ggplot2)

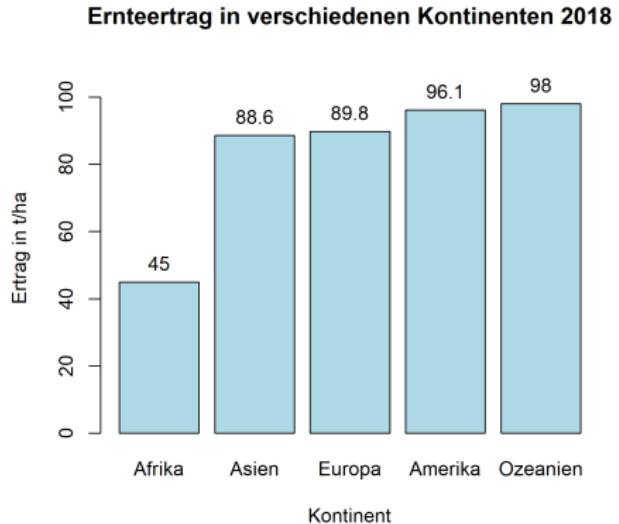


Abbildung: baseR

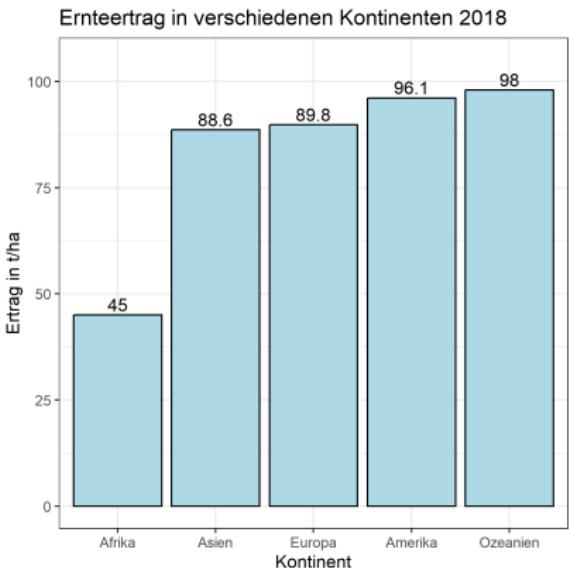


Abbildung: ggplot2

8. Lagemaße und Streuungsmaße

Yichen Han



14. Dezember 2023

1 Wiederholung

2 Blatt 8



Abbildung: Quiz 9

Stichwörter

1. Modus, Median, Quantil, Mittelwerte, Erwartungswert
2. Spannweite, IQR, Standardabweichung, Varianz, MAD, MedAD, Schiefe, Kurtosis
3. Konzentrationsmaße: Lorenz-Kurve, Gini-Koeffizient, Herfindahl-Index
4. Verschiebungssatz, Tschebyscheffsche Ungleichung, Boxplot

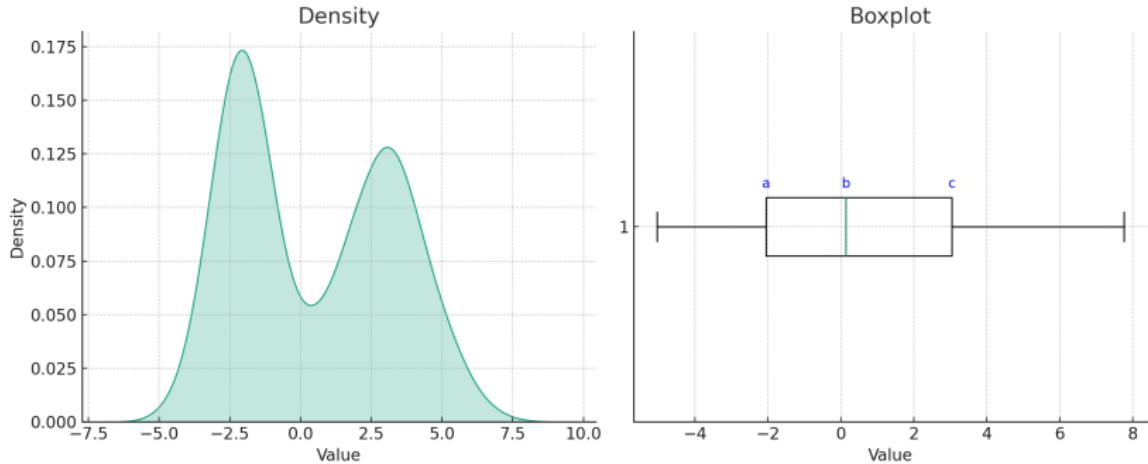


Abbildung: Beispiel 1: eine bimodale Verteilung

Beispiel 2: siehe R-Simulation.

1 Wiederholung

2 Blatt 8

Aufgabe 1.1

Jahr	2012	2013	2014	2015	2016
Umsatz (in Euro)	510000	530700	590200	640800	?

Tabelle: Umsatzentwicklung über die Jahre

- (a) Angenommen, der Umsatz im Jahr 2016 konnte um 20 Prozent im Vergleich zum Vorjahr gesteigert werden. Ermitteln Sie den durchschnittlichen Umsatz für die Jahre 2012 bis 2016.

Aufgabe 1.1

Jahr	2012	2013	2014	2015	2016
Umsatz (in Euro)	510000	530700	590200	640800	?

Tabelle: Umsatzentwicklung über die Jahre

- (a) Angenommen, der Umsatz im Jahr 2016 konnte um 20 Prozent im Vergleich zum Vorjahr gesteigert werden. Ermitteln Sie den durchschnittlichen Umsatz für die Jahre 2012 bis 2016.

$$U_{2016} = (1 + 0.2) \times U_{2015} = 1.2 \times 640800 = 768960,$$

$$\bar{U} = \frac{1}{5} \sum_{i=1}^5 U_i = 608132.$$

Aufgabe 1.2

Jahr	2012	2013	2014	2015	2016
Umsatz (in Euro)	510000	530700	590200	640800	768960

Tabelle: Umsatzentwicklung über die Jahre

- (b) Berechnen Sie das mittlere jährliche Umsatzwachstum in Prozent.

Aufgabe 1.2

Jahr	2012	2013	2014	2015	2016
Umsatz (in Euro)	510000	530700	590200	640800	768960

Tabelle: Umsatzentwicklung über die Jahre

- (b) Berechnen Sie das mittlere jährliche Umsatzwachstum in Prozent.

Der t -te Wachstumsfaktor ist definiert als: $i_t = \frac{x_t}{x_{t-1}}$ für $t = 1, \dots, T$.

Aufgabe 1.2

Jahr	2012	2013	2014	2015	2016
Umsatz (in Euro)	510000	530700	590200	640800	768960

Tabelle: Umsatzentwicklung über die Jahre

- (b) Berechnen Sie das mittlere jährliche Umsatzwachstum in Prozent.

Der t -te Wachstumsfaktor ist definiert als: $i_t = \frac{x_t}{x_{t-1}}$ für $t = 1, \dots, T$.

Das passende Mittel: Geometrisches Mittel! (Wachstumsfaktor ist ein multiplikativer Faktor).

