

Übungen zur Deskr. Statistik und Stochastik

Wolfgang Bischof

Sommersemester 2024

Technische Hochschule Augsburg

Data Science

11. März 2025

Klassifikation der Aufgaben:

- ○ ○ ○ leichte Aufgaben
- ● ○ ○ mittelschwere Aufgaben
- ● ● ○ schwere Aufgaben
- ● ● ● Aufgaben zur weiteren Vertiefung

Aufgabe 1.1: R als Taschenrechner (2 Punkte)



Berechnen Sie:

(i) $\sin(2^3\pi) + e$

(ii) $2 + 3^8$

(iii) $2^{\tan(17+e^3)}$

Weisen Sie dem Vektor a die Zahlen 1,2 und 3 zu.

- (i) Welches Ergebnis erhält man bei $a * a$?
- (ii) Multiplizieren Sie a komponentenweise mit 5.
- (iii) Quadrieren Sie alle Einträge von a .
- (iv) Erhöhen Sie alle Einträge um 1.
- (v) Berechnen Sie das Skalarprodukt $a \cdot a$.



Berechnen Sie:

(i) $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$

(ii) $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1}$

Als Vorarbeit generieren Sie bitte mit den Befehlen

```
set.seed(1907) # für die Reproduzierbarkeit
schaeden <- rnorm(10000, 10000, 2000) # Schadenaufwände von 10000 KFZ-Schäden
alter_vn <- runif(10000, 18, 80) # Alter des Versicherungsnehmers
geschlecht <- sample( c("m","w","d"), 10000, replace=TRUE, prob=c(51, 48, 1))
alter_kfz <- runif(10000, 0, 15) # Alter des Fahrzeugs
```

einen Datensatz aus zehntausend KFZ-Schäden mit Angaben zum Alter und Geschlecht des Versicherungsnehmers und zum Fahrzeugalter. Dabei entspricht ein Index einem Schaden, z.B. gibt `schaeden[5764]` die Schadenhöhe des Schadens mit Nummer 5764 und `geschlecht[5764]` das Geschlecht des Versicherungsnehmers zu diesem Schaden an.

- (i) Ermitteln Sie die Nummer des kleinsten und größten Schadens und das Alter des Versicherungsnehmers und des Fahrzeuges zum Schaden mit dem kleinsten bzw. größten Aufwand. Sie dürfen dabei davon ausgehen, dass alle Schäden verschieden hoch sind.
- (ii) Ermitteln Sie das arithmetische Mittel und den Median sowie die Stichproben-Standardabweichung und -Varianz aller Schäden.
- (iii) Wie viele Schäden über 10000 mit einem männlichen Versicherungsnehmer gibt es? Bei wie vielen davon war das Fahrzeug unter 20 [Jahren] alt?
- (iv) Errechnen Sie das 40%-Quantil und erstellen Sie eine Liste der Quantile von 5% bis 95%, wobei die Schrittweite 10% betrage.
- (v) Erstellen Sie einen Vektor der Großschäden, die größer als 15000 sind.
- (vi) Zerlegen Sie die Schäden mittels Indexselektion in die zwei Teilmengen „männlich“ und „weiblich“.

- (i) Ein Geschäftsführer eines großen Supermarkts erfasst 20 Tage lang jeweils die Bediendauern von 1000 Kunden. Sie können seine Werte mit den Befehlen

```
set.seed(1972) # für Reproduzierbarkeit
Bediendauern <- matrix( rexp(20000), ncol=20 )
```

simulieren, wobei in jeder Spalte die 1000 Beobachtungen eines Tages stehen.

Ermitteln Sie für jeden Tag die Summe, das arithmetische Mittel sowie die Stichprobenstandardabweichung der Bediendauern der 1000 Kunden.

- (ii) Nun erfasst der Geschäftsführer 20 Tage lang jeweils die Bediendauern von allen Kunden, die in seinem Supermarkt einkaufen. Die Anzahl der Kunden pro Tag können Sie mit dem Befehl

```
Anzahl_Kunden <- rpois(20, 2000)
```

simulieren.

Wie können Sie die Liste der Bediendauern aller Kunden simulieren, wenn je Listenelement alle Bediendauern eines Tages stehen sollen?

Berechnen Sie damit für jeden Tag die Summe, das arithmetische Mittel sowie die Stichprobenstandardabweichung der Bediendauern aller Kunden des jeweiligen Tages.

Hinweise:

- Die Bediendauern von n Kunden können Sie durch `rexp(n)` simulieren.
- Verwenden Sie den Befehl `lapply` zweimal: zunächst zur Simulation der Bediendauern der Kunden und anschließend nochmal für die Berechnung der Summe, des arithmetischen Mittels sowie der Stichprobenstandardabweichung der Bediendauern.

- (iii) Dem Geschäftsführer ist aufgefallen, dass die Bediendauern im Laufe eines Tages immer kürzer werden. Offenbar gewöhnt sich das Personal an die Bedienung und wird immer schneller. Tatsächlich hat er beobachtet, dass er die Bediendauer des n -ten Kunden eines Tages mit dem Befehl

```
rexp(1, 1/(0.5+exp(-n)))
```

simulieren kann (nur zu Ihrer Information: im Mittel ist die Bediendauer des n -ten Kunden damit $0.5 + e^{-n/5}$; die sogenannte Ankunftsrate beträgt $1/(0.5 + e^{-n/5})$).

Simulieren Sie erneut die Liste der Bediendauern aller Kunden, wobei je Listenelement alle Bediendauern eines Tages stehen sollen und die mittleren Bediendauern je Tag nun entsprechend obigem Bildungsgesetz sinken sollen.

Berechnen Sie damit für jeden Tag die Summe, das arithmetische Mittel sowie die Stichprobenstandardabweichung der Bediendauern aller Kunden des jeweiligen Tages.

Berechnen Sie ferner zur Überprüfung, ob die mittleren Bediendauern tatsächlich sinken, das arithmetische Mittel der Bediendauern der ersten, dritten und zehnten Tages.

Hinweise:

- Verwenden Sie die Funktion:

```
erzeuge_Bediendauern_eines_Tages <- function( N )
{
  return ( rexp( N, 1/(0.5+exp(-(1:N)/5)) ) )
}
```

Beschreiben Sie den Rückgabewert dieser Funktion.

Verwenden Sie anschließend diese Funktion in Kombination mit dem `lapply`-Befehl, um die Bediendauern aller Kunden zu simulieren.

- Eine Liste kann mit dem Befehl `unlist` in einen Vektor umgewandelt werden.

Aufgabe 2.1: Kontingenztabellen (Teil (iii): R) (8 Punkte)



In der folgenden Kontingenztafel sind die Informationen über das Geschlecht (weiblich bzw. männlich) und den Überlebensstatus (überlebend bzw. nicht-überlebend) zum Untergang der Titanic zusammengefasst.

	überlebend	nicht-überlebend	Σ
männlich	367	1364	
weiblich	344	126	
Σ			

- (i) Spezifizieren Sie für die gegebenen Daten die Begriffe bivariate Stichprobe, Merkmale, Merkmalsträger, Skalenniveaus und Stichprobenumfang.
- (ii) Bestimmen Sie die bedingten relativen Häufigkeitsverteilungen.
- (iii) Stellen Sie eine bedingte relative Häufigkeitsverteilung grafisch dar.
Hinweis: Im R-internen Datensatz `titanic` finden Sie u.a. die oben genannten Merkmale für alle Passagiere.
- (iv) Vergleichen Sie die Zellhäufigkeiten in der Kontingenztafel mit den erwarteten Häufigkeiten bei Unabhängigkeit von Geschlecht und Überlebensstatus.

- (i) Lesen Sie die Datei „KfzDat.csv“ in das data frame `dat.kfz` ein. Die Datei finden Sie in unserem Moodle-Kursraum.
- (ii) Verschaffen Sie sich einen Überblick über den Datensatz (z.B. mithilfe der Funktionen `summary` und `str`).
- (iii) Ordnen Sie die Werte des Merkmals `TypklasseVK` aus `dat.kfz` einer der Kategorien 0-10, 11-15, 16-20, 21-25, 26-34 zu und überschreiben Sie den Wert des Merkmals `TypklasseVK` mit der entsprechenden Kategorie. Überprüfen Sie mit der Funktion `summary`, ob alle Werte einer Kategorie zugeordnet wurden.
- (iv) Warum ergibt
- ```
PS <- pull(dat.kfz, "PS")
length(PS[PS<=100])+length(PS[PS>100])
```
- mehr als die Gesamtanzahl aller Beobachtungen `nrow(dat.kfz.ps)`?
- Mit welchem Ausdruck kann man die Anzahl der Einträge mit mehr als 100 PS ermitteln?
- Hinweis: Mit der Funktion `pull` kann man ein Merkmal eines Tibbles als Vektor extrahieren.*
- (v) Löschen Sie alle Ausreißer mit  $PS > 1000$ , alle Beobachtungen mit negativem  $PS$ -Wert, sowie alle Beobachtungen bei denen irgendein Wert fehlt. Kontrollieren Sie Ihr Ergebnis mit `summary( dat.kfz )`.
- (vi) Erstellen Sie ein neues Merkmal `PS.Kat` vom Datentyp Faktor, in der Sie die Werte des Merkmals `PS` einer der Kategorien 0-70, 71-90, 91-110, 111-130, 131-150, 151-200, 201-1000 zuordnen.
- (vii) Erstellen Sie eine zweidimensionale Kontingenztafel für das arithmetische Mittel des Preises je PS-/VK-Typklasse-Kategorie-Kombination.
- Erstellen Sie ferner eine Kontingenztafel für die absoluten Häufigkeiten je PS-/VK-Typklasse-Kategorie-Kombination.
- Hinweis: Verwenden Sie die Hilfefunktion, um sich die Varianten der Funktionen `tapply` und `table` für zwei Faktoren anzusehen.*