Aufgabe 1.2

Jahr	2012	2013	2014	2015	2016
Umsatz (in Euro)	510000	530700	590200	640800	768960

Tabelle: Umsatzentwicklung über die Jahre

- (b) Berechnen Sie das mittlere jährliche Umsatzwachstum in Prozent.

Der t -te Wachstumsfaktor ist definiert als: $i_t = \frac{x_t}{x_{t-1}}$ für $t = 1, \dots, T$.

Das passende Mittel: Geometrisches Mittel! (Wachstumsfaktor ist ein multiplikativer Faktor).

$$\begin{aligned}\bar{x}_G &= \sqrt[T]{i_1 \cdot \dots \cdot i_T} = \sqrt[T]{\frac{x_1}{x_0} \cdot \frac{x_2}{x_1} \cdot \dots \cdot \frac{x_T}{x_{T-1}}} = \sqrt[T]{\frac{x_T}{x_0}} \\ &= \sqrt[4]{\frac{x_4}{x_0}} = \sqrt[4]{\frac{768960}{510000}} \approx 1.1081.\end{aligned}$$

Aufgabe 2.1

Feinstaub: 20, 27, 22, 20, 22, 20, 19, 20, 17, 27, 23, 27, 22, 27, 25.

- (a) Geben Sie das arithmetische Mittel sowie den Median und die beiden Quartile an.

Aufgabe 2.1

Feinstaub: 20, 27, 22, 20, 22, 20, 19, 20, 17, 27, 23, 27, 22, 27, 25.

- (a) Geben Sie das arithmetische Mittel sowie den Median und die beiden Quartile an.

Stichprobenumfang $n = 15$,

Geordnete Werte: 17, 19, 20, 20, 20, 20, 22, 22, 22, 23, 25, 27, 27, 27, 27.

$$\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = \frac{1}{15} \cdot 338 \approx 22.53$$

Aufgabe 2.1

Feinstaub: 20, 27, 22, 20, 22, 20, 19, 20, 17, 27, 23, 27, 22, 27, 25.

- (a) Geben Sie das arithmetische Mittel sowie den Median und die beiden Quartile an.

Stichprobenumfang $n = 15$,

Geordnete Werte: 17, 19, 20, 20, 20, 20, 22, 22, 22, 23, 25, 27, 27, 27, 27.

$$\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = \frac{1}{15} \cdot 338 \approx 22.53$$

p -Quantile:

$$\tilde{x}_p = \begin{cases} x_{(k)} & np \notin \mathbb{Z}, k \text{ kleinste ganze Zahl } > np \\ \in [x_{(k)}; x_{(k+1)}] & k = np \in \mathbb{Z} \end{cases}$$

$$\tilde{x}_{0.25} = x_{(4)} = 20, \quad \tilde{x}_{med} = x_{(8)} = 22, \quad \tilde{x}_{0.75} = x_{(12)} = 27$$

Aufgabe 3.1

Sei Z eine Zufallsvariable, die das Minimum von drei Würfen eines sechsseitigen Würfels beschreibt.

$$F(z) = \begin{cases} 0 & \text{für } z < 1 \\ \frac{91}{216} & \text{für } 1 \leq z < 2 \\ \frac{152}{216} & \text{für } 2 \leq z < 3 \\ \frac{189}{216} & \text{für } 3 \leq z < 4 \\ \frac{208}{216} & \text{für } 4 \leq z < 5 \\ \frac{215}{216} & \text{für } 5 \leq z < 6 \\ 1 & \text{für } z \geq 6. \end{cases}$$

Berechnen Sie $E(Z)$.

Aufgabe 3.2

$$f(z) = \begin{cases} 0 & \text{für } z \notin \{1, 2, \dots, 6\} \\ \frac{91}{216} & \text{für } z = 1 \\ \frac{152-91}{216} = \frac{61}{216} & \text{für } z = 2 \\ \frac{189-152}{216} = \frac{37}{216} & \text{für } 3 \leq z < 4 \\ \frac{208-189}{216} = \frac{19}{216} & \text{für } 4 \leq z < 5 \\ \frac{215-208}{216} = \frac{7}{216} & \text{für } 5 \leq z < 6 \\ \frac{216-215}{216} = \frac{1}{216} & \text{für } z = 6. \end{cases}$$

Aufgabe 3.2

$$f(z) = \begin{cases} 0 & \text{für } z \notin \{1, 2, \dots, 6\} \\ \frac{91}{216} & \text{für } z = 1 \\ \frac{152-91}{216} = \frac{61}{216} & \text{für } z = 2 \\ \frac{189-152}{216} = \frac{37}{216} & \text{für } 3 \leq z < 4 \\ \frac{208-189}{216} = \frac{19}{216} & \text{für } 4 \leq z < 5 \\ \frac{215-208}{216} = \frac{7}{216} & \text{für } 5 \leq z < 6 \\ \frac{216-215}{216} = \frac{1}{216} & \text{für } z = 6. \end{cases}$$

$$E(Z) = \sum_{z \in T_Z} zf(z) = 1 \cdot \frac{91}{216} + 2 \cdot \frac{61}{216} + \dots + 6 \cdot \frac{1}{216} \approx 2.042$$

Aufgabe 4.1

Stunden	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	Σ
1960	5	3	10	9	13	18	21	27	12	3	$n_1 = 121$
1980	6	7	5	20	29	27	13	5	3	2	$n_2 = 117$
2000	35	24	13	8	9	4	2	1	0	1	$n_3 = 97$

Gruppierte Lagemaße

- **Modus:** Bestimme Modalklasse (Klasse mit der größten Beobachtungszahl) und verwende Klassenmitte (m_i) als Modus.
- **Median:** Bestimme Einfallsklasse $[c_{i-1}, c_i)$ des Medians und daraus

$$\tilde{x}_{\text{med,grupp}} = c_{i-1} + \frac{d_i \cdot (0.5 - F(c_{i-1}))}{f_i}.$$

- **Arithmetisches Mittel:**

$$\bar{x}_{\text{grupp}} = \sum_{i=1}^k f_i m_i.$$

Aufgabe 4.2

Stunden	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	Σ
1960	5	3	10	9	13	18	21	27	12	3	$n_1 = 121$
1980	6	7	5	20	29	27	13	5	3	2	$n_2 = 117$
2000	35	24	13	8	9	4	2	1	0	1	$n_3 = 97$

- (a) Bestimmen Sie aus den klassierten Daten jeweils Modus und Median für die drei Jahre.

Aufgabe 4.2

Stunden	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	Σ
1960	5	3	10	9	13	18	21	27	12	3	$n_1 = 121$
1980	6	7	5	20	29	27	13	5	3	2	$n_2 = 117$
2000	35	24	13	8	9	4	2	1	0	1	$n_3 = 97$

- (a) Bestimmen Sie aus den klassierten Daten jeweils Modus und Median für die drei Jahre.

Aufgabe 4.2

Stunden	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	Σ
1960	5	3	10	9	13	18	21	27	12	3	$n_1 = 121$
1980	6	7	5	20	29	27	13	5	3	2	$n_2 = 117$
2000	35	24	13	8	9	4	2	1	0	1	$n_3 = 97$

- (a) Bestimmen Sie aus den klassierten Daten jeweils Modus und Median für die drei Jahre.