## Aufgabe 2.3: Potenzmenge und relative Häufigkeit (7 Punkte)



Gegeben sei eine Stichprobe  $x = (x_1, \dots, x_n)$ ,  $n \in \mathbb{N}$ , mit der Ausprägungsmenge  $A = \{a_1, \dots, a_m\}$ ,  $2 \leq m$ . Für  $B \subseteq A$  sei

$$f : \mathcal{P}(A) \rightarrow [0,1], f(B) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_B(x_i),$$

wobei  $\mathcal{P}(A) := \{B : B \subseteq A\}$  die Potenzmenge von  $A$  bezeichnet.

- (i) Geben Sie exemplarisch 3 Elemente von  $\mathcal{P}(A)$  an und bestimmen Sie die Mächtigkeit  $|\mathcal{P}(A)|$ .
- (ii) Zeigen Sie, dass für die Abbildung  $f$  die folgenden Eigenschaften (a) und (b) gelten:
  - (a)  $f(A) = 1$ .
  - (b) Für alle  $B_1, B_2 \subseteq A$ ,  $B_1 \cap B_2 = \emptyset$ , gilt, dass  $f(B_1 \cup B_2) = f(B_1) + f(B_2)$ .
- (iii) Zeigen Sie, dass für  $B \subseteq A$  im Allgemeinen weder  $f(B) = \frac{|B|}{n}$  noch  $f(B) = \frac{|B|}{m}$  gilt.

(i) Seien  $\Omega$  eine beliebige, nichtleere Menge und  $A, B$  zwei Teilmengen von  $\Omega$ . Zeigen oder widerlegen Sie:

- (a)  $\mathbb{1}_{A \cap B} = \mathbb{1}_A \cdot \mathbb{1}_B$
- (b)  $\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B$
- (c) Für alle natürlichen Zahlen  $n$  gilt:  $(\mathbb{1}_A)^n = \mathbb{1}_A$
- (d)  $\mathbb{1}_{A \setminus B} = \mathbb{1}_A - \mathbb{1}_B$

*Hinweise:*

- Überlegen Sie sich, um welche Gleichheiten es sich bei diesen vier Teilaufgaben genau handelt und was sie bedeuten.
  - Verwenden Sie, dass Indikatorfunktionen lediglich die Werte 0 und 1 annehmen können.
- (ii) Die Indikatorfunktion ist oft nützlich, um Fallunterscheidungen zusammenzufassen. Schreiben Sie folgende Funktion ohne explizite Fallunterscheidung mit einer Indikatorfunktion:

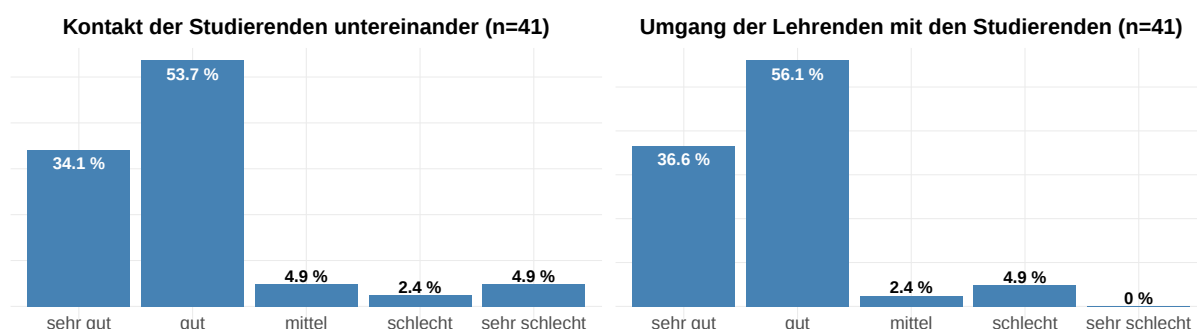
$$f : \mathbb{R} \rightarrow \mathbb{R}; \quad x \mapsto \begin{cases} 0 & \text{falls } x < 0 \\ x & \text{sonst} \end{cases}$$

- (i) Lesen Sie die Datei „AbsolventenDat.csv“ in den data frame `dat.absolventen` ein. Beachten Sie, dass fehlende Werte teils mit „-“ und teils mit „k.A.“ gekennzeichnet sind. Die Datei finden Sie in unserem Moodle-Kursraum.
  - (ii) Lesen Sie die Komponente „Geschlecht“ des data frames in den Vektor `geschlecht` ein und erstellen Sie eine Häufigkeitstabelle dieses Vektors.
  - (iii) Erzeugen Sie mit dem Befehl `barplot` oder mit `geom_bar` ein einfaches Säulendiagramm der Häufigkeitstabelle des Merkmals `geschlecht` ohne weitere Formatierung.
  - (iv) Definieren Sie einen Vektor `plot.erg`, in den Sie das Ergebnis des `barplot`-Kommandos schreiben. Lesen Sie die Hilfe zum Kommando `barplot`, um zu verstehen, welche Werte der Vektor `plot.erg` enthält.
  - (v) Verwenden Sie den Vektor `plot.erg` und die Häufigkeitstabelle aus Teilaufgabe (ii), um die Säulen mit dem `text`-Kommando mit dem String  

```
paste(round(prop.table(table(geschlecht))*100,1),"%")
```

der die relativen Häufigkeiten der beiden Geschlechter in Prozent enthält, zu beschriften.
  - (vi) Schreiben Sie eine Funktion `pretty.barplot( dat.kat )`, die für einen Vektor `dat.kat` mit kategorialen Daten ein Säulendiagramm zeichnet, das mit den relativen Häufigkeiten beschriftet ist. Testen Sie Ihre Funktion mit dem Vektor `geschlecht`.
  - (vii) Verwenden Sie nun die Option `main` des `barplot`-Befehls, um die `pretty.barplot`-Funktion um das Argument `titel` zu erweitern, so dass der String `titel` als Überschrift des Säulendiagramms ausgegeben wird. Verwenden Sie ferner die Option `axes`, damit keine y-Achse gezeichnet wird. Testen Sie Ihre Funktion mit dem Vektor `geschlecht` und dem Titel „Geschlecht“.
  - (viii) Im Titel soll nun neben dem übergebenen String `Titel` noch die Anzahl der gültigen Werte ausgegeben werden. Testen Sie Ihre Funktion wieder mit dem Vektor `geschlecht` und dem Titel „Geschlecht“.
- Hinweise:*
- i.) Verwenden Sie die Befehle `length` und `is.na`, um herauszufinden, wie viele gültige Werte der Vektor `geschlecht` enthält.
  - ii.) Mit dem Befehl `paste` können mehrere Strings zu einem zusammen gebaut werden.
- (ix) Verwenden Sie die Funktion `pretty.barplot`, um Säulendiagramme für alle kategorialen Komponenten von `dat.absolventen` zu zeichnen.
  - (x) Verschönern Sie die Säulendiagramme aus Teilaufgabe (ix) noch weiter. Achten Sie dabei auf folgende Details:
    - Prinzipiell soll die Prozentzahl oben in weißer Farbe in die dunkelblauen Säulen geschrieben werden. Wenn eine Säule zu klein für die Beschriftung ist, soll die relative Anzahl in schwarzer Farbe über die Säule geschrieben werden.
    - Achten Sie auf die korrekte Reihenfolge der Kategorien in dem Säulendiagramm. Beispielsweise sollte die Reihenfolge bei den Beurteilungsfragen „sehr gut“, „gut“, „mittel“, „schlecht“, „sehr schlecht“ sein. Außerdem sollen bei den Beurteilungsfragen auch Ausprägungen auf der x-Achse aufgeführt werden, zu denen es keine Beobachtungen gibt, um die verschiedenen Fragen besser vergleichen zu können.