Jahr	Modalklasse	Modus (Klassenmitte)
1960	[7, 8)	7.5
1980	[4, 5)	4.5
2000	[0, 1)	0.5

Aufgabe 4.2

Stunden	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	Σ
1960	5	3	10	9	13	18	21	27	12	3	$n_1 = 121$
1980	6	7	5	20	29	27	13	5	3	2	$n_2 = 117$
2000	35	24	13	8	9	4	2	1	0	1	$n_3 = 97$

- (a) Bestimmen Sie aus den klassierten Daten jeweils Modus und Median für die drei Jahre.

Jahr	Modalklasse	Modus (Klassenmitte)
1960	[7, 8)	7.5
1980	[4, 5)	4.5
2000	[0, 1)	0.5

Median:

Berechnung der relativen und kumulierten relativen Häufigkeiten für die einzelnen Klassen: (nächste Seite)

Aufgabe 4.3

Klasse j	1	2	3	4	5	6	7	8	9	10
Stunden	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)
f_m^{1960}	0.04	0.02	0.08	0.07	0.11	0.15	0.17	0.22	0.10	0.02
F_m^{1960}	0.04	0.07	0.15	0.22	0.33	0.48	0.65	0.88	0.98	1.00
f_m^{1980}	0.05	0.06	0.04	0.17	0.25	0.23	0.11	0.04	0.03	0.02
F_m^{1980}	0.05	0.11	0.15	0.32	0.57	0.80	0.91	0.96	0.98	1.00
f_m^{2000}	0.36	0.25	0.13	0.08	0.09	0.04	0.02	0.01	0.00	0.01
F_m^{2000}	0.36	0.61	0.74	0.82	0.92	0.96	0.98	0.99	0.99	1.00

Aufgabe 4.3

Klasse j Stunden	1 [0,1)	2 [1,2)	3 [2,3)	4 [3,4)	5 [4,5)	6 [5,6)	7 [6,7)	8 [7,8)	9 [8,9)	10 [9,10)
f_m^{1960}	0.04	0.02	0.08	0.07	0.11	0.15	0.17	0.22	0.10	0.02
F_m^{1960}	0.04	0.07	0.15	0.22	0.33	0.48	0.65	0.88	0.98	1.00
f_m^{1980}	0.05	0.06	0.04	0.17	0.25	0.23	0.11	0.04	0.03	0.02
F_m^{1980}	0.05	0.11	0.15	0.32	0.57	0.80	0.91	0.96	0.98	1.00
f_m^{2000}	0.36	0.25	0.13	0.08	0.09	0.04	0.02	0.01	0.00	0.01
F_m^{2000}	0.36	0.61	0.74	0.82	0.92	0.96	0.98	0.99	0.99	1.00

$$1960: \tilde{x}_{med,1} = 6 + \frac{0.5 - 0.48}{0.17} \approx 6.12$$

$$1980: \tilde{x}_{med,2} = 4 + \frac{0.5 - 0.32}{0.25} \approx 4.72$$

$$2000: \tilde{x}_{med,3} = 1 + \frac{0.5 - 0.36}{0.25} \approx 1.56$$

Aufgabe 4.4

- (b) Wie drücken sich die im Laufe der Zeit veränderten Hörgewohnheiten durch die drei unter (a) berechneten Lagemaße aus?

Aufgabe 4.4

(b) Wie drücken sich die im Laufe der Zeit veränderten Hörgewohnheiten durch die drei unter (a) berechneten Lagemaße aus?

Jahr	Modus (Klassenmitte)	Median
1960	7.5	6.12
1980	4.5	4.72
2000	0.5	1.56

Aufgabe 4.4

- (b) Wie drücken sich die im Laufe der Zeit veränderten Hörgewohnheiten durch die drei unter (a) berechneten Lagemaße aus?

Jahr	Modus (Klassenmitte)	Median
1960	7.5	6.12
1980	4.5	4.72
2000	0.5	1.56

Modus und Median wurden über die Zeit kleiner, dies drückt einen rückwärtigen Radiokonsum im Zeitverlauf aus. Dieser Effekt fällt beim Modus stärker aus.

Aufgabe 4.5

- (c) Wie können Sie aus diesen klassierten Daten arithmetische Mittelwerte für jedes Jahr berechnen? Welche zusätzlichen Annahmen müssen Sie dafür treffen?

Arithmetisches Mittel (Verwendung der Klassenmittelpunkte):

$$1960: \bar{x}_1 = \frac{1}{121} (0.5 \cdot 5 + \dots + 9.5 \cdot 3) \approx 5.71$$

$$1980: \bar{x}_2 = \frac{1}{117} (0.5 \cdot 6 + \dots + 9.5 \cdot 2) \approx 4.63$$

$$2000: \bar{x}_3 = \frac{1}{97} (0.5 \cdot 35 + \dots + 9.5 \cdot 1) \approx 2.13$$

Aufgabe 4.6

Die Formeln aus Fahrmeir sind im Prinzip *lineare Interpolation* der Daten.

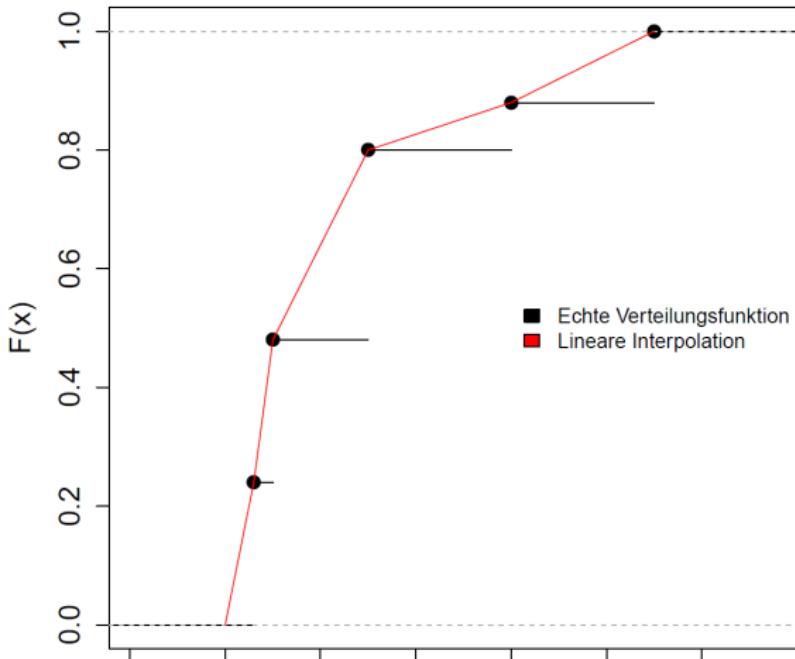


Abbildung: Beispiel: Lineare Interpolation einer empirischen Verteilung

Aufgabe 4.7

$$\tilde{x}_{\text{med,grupp}} = \overbrace{c_{i-1}}^{\text{Basis}} + \underbrace{\frac{d_i \cdot (0.5 - F(c_{i-1}))}{f_i}}_{\text{Modifikation unter Gleichverteilung}}$$