Beispielsweise könnte die Grafik so aussehen:



*Hinweis: Wenn Sie wollen, dürfen Sie gerne die Pakete `tidyverse` und `ggplot2` verwenden.*



## Aufgabe 3.1: Quantile, emp. Vtlgsfkt. und Boxplot (teilw. R) (5 Punkte)



Gegeben ist der folgende Datenvektor:

```
x <- c(1, 2, 1, 2, 10, 10, 20, 1, 2, 1)
```

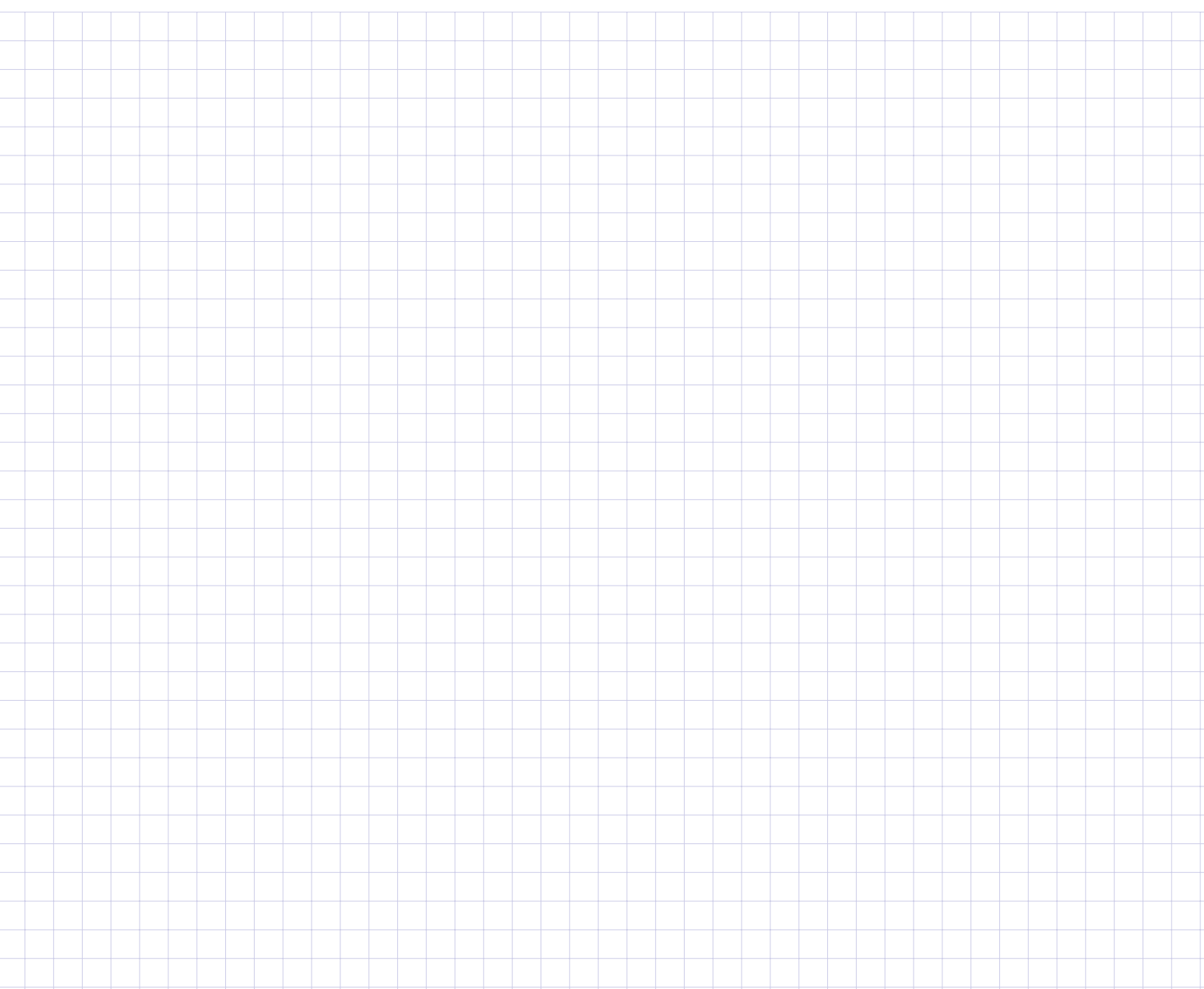
Lösen Sie folgende Aufgaben jeweils zuerst auf Papier und dann mit R:

- (i) Geben Sie die empirische Verteilungsfunktion  $F_{10}(x)$  an und zeichnen Sie den Graph von  $F$ . Mit R brauchen Sie nur den Graphen zu zeichnen.
- (ii) Berechnen Sie jeweils alle der folgenden empirischen Quantile  $x_{p_i}$  mit

```
p <- c(0.2, 0.25, 0.4, 0.5, 0.6, 0.75, 0.99)
```

wobei  $p_i$  der  $i$ -te Vektoreintrag ist ( $i = 1, \dots, 7$ ).

- (iii) Zeichnen Sie einen Boxplot zum Datenvektor  $x$ .



50 Hochschulabsolventen wurden bezüglich ihrer Studiendauer (in Semestern) befragt.

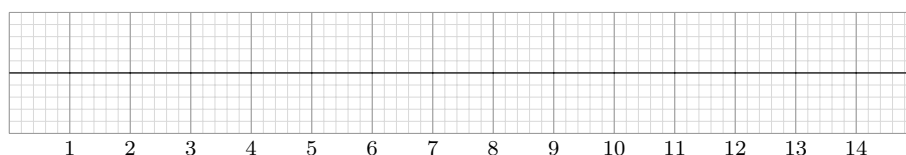
| Anzahl Semester    | $a_i$         | 6 | 7    | 8    | 9  | 10   | 11 | 14   |
|--------------------|---------------|---|------|------|----|------|----|------|
| Anzahl Absolventen | $h_i$         | 3 | 9    |      | 11 |      | 2  | 1    |
|                    | $H_i$         |   |      | 30   | 41 |      |    |      |
|                    | $f_i$         |   |      | 0,36 |    |      |    | 0,02 |
|                    | $F_{50}(a_i)$ |   | 0,24 |      |    | 0,94 |    | 1    |

- (i) Beim Aufschreiben und Auswerten sind einige Teile der Tabelle verloren gegangen. Vervollständigen Sie diese.

Hinweis:  $H_i$  bezeichnet die kumulative, absolute Häufigkeit bis zur Ausprägung  $a_i$ , also  $H_i := \sum_{j:a_j \leq a_i} h_j$ .

- (ii) Bestimmen Sie händisch alle fünf Quartile.

- (iii) Zeichnen Sie einen Boxplot der Daten:



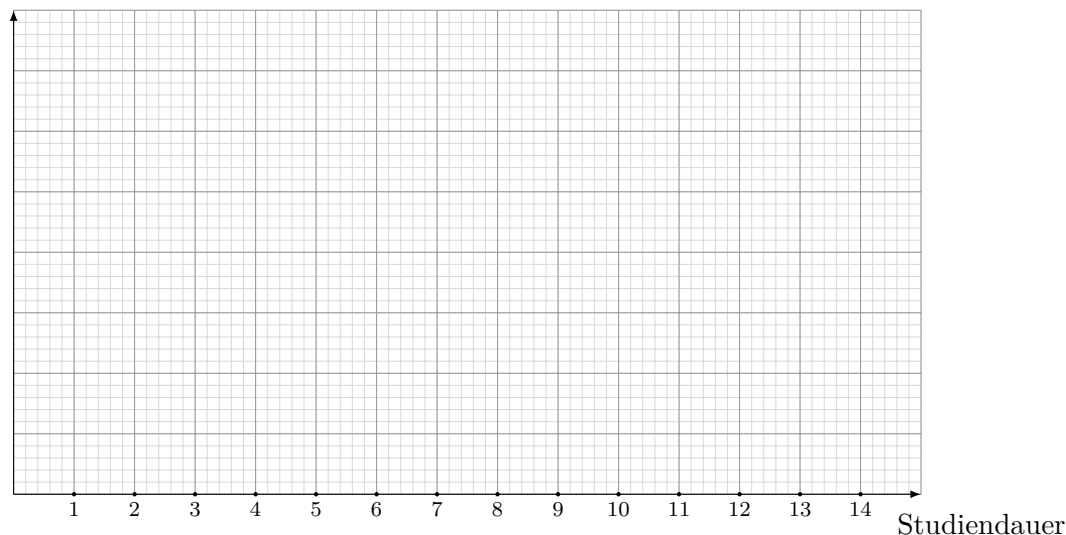
- (iv) Geben Sie jeweils *einen* R-Befehl an mit dem Sie

- die obigen Daten einem R-Objekt (Vektor)  $x$  zuweisen,
- alle Quartile ausgeben und
- das (mittlere) 25%, 50%, 75% und 95%-Quantil gleichzeitig aus dem Vektor  $x$  berechnen können.