Aufgabe 4.7

$$\tilde{x}_{\text{med,grupp}} = \overbrace{c_{i-1}}^{\text{Basis}} + \underbrace{\frac{d_i \cdot (0.5 - F(c_{i-1}))}{f_i}}_{\text{Modifikation unter Gleichverteilung}}$$

Annahmen der gruppierten Lagemaße

- gleiche Klassenbreite (optional)
- Die Merkmalswerte sind innerhalb aller Klassen gleichverteilt. (zentral)

9. Maße (weiterführend)

Yichen Han



21. Dezember 2023

1 Wiederholung

2 Blatt 9



Abbildung: Quiz 10

Stichwörter

1. Modus, Median, Quantil, Mittelwerte, Erwartungswert
2. Spannweite, IQR, Standardabweichung, Varianz, MAD, MedAD, Schiefe, Kurtosis, Variationskoeffizient
3. Konzentrationsmaße: Lorenz-Kurve, Gini-Koeffizient, Herfindahl-Index
4. Verschiebungssatz, Tschebyscheffsche Ungleichung, Boxplot

1 Wiederholung

2 Blatt 9

Aufgabe 1.1

stündliche Mittelwerte für die Temperatur in $^{\circ}\text{C}$:

7.2	7.2	6.8	7.2	7.4	7.1	7.0	7.2	7.5	7.7	7.9	8.3
7.9	7.6	7.2	7.1	6.9	6.9	6.7	6.2	6.4	6.2	6.2	6.2

- (a) Modus, arithmetische Mittel, Median, IQR, Varianz. Charakterisieren.

Aufgabe 1.1

stündliche Mittelwerte für die Temperatur in °C:

7.2	7.2	6.8	7.2	7.4	7.1	7.0	7.2	7.5	7.7	7.9	8.3
7.9	7.6	7.2	7.1	6.9	6.9	6.7	6.2	6.4	6.2	6.2	6.2

- (a) Modus, arithmetische Mittel, Median, IQR, Varianz. Charakterisieren.

Beobachtungen: $n = 24$

Geordnete Urliste:

6.2	6.2	6.2	6.2	6.4	6.7	6.8	6.9	6.9	7.0	7.1	7.1
7.2	7.2	7.2	7.2	7.2	7.4	7.5	7.6	7.7	7.9	7.9	8.3

Aufgabe 1.1

stündliche Mittelwerte für die Temperatur in °C:

7.2	7.2	6.8	7.2	7.4	7.1	7.0	7.2	7.5	7.7	7.9	8.3
7.9	7.6	7.2	7.1	6.9	6.9	6.7	6.2	6.4	6.2	6.2	6.2

(a) Modus, arithmetische Mittel, Median, IQR, Varianz. Charakterisieren.

Beobachtungen: $n = 24$

Geordnete Urliste:

6.2	6.2	6.2	6.2	6.4	6.7	6.8	6.9	6.9	7.0	7.1	7.1
7.2	7.2	7.2	7.2	7.2	7.4	7.5	7.6	7.7	7.9	7.9	8.3

Modus: $x_{mod} = 7.2$ °C

Aufgabe 1.1

stündliche Mittelwerte für die Temperatur in $^{\circ}\text{C}$:

7.2	7.2	6.8	7.2	7.4	7.1	7.0	7.2	7.5	7.7	7.9	8.3
7.9	7.6	7.2	7.1	6.9	6.9	6.7	6.2	6.4	6.2	6.2	6.2

(a) Modus, arithmetische Mittel, Median, IQR, Varianz. Charakterisieren.

Beobachtungen: $n = 24$

Geordnete Urliste:

6.2	6.2	6.2	6.2	6.4	6.7	6.8	6.9	6.9	7.0	7.1	7.1
7.2	7.2	7.2	7.2	7.2	7.4	7.5	7.6	7.7	7.9	7.9	8.3

Modus: $x_{mod} = 7.2 \text{ } ^{\circ}\text{C}$

Arithmetisches Mittel: $\bar{x} = \frac{1}{24} \sum_{i=1}^{24} x_i = \frac{1}{24} \cdot 170 \approx 7.08 \text{ } ^{\circ}\text{C}$

Aufgabe 1.1

stündliche Mittelwerte für die Temperatur in $^{\circ}\text{C}$:

7.2	7.2	6.8	7.2	7.4	7.1	7.0	7.2	7.5	7.7	7.9	8.3
7.9	7.6	7.2	7.1	6.9	6.9	6.7	6.2	6.4	6.2	6.2	6.2

(a) Modus, arithmetische Mittel, Median, IQR, Varianz. Charakterisieren.

Beobachtungen: $n = 24$

Geordnete Urliste:

6.2	6.2	6.2	6.2	6.4	6.7	6.8	6.9	6.9	7.0	7.1	7.1
7.2	7.2	7.2	7.2	7.2	7.4	7.5	7.6	7.7	7.9	7.9	8.3

Modus: $x_{mod} = 7.2 \text{ } ^{\circ}\text{C}$

Arithmetisches Mittel: $\bar{x} = \frac{1}{24} \sum_{i=1}^{24} x_i = \frac{1}{24} \cdot 170 \approx 7.08 \text{ } ^{\circ}\text{C}$

Median: $\tilde{x}_{med} = \frac{1}{2}[x_{(12)} + x_{(13)}] = \frac{1}{2}[7.1 + 7.2] = 7.15 \text{ } ^{\circ}\text{C}$

IQR: $d_{Q,x} = \tilde{x}_{0.75} - \tilde{x}_{0.25} = \frac{1}{2}(x_{(18)} + x_{(19)}) - \frac{1}{2}(x_{(6)} + x_{(7)}) = 7.45 - 6.75 = 0.7$

Aufgabe 1.2

Varianz:

$$\tilde{S}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1211.86}{24} - 7.08^2 \approx 0.32 \text{ } {}^\circ\text{C}^2$$

Aufgabe 1.2

Varianz:

$$\tilde{s}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1211.86}{24} - 7.08^2 \approx 0.32 \text{ } ^\circ C^2$$

Charakterisierung der Verteilung:

Beobachtung: Modus \approx Median \approx Mittelwert

Aufgabe 1.2

Varianz:

$$\tilde{S}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1211.86}{24} - 7.08^2 \approx 0.32 \text{ } ^\circ C^2$$

Charakterisierung der Verteilung:

Beobachtung: Modus \approx Median \approx Mittelwert

Da arithmetisches Mittel, Median und Modus sehr ähnlich sind, kann von einer **symmetrischen** Verteilung ausgegangen werden.