- (v) Die Daten sollen nun klassiert werden in *schnelle Absolventen* (mehr als vier, aber weniger als acht Semester), *normale Absolventen* und *langsame Absolventen* (neun bis maximal 14 Semester). Füllen Sie dazu die Tabelle auf der rechten Seite aus und zeichnen Sie das zugehörige Histogramm. Die Fläche des Histogramms soll auf 1 normiert sein.

| Klassen | [4, 8) |  | [9, 14] |
|---------|--------|--|---------|
| Anzahl  |        |  |         |
| Breite  |        |  |         |
| Höhe    |        |  |         |

Häufigkeitsdichte



Bemerkung: Beschriften Sie die Achsen!

## Aufgabe 3.3: Empirische Verteilungs- und Quantilfunktion (8 Punkte)



Um den Zeitaufwand für einen bestimmten Arbeitsvorgang einzuschätzen, wurde die Dauer  $x_i, i = 1, \dots, 100$ , des Arbeitsvorgangs in einer Stichprobe  $x = (x_1, \dots, x_{100})$  bei  $n = 100$  verschiedenen Mitarbeitern erfasst. Die Stichprobe ist in der folgenden Häufigkeitstabelle zusammengefasst:

| Zeitdauer $a_i$ (in Minuten) | 4  | 6  | 9  | 12 |
|------------------------------|----|----|----|----|
| absolute Häufigkeit $h(a_i)$ | 30 | 20 | 40 | 10 |

- (i) Bestimmen Sie die empirische Verteilungsfunktion  $F_{100}$  der Stichprobe  $x$  und skizzieren Sie  $F_{100}$ .
- (ii) Bestimmen Sie  $F_{100}(10), F_{100}(15)$  und interpretieren Sie die Ergebnisse.
- (iii) Bestimmen Sie  $F_{100}^{-1}(\{\frac{1}{2}\})$  und  $F_{100}^{-1}(\{\frac{3}{4}\})$ , wobei  $F_{100}^{-1}(I), I \subseteq [0; 1]$ , das Urbild von  $I$  unter  $F_{100}$  bezeichnet.
- (iv) Entscheiden und begründen Sie, ob  $F_{100} : \mathbb{R} \rightarrow F_{100}(\mathbb{R})$  eine inverse Abbildung besitzt, also bijektiv ist.
- (v) Bestimmen Sie  $F_{100}^{-1}(\frac{1}{2})$  und  $F_{100}^{-1}(\frac{3}{4})$ , wobei  $F_{100}^{-1}$ , wie in der Vorlesung definiert, die Quantilfunktion bezeichnet.

Die Jahresumsätze von 1000 Unternehmen sind in der folgenden Tabelle klassifiziert gegeben:

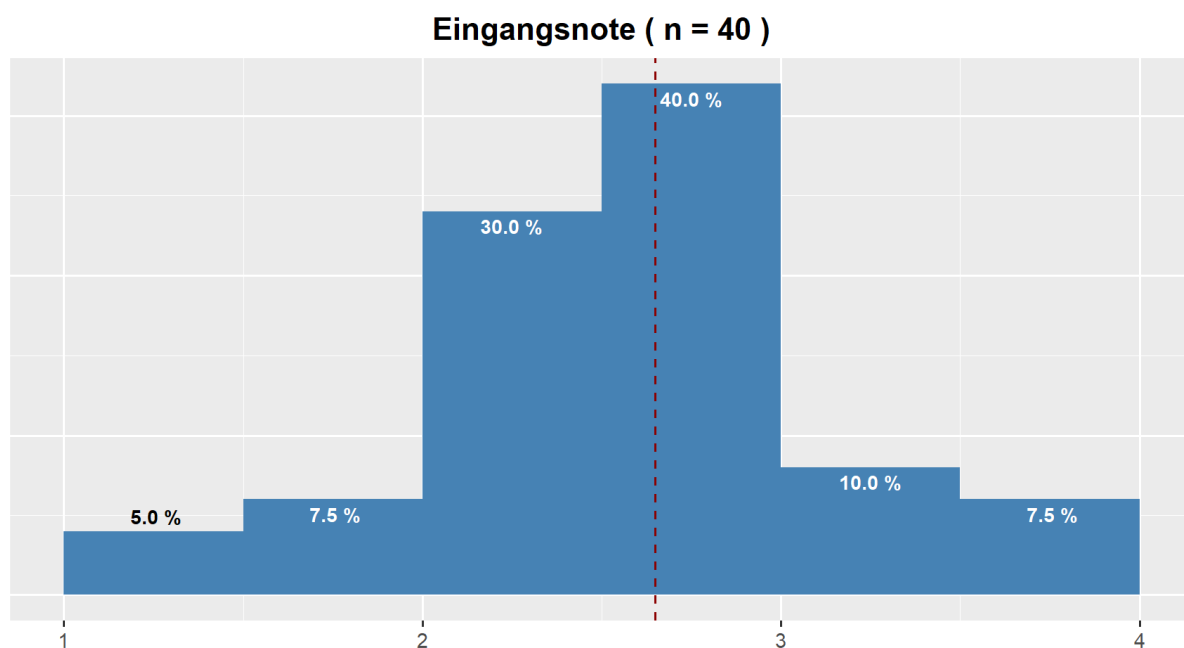
| Umsatz [in Mio. Euro ] | [1,5) | [5,10) | [10,20) | [20,40] |
|------------------------|-------|--------|---------|---------|
| Anzahl Unternehmen     | 490   | 410    | 50      | 50      |