Aufgabe 1.3

- Y : Stündliches Temperaturmittel in München am 27.11.2012 (in °F)
- Modus:

$$y_{mod} = a + bx_{mod} = 32 + \frac{9}{5} \cdot 7.2 = 44.96 \quad [\text{°F}]$$

- Arithmetisches Mittel:

$$\bar{y} = a + b\bar{x} = 32 + \frac{9}{5} \cdot 7.08 \approx 44.74 \quad [\text{°F}]$$

- Median:

$$\tilde{y}_{med} = a + b\tilde{x}_{med} = 32 + \frac{9}{5} \cdot 7.15 = 44.87 \quad [\text{°F}]$$

- Unteres Quartil:

$$\tilde{y}_{0.25} = a + b\tilde{x}_{0.25} = 32 + \frac{9}{5} \cdot 6.75 = 44.15 \quad [\text{°F}]$$

- Oberes Quartil:

$$\tilde{y}_{0.75} = a + b\tilde{x}_{0.75} = 32 + \frac{9}{5} \cdot 7.45 = 45.41 \quad [\text{°F}]$$

- Interquartilsabstand:

$$d_{Q,Y} = \tilde{y}_{0.75} - \tilde{y}_{0.25} = a + b\tilde{x}_{0.75} - (a + b\tilde{x}_{0.25}) = b(\tilde{x}_{0.75} - \tilde{x}_{0.25}) = bd_{Q,X} = \frac{9}{5} \cdot 0.7 = 1.26 \quad [\text{°F}]$$

- Varianz

$$\tilde{S}_Y^2 = b^2 \tilde{S}_X^2 = \left(\frac{9}{5}\right)^2 \cdot 0.32 \approx 1.04 \quad [\text{°F}^2]$$

Aufgabe 1.4

(c) Geben Sie den MAD und den MedAD für die Temperatur in °C an.

Aufgabe 1.4

(c) Geben Sie den MAD und den MedAD für die Temperatur in °C an.

- MAD:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{24} \sum_{i=1}^{24} |x_i - 7.08| \approx 0.44$$

Aufgabe 1.4

(c) Geben Sie den MAD und den MedAD für die Temperatur in °C an.

- MAD:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{24} \sum_{i=1}^{24} |x_i - 7.08| \approx 0.44$$

- MedAD:

$$MedAD = median(|x_i - x_{med}|) = median(|x_i - 7.15|)$$

Aufgabe 1.4

(c) Geben Sie den MAD und den MedAD für die Temperatur in °C an.

- MAD:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{24} \sum_{i=1}^{24} |x_i - 7.08| \approx 0.44$$

- MedAD:

$$MedAD = median(|x_i - x_{med}|) = median(|x_i - 7.15|)$$

Geordnete Urliste der $|x_i - 7.15|$:

0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.15	0.25	0.25	0.25	0.35
0.35	0.45	0.45	0.55	0.75	0.75	0.75	0.95	0.95	0.95	0.95	1.15	

Aufgabe 1.4

(c) Geben Sie den MAD und den MedAD für die Temperatur in °C an.

- MAD:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{24} \sum_{i=1}^{24} |x_i - 7.08| \approx 0.44$$

- MedAD:

$$MedAD = median(|x_i - x_{med}|) = median(|x_i - 7.15|)$$

Geordnete Urliste der $|x_i - 7.15|$:

0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.15	0.25	0.25	0.25	0.35
0.35	0.45	0.45	0.55	0.75	0.75	0.75	0.95	0.95	0.95	0.95	1.15

⇒

$$MedAD = \frac{1}{2}(0.35 + 0.35) = 0.35$$

Aufgabe 1.5

- (d) Der Variationskoeffizient gilt als skalierungsunabhängiges Streuungsmaß. Kann daraus abgeleitet werden, dass die Variationskoeffizienten für die Temperatur in °F und in °C gleich sind? Begründen Sie.

Aufgabe 1.5

- (d) Der Variationskoeffizient gilt als skalierungsunabhängiges Streuungsmaß. Kann daraus abgeleitet werden, dass die Variationskoeffizienten für die Temperatur in °F und in °C gleich sind? Begründen Sie.

$$y[\text{°F}] = \frac{9}{5}x[\text{°C}] + 32$$

Aufgabe 1.5

- (d) Der Variationskoeffizient gilt als skalierungsunabhängiges Streuungsmaß. Kann daraus abgeleitet werden, dass die Variationskoeffizienten für die Temperatur in °F und in °C gleich sind? Begründen Sie.

$$y[\text{°F}] = \frac{9}{5}x[\text{°C}] + 32$$

⇒ Nein! Die Streuung der beiden Verteilungen kann ohne vorheriges Umrechnen nicht mit Hilfe des Variationskoeffizienten verglichen werden. $v_X = v_Y$ gilt nur, wenn $a = 0$ gilt.

Aufgabe 2.1

Folien 385, 388-390

Schüler	1	2	3	4	5
Taschengeld	50	80	20	65	40

- (a) Berechnen und interpretieren Sie ein auf den Bereich $[0, 1]$ normiertes Maß für die Konzentration des Taschengeldes. Stellen Sie die Situation graphisch dar.

Aufgabe 2.1

Folien 385, 388-390

Schüler	1	2	3	4	5
Taschengeld	50	80	20	65	40

- (a) Berechnen und interpretieren Sie ein auf den Bereich $[0, 1]$ normiertes Maß für die Konzentration des Taschengeldes. Stellen Sie die Situation graphisch dar.

$$\text{normierte Gini-Koeffizient } G^+ = \frac{n}{n-1} G \in [0, 1].$$

Aufgabe 2.2

i	$x(i)$	$u_i = \frac{i}{n}$	$v_i = \frac{\sum_{j=1}^i x(j)}{\sum_{j=1}^n x(j)}$
1	20	$\frac{1}{5} = 0.2$	$\frac{20}{255} \approx 0.078$
2	40	$\frac{2}{5} = 0.4$	$\frac{60}{255} \approx 0.235$
3	50	$\frac{3}{5} = 0.6$	$\frac{110}{255} \approx 0.431$
4	65	$\frac{4}{5} = 0.8$	$\frac{175}{255} \approx 0.686$
5	80	$\frac{5}{5} = 1$	$\frac{255}{255} = 1$

Aufgabe 2.2

i	$x(i)$	$u_i = \frac{i}{n}$	$v_i = \frac{\sum_{j=1}^i x(j)}{\sum_{j=1}^n x(j)}$
1	20	$\frac{1}{5} = 0.2$	$\frac{20}{255} \approx 0.078$
2	40	$\frac{2}{5} = 0.4$	$\frac{60}{255} \approx 0.235$
3	50	$\frac{3}{5} = 0.6$	$\frac{110}{255} \approx 0.431$
4	65	$\frac{4}{5} = 0.8$	$\frac{175}{255} \approx 0.686$
5	80	$\frac{5}{5} = 1$	$\frac{255}{255} = 1$

$$G = 1 - \frac{1}{n} \sum_{j=1}^n (v_{j-1} + v_j) = 1 - \frac{1}{5} ((0 + 0.078) + \dots) \approx 0.227$$

Aufgabe 2.2

i	$x(i)$	$u_i = \frac{i}{n}$	$v_i = \frac{\sum_{j=1}^i x(j)}{\sum_{j=1}^n x(j)}$
1	20	$\frac{1}{5} = 0.2$	$\frac{20}{255} \approx 0.078$
2	40	$\frac{2}{5} = 0.4$	$\frac{60}{255} \approx 0.235$
3	50	$\frac{3}{5} = 0.6$	$\frac{110}{255} \approx 0.431$
4	65	$\frac{4}{5} = 0.8$	$\frac{175}{255} \approx 0.686$
5	80	$\frac{5}{5} = 1$	$\frac{255}{255} = 1$

$$G = 1 - \frac{1}{n} \sum_{j=1}^n (v_{j-1} + v_j) = 1 - \frac{1}{5} ((0 + 0.078) + \dots) \approx 0.227$$

Normierter Gini-Koeffizient: $G^+ = \frac{n}{n-1} G = \frac{5}{4} \cdot 0.227 \approx 0.284$

Aufgabe 2.2

i	$x(i)$	$u_i = \frac{i}{n}$	$v_i = \frac{\sum_{j=1}^i x(j)}{\sum_{j=1}^n x(j)}$
1	20	$\frac{1}{5} = 0.2$	$\frac{20}{255} \approx 0.078$
2	40	$\frac{2}{5} = 0.4$	$\frac{60}{255} \approx 0.235$
3	50	$\frac{3}{5} = 0.6$	$\frac{110}{255} \approx 0.431$
4	65	$\frac{4}{5} = 0.8$	$\frac{175}{255} \approx 0.686$
5	80	$\frac{5}{5} = 1$	$\frac{255}{255} = 1$

$$G = 1 - \frac{1}{n} \sum_{j=1}^n (v_{j-1} + v_j) = 1 - \frac{1}{5} ((0 + 0.078) + \dots) \approx 0.227$$

Normierter Gini-Koeffizient: $G^+ = \frac{n}{n-1} G = \frac{5}{4} \cdot 0.227 \approx 0.284$

Relativ geringe Konzentration des monatlichen Taschengeldes.

Aufgabe 2.3

- (b) Wie ändert sich das in (a) berechnete Konzentrationsmaß, wenn jeder Schüler
- (i) 10 Euro mehr
 - (ii) das doppelte von seinem ursprünglichen Taschengeld bekommt?

Aufgabe 2.3

- (b) Wie ändert sich das in (a) berechnete Konzentrationsmaß, wenn jeder Schüler
- (i) 10 Euro mehr
 - (ii) das doppelte von seinem ursprünglichen Taschengeld bekommt?
- (i) Der normierte Gini-Koeffizient wird geringer, wenn jeder Schüler 10 Euro mehr Taschengeld bekommt. Die Konzentration nimmt ab, da der Anteil, den die Schüler mit dem geringsten Taschengeld erhalten, größer wird.

Aufgabe 2.3

(b) Wie ändert sich das in (a) berechnete Konzentrationsmaß, wenn jeder Schüler

- (i) 10 Euro mehr
- (ii) das doppelte von seinem ursprünglichen Taschengeld bekommt?

- (i) Der normierte Gini-Koeffizient wird geringer, wenn jeder Schüler 10 Euro mehr Taschengeld bekommt. Die Konzentration nimmt ab, da der Anteil, den die Schüler mit dem geringsten Taschengeld erhalten, größer wird.
- (ii) Der normierte Gini-Koeffizient bleibt gleich, da die Konzentration unverändert bleibt. Ändern sich die Werte aller Schüler um einen konstanten Faktor, so bleiben die Anteile am gesamten Taschengeld für alle identisch.

Aufgabe 2.4

- (c) Statt fünf Schülern werden nun 485 Schüler betrachtet. Ändert sich das Konzentrationsmaß, wenn jeweils 97 Schüler ein monatliches Taschengeld von 20 Euro, von 40 Euro, von 50 Euro, von 65 Euro und von 80 Euro bekommen?

Aufgabe 2.4

- (c) Statt fünf Schülern werden nun 485 Schüler betrachtet. Ändert sich das Konzentrationsmaß, wenn jeweils 97 Schüler ein monatliches Taschengeld von 20 Euro, von 40 Euro, von 50 Euro, von 65 Euro und von 80 Euro bekommen?
- Der Gini-Koeffizient ändert sich nicht, da die prozentuale Aufteilung der Taschengeldbeträge auf fünf Gruppen aus jeweils 97 Schülern identisch zur Aufteilung auf fünf Schüler ist.

Aufgabe 2.4

- (c) Statt fünf Schülern werden nun 485 Schüler betrachtet. Ändert sich das Konzentrationsmaß, wenn jeweils 97 Schüler ein monatliches Taschengeld von 20 Euro, von 40 Euro, von 50 Euro, von 65 Euro und von 80 Euro bekommen?
- Der Gini-Koeffizient ändert sich nicht, da die prozentuale Aufteilung der Taschengeldbeträge auf fünf Gruppen aus jeweils 97 Schülern identisch zur Aufteilung auf fünf Schüler ist.
 - Der normierte Gini-Koeffizient ändert sich jedoch, da dieser von der Stichprobengröße abhängt.

Aufgabe 3,4

Siehe Tafel



11. Stetige Verteilungen & Zufallsvektoren

Yichen Han



18. Januar 2024

1 Lehrevaluation

2 Blatt 11

Interessant zu überlegen: (Stoff für höhere Semester)
Welche Bias könnte so eine Umfrage haben?
sprich: Die Ergebnisse sind nicht 100% zuverlässig, weil...



1 Lehrevaluation

2 Blatt 11

Aufgabe 1.1

Sei $Z = X + Y$ mit $X, Y \stackrel{iid}{\sim} \mathcal{E}(\lambda)$. Zeigen Sie, dass Z Gamma-verteilt ist mit $Z \sim \mathcal{G}(\alpha = 2, \beta = \lambda)$.

Aufgabe 1.1

Sei $Z = X + Y$ mit $X, Y \stackrel{iid}{\sim} \mathcal{E}(\lambda)$. Zeigen Sie, dass Z Gamma-verteilt ist mit $Z \sim \mathcal{G}(\alpha = 2, \beta = \lambda)$.

Faltung: $f_Z(z = x + y) = \int f_X(x)f_Y(z - x) dx$ (1)

$$\mathcal{E}(\lambda) : f(x) = \lambda \exp(-\lambda x) \mathbb{I}_{[0, \infty]}(x) \quad (2)$$

$$G(\alpha, \beta) : f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathbb{I}_{[0, \infty]}(x) \quad (3)$$

$$\Gamma(\alpha) := \alpha!, \quad \forall \alpha = 1, 2, 3, \dots \quad (4)$$

Aufgabe 1.2

Zu zeigen: $f_Z(z) = \lambda^2 z \exp(-\lambda z)$ für $z \in \mathbb{R}^+$

Aufgabe 1.2

Zu zeigen: $f_Z(z) = \lambda^2 z \exp(-\lambda z)$ für $z \in \mathbb{R}^+$

$$\begin{aligned}\int_{\mathbb{R}} f(z-x)f(x)dx &= \int_0^z (\lambda \exp(-\lambda(z-x)))(\lambda \exp(-\lambda x))dx \\ &= \int_0^z \lambda^2 \exp(-\lambda z)dx \\ &= \lambda^2 \exp(-\lambda z) \int_0^z 1dx \\ &= \lambda^2 z \exp(-\lambda z).\end{aligned}$$

Aufgabe 2.1

Sei die gemeinsame Dichte von (X, Y) gegeben als

$$f(x, y) = \begin{cases} \frac{1}{x} & \text{für } 0 < y \leq x < 1 \\ 0 & \text{sonst} \end{cases}$$

- ① Was ist der Träger von (X, Y) ?
- ② Zeigen Sie dass die Randverteilung von X eine Gleichverteilung auf $(0, 1)$ ist.
- ③ Bestimmen Sie die Randverteilung von Y .
- ④ Zeigen Sie dass die bedingte Verteilung von $Y|X=x$ eine Gleichverteilung auf $(0, x)$ ist.
- ⑤ Bestimmen Sie die bedingte Verteilung von $X|Y=y$.

Aufgabe 2.1

Sei die gemeinsame Dichte von (X, Y) gegeben als

$$f(x, y) = \begin{cases} \frac{1}{x} & \text{für } 0 < y \leq x < 1 \\ 0 & \text{sonst} \end{cases}$$

- ① Was ist der Träger von (X, Y) ?
- ② Zeigen Sie dass die Randverteilung von X eine Gleichverteilung auf $(0, 1)$ ist.
- ③ Bestimmen Sie die Randverteilung von Y .
- ④ Zeigen Sie dass die bedingte Verteilung von $Y|X=x$ eine Gleichverteilung auf $(0, x)$ ist.
- ⑤ Bestimmen Sie die bedingte Verteilung von $X|Y=y$.

$$\text{Randverteilung: } f_Y(y) = \int f_{X,Y}(x, y) dx \quad (5)$$

$$\text{bedingte Verteilung: } f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad (6)$$

Aufgabe 2.2

- ① Der Träger ist $\{(x, y) : 0 < y \leq x \wedge 0 < x \cancel{<} 1\}$

Aufgabe 2.2

- ① Der Träger ist $\{(x, y) : 0 < y \leq x \wedge 0 < x \leq 1\}$
- ② Die Randverteilung von X ist

$$f_X(x) = \int_0^x \frac{1}{x} dy = \frac{1}{x} [x - 0] = 1 \quad \text{für } 0 < x < 1,$$

eine Gleichverteilung auf $(0, 1)$.

Aufgabe 2.2

- ① Der Träger ist $\{(x, y) : 0 < y \leq x \wedge 0 < x \leq 1\}$
- ② Die Randverteilung von X ist

$$f_X(x) = \int_0^x \frac{1}{x} dy = \frac{1}{x} [x - 0] = 1 \quad \text{für } 0 < x < 1,$$

eine Gleichverteilung auf $(0, 1)$.

- ③ Die Randverteilung von Y ist

$$f_Y(y) = \int_y^1 \frac{1}{x} dx = [\log(x)]_y^1 = \log\left(\frac{1}{y}\right) \quad \text{für } 0 < y < 1.$$

Aufgabe 2.3

- ① Für die bedingte Dichte von Y , gegeben $X = x$ ergibt sich:

$$\begin{aligned}f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\&= \frac{\frac{1}{x}}{1} \quad \text{für } 0 < y \leq x \\&= \begin{cases} \frac{1}{x} & \text{für } 0 < y \leq x \\ 0 & \text{sonst} \end{cases}\end{aligned}$$

d.h. $Y|X=x$ ist gleichverteilt auf $(0, x)$ ($Y|X \sim \mathcal{U}(0, x)$).

Aufgabe 2.3

- ① Für die bedingte Dichte von Y , gegeben $X = x$ ergibt sich:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{\frac{1}{x}}{1} \quad \text{für } 0 < y \leq x \\ &= \begin{cases} \frac{1}{x} & \text{für } 0 < y \leq x \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

d.h. $Y|X=x$ ist gleichverteilt auf $(0, x)$ ($Y|X \sim \mathcal{U}(0, x)$).

- ② Für die Dichte von X , gegeben $Y = y$ erhält man:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{\frac{1}{x}}{\log(\frac{1}{y})} \quad \text{für } y \leq x < 1 \\ &= \begin{cases} -1/(x \log(y)) & \text{für } y \leq x < 1 \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

Aufgabe 3.1

Für die zweidimensionale Standardnormalverteilung finden Sie in
Entsprechung zu den Vorlesungs-Folien

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(\frac{2\rho xy - x^2 - y^2}{2(1-\rho)^2}\right)$$

als Dichte.

- (a) Leiten Sie diese Gleichung aus der allgemeinen Form her:

$$f(z) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \cdot \exp\left(\frac{-1}{2}(z - \mu)^T \Sigma^{-1} (z - \mu)\right)$$

Beachten Sie dabei, dass es sich bei z und μ um Vektoren der Dimension d handelt.

Im Fall $d = 2$ haben wir $\Sigma = \begin{pmatrix} 1 & \rho \\ \bar{\rho} & 1 \end{pmatrix}$ als zweidimensionale Kovarianzmatrix.

$$\text{Cov}(Y_1, X_1) - \text{Var}(Y_1)$$

Aufgabe 3.2

$(X, Y) =: \mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma})$, mit $X, Y \sim \mathcal{N}(0, 1)$

$$f(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \cdot \exp \left(\frac{-1}{2} (\mathbf{z} - \underline{\boldsymbol{\mu}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \underline{\boldsymbol{\mu}}) \right)$$

Aufgabe 3.2

$(X, Y) =: \mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma})$, mit $X, Y \sim \mathcal{N}(0, 1)$

$$\begin{aligned}f(\mathbf{z}) &= \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \cdot \exp\left(\frac{-1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right) \\&= \frac{1}{2\pi\sqrt{\det(\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})}} \cdot \exp\left(\frac{-1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1} \mathbf{z}\right)\end{aligned}$$

Definition: Seien A eine $m \times n$ -Matrix und B eine $n \times p$ -Matrix. Das Produkt AB ist definiert als die $m \times p$ -Matrix C , wobei jedes Element c_{ij} gegeben ist durch:

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}$$

Beispiel: Vektor-Matrix-Multiplikation

Betrachten Sie den Vektor \mathbf{x} als $1 \times n$ -Matrix und A als $n \times m$ -Matrix. Die Multiplikation $\mathbf{x}A$ ist definiert als:

$$\mathbf{x}A = \left(\sum_{k=1}^n x_k a_{k1}, \sum_{k=1}^n x_k a_{k2}, \dots, \sum_{k=1}^n x_k a_{km} \right)$$

Determinante: Gegeben sei eine Matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Die Determinante von A ist definiert als:

$$\det(A) = ad - bc$$

Inverse: Die Inverse einer 2×2 -Matrix A , vorausgesetzt, dass $\det(A) \neq 0$, ist gegeben durch:

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Beispiel: Betrachten Sie die Matrix $B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Die Determinante von B ist $1 \cdot 4 - 2 \cdot 3 = -2$. Die Inverse von B , falls $\det(B) \neq 0$, ist:

$$B^{-1} = \frac{1}{-2} \begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1.5 & -0.5 \end{pmatrix}$$

Aufgabe 3.2

$(X, Y) =: \mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma})$, mit $X, Y \sim \mathcal{N}(0, 1)$

$$\begin{aligned}f(\mathbf{z}) &= \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \cdot \exp\left(\frac{-1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right) \\&= \frac{1}{2\pi \sqrt{\det(\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})}} \cdot \exp\left(\frac{-1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1} \mathbf{z}\right) \\&= \frac{1}{2\pi \sqrt{1 - \rho^2}} \cdot \exp\left(\frac{-1}{2} \begin{pmatrix} x & y \end{pmatrix} \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\right)\end{aligned}$$

Aufgabe 3.2

$(X, Y) =: \mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma})$, mit $X, Y \sim \mathcal{N}(0, 1)$

$$\begin{aligned}f(\mathbf{z}) &= \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \cdot \exp\left(\frac{-1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right) \\&= \frac{1}{2\pi\sqrt{\det(\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})}} \cdot \exp\left(\frac{-1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1} \mathbf{z}\right) \\&= \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(\frac{-1}{2} \begin{pmatrix} x & y \end{pmatrix} \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\right) \\&= \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(\frac{-1}{2(1-\rho^2)} \begin{pmatrix} x - \rho y & y - \rho x \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\right)\end{aligned}$$

Aufgabe 3.2

$$(X, Y) =: \mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma}), \text{ mit } X, Y \sim \mathcal{N}(0, 1)$$

$$\begin{aligned}f(\mathbf{z}) &= \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \cdot \exp\left(\frac{-1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right) \\&= \frac{1}{2\pi\sqrt{\det(\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})}} \cdot \exp\left(\frac{-1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1} \mathbf{z}\right) \\&= \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(\frac{-1}{2} \begin{pmatrix} x & y \end{pmatrix} \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\right) \\&= \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(\frac{-1}{2(1-\rho^2)} (x - \rho y \quad y - \rho x) \begin{pmatrix} x \\ y \end{pmatrix}\right) \\f(x, y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(\frac{2\rho xy - x^2 - y^2}{2(1-\rho)^2}\right)\end{aligned}$$

Aufgabe 3.3

(b) Wie ist der Parameter ρ zu interpretieren?

Aufgabe 3.3

(b) Wie ist der Parameter ρ zu interpretieren?

Der Parameter ρ gibt die **Kovarianz** zwischen den Zufallskomponenten / ZV X, Y an.

Aufgabe 3.3

- (b) Wie ist der Parameter ρ zu interpretieren?

Der Parameter ρ gibt die **Kovarianz** zwischen den Zufallskomponenten / ZV X, Y an.

- (c) Zeigen Sie für $d = 2$ dass aus $\rho = 0$ die stochastische Unabhängigkeit von X und Y folgt.

Aufgabe 3.3

- (b) Wie ist der Parameter ρ zu interpretieren?

Der Parameter ρ gibt die **Kovarianz** zwischen den Zufallskomponenten / ZV X, Y an.

- (c) Zeigen Sie für $d = 2$ dass aus $\rho = 0$ die stochastische Unabhängigkeit von X und Y folgt.

$$X \perp Y \Leftrightarrow f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$\begin{aligned}f_X(x)f_Y(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \\&= \frac{1}{2\pi} \exp\left(\frac{-x^2 - y^2}{2}\right) \\&\stackrel{\rho=0}{=} f_{X,Y}(x,y)\end{aligned}$$

Aufgabe 4.1

Sei $N \sim \mathcal{P}(\lambda)$ die Poisson-verteilte Anzahl an Kund:innen an einem gegebenen Tag. Seien die Umsätze U_i für die i -te Person des Tages unabhängig und identisch Gamma-verteilt: $U_i \stackrel{iid}{\sim} \mathcal{G}(\alpha, \beta)$

- ① Bestimmen Sie den Erwartungswert des täglichen Gesamtumsatzes.
- ② Bestimmen Sie die Varianz des täglichen Gesamtumsatzes.

Hinweis: Die Summe von n unabhängigen $\mathcal{G}(\alpha, \beta)$ -Zufallsvariablen ist $\mathcal{G}(n\alpha, \beta)$ -verteilt.

Aufgabe 4.1

Sei $N \sim \mathcal{P}(\lambda)$ die Poisson-verteilte Anzahl an Kund:innen an einem gegebenen Tag. Seien die Umsätze U_i für die i -te Person des Tages unabhängig und identisch Gamma-verteilt: $U_i \stackrel{iid}{\sim} \mathcal{G}(\alpha, \beta)$

- ① Bestimmen Sie den Erwartungswert des täglichen Gesamtumsatzes.
- ② Bestimmen Sie die Varianz des täglichen Gesamtumsatzes.

Hinweis: Die Summe von n unabhängigen $\mathcal{G}(\alpha, \beta)$ -Zufallsvariablen ist $\mathcal{G}(n\alpha, \beta)$ -verteilt.

Gesamtumsatz: $U = \sum_{i=1}^N U_i$, mit $U|N=n \sim \mathcal{G}(n\alpha, \beta)$

Aufgabe 4.1

Sei $N \sim \mathcal{P}(\lambda)$ die Poisson-verteilte Anzahl an Kund:innen an einem gegebenen Tag. Seien die Umsätze U_i für die i -te Person des Tages unabhängig und identisch Gamma-verteilt: $U_i \stackrel{iid}{\sim} \mathcal{G}(\alpha, \beta)$

- ① Bestimmen Sie den Erwartungswert des täglichen Gesamtumsatzes.
- ② Bestimmen Sie die Varianz des täglichen Gesamtumsatzes.

Hinweis: Die Summe von n unabhängigen $\mathcal{G}(\alpha, \beta)$ -Zufallsvariablen ist $\mathcal{G}(n\alpha, \beta)$ -verteilt.

Gesamtumsatz: $U = \sum_{i=1}^N U_i$, mit $U|(N=n) \sim \mathcal{G}(n\alpha, \beta)$

$$E(U|(N)) = \frac{N\alpha}{\beta}$$

Satz vom iterierten Erwartungswert:

$$E(U) = E_N(E(U|N)) = E_N\left(\frac{N\alpha}{\beta}\right) = \frac{\alpha}{\beta}E(N) = \frac{\alpha}{\beta}\lambda$$

Aufgabe 4.2

$$\text{Var}(U|N) = \frac{N\alpha}{\beta^2}$$

Satz von der totalen Varianz:

$$\begin{aligned}\text{Var}(U) &= E(\text{Var}(U|N)) + \text{Var}(E(U|N)) \\ &= E\left(\frac{N\alpha}{\beta^2}\right) + \text{Var}\left(\frac{N\alpha}{\beta}\right) \\ &= \frac{\lambda\alpha}{\beta^2} + \frac{\lambda\alpha^2}{\beta^2} \quad \text{Varianz quadratisch!} \\ &= \frac{\lambda\alpha(1+\alpha)}{\beta^2}\end{aligned}$$

Aufgabe 4.3

Aus der Angabe können wir schließen, dass $E(N) = 120 = \lambda$ und $E(U_i) = 10 = \frac{\alpha}{\beta}$ und $Var(U_i) = 10 = \frac{\alpha}{\beta^2}$, also $\alpha = 10; \beta = 1$. Also ist hier $E(U) = 1200$ und $Var(U) = 13200$
Siehe R Simulation.