- (i) Spezifizieren Sie für die gegebenen Daten die Begriffe Stichprobe, Merkmal, Merkmalsträger, Skalenniveau und Stichprobenumfang.
- (ii) Skizzieren Sie mit der gegebenen Intervalleinteilung das zugehörige Histogramm der absoluten Klassenhäufigkeiten, wobei die Höhe  $H_1$  des Histogrammrechtecks der ersten Klasse  $[1,5)$  [Mio. Euro]  $H_1 = 20$  [cm] sein soll.

### Aufgabe 3.5: Histogramme (keine Punkte)



- (i) Lesen Sie die Datei `AbsolventenDat.csv`, die Sie in unserem Moodle-Kursraum finden, ein. Achten Sie dabei auf das korrekte Einlesen von fehlenden Werten.
- (ii) Lassen Sie sich mit `summary` die wichtigsten Kenngrößen des Merkmals „Eingangsnote“ des data frames `dat.absolventen` ausgeben.
- (iii) Erzeugen Sie ein einfaches Histogramm des Merkmals „Eingangsnote“ ohne weitere Formatierung. Dabei sollen alle Intervalle die Länge 0,5 haben und das erste bei 1 starten.
- (iv) Verwenden Sie das `stat_bin`-Kommando, um ein Histogramm zu zeichnen, bei dem die Rechtecke mit den entsprechenden relativen Häufigkeiten (in Prozent) beschriftet sind. Über 5% soll die Häufigkeit im Rechteck angegeben werden. Darunter über dem Rechteck. Die *y*-Achse soll mit „Anzahl“ statt „count“ beschriftet werden.
- (v) Schreiben Sie eine Funktion `pretty.hist( dat.num )`, die für einen numerischen Vektor `dat.num` ein Histogramm zeichnet, das mit den relativen Häufigkeiten beschriftet ist.  
Testen Sie Ihre Funktion mit dem Vektor `dat.absolventen$Abschlussnote`.
- (vi) Erweitern Sie nun die `pretty.hist`-Funktion um das Argument `titel`. Im Titel des Histogramms soll neben dem übergebenen String `titel` noch die Anzahl der gültigen Werte ausgegeben werden. Außerdem sollen die x-Achse keinen Titel haben und die y-Achse ganz ausgeblendet wird. Testen Sie Ihre Funktion wieder mit dem Merkmal `Eingangsnote` und dem Titel „Eingangsnote“.
- (vii) Zeichnen Sie in das Histogramm beim arithmetische Mittel (vulgo: Durchschnitt) eine gestrichelte senkrechte Linie ein und beschriften Sie diese Linie mit dem Durchschnittswert.  
Beispielsweise könnte die fertige Grafik so aussehen:



## Aufgabe 4.1: Säulendiagramme (2 Punkte)



Betrachten Sie den Datenvektor  $x = (x_1, \dots, x_{20})$ , wobei jedes  $x_i$  einen der vier Werte 1, 2, 3 und 4 annimmt. Zeichnen Sie jeweils ein passendes Säulendiagramme für die absoluten Häufigkeiten  $h_i$  ( $i = 1, \dots, 4$ ), wobei folgende Eigenschaften erfüllt sein müssen:

- (i)  $h_i$  ist symmetrisch zum arithmetischen Mittel  $\bar{x} = 3$  und  $h_3 = 0$ .
- (ii)  $x_{mod} = 1$  und  $x_{\frac{1}{2}} = \frac{3}{2}$ .

#### Aufgabe 4.2: Lagemaße (3 Punkte)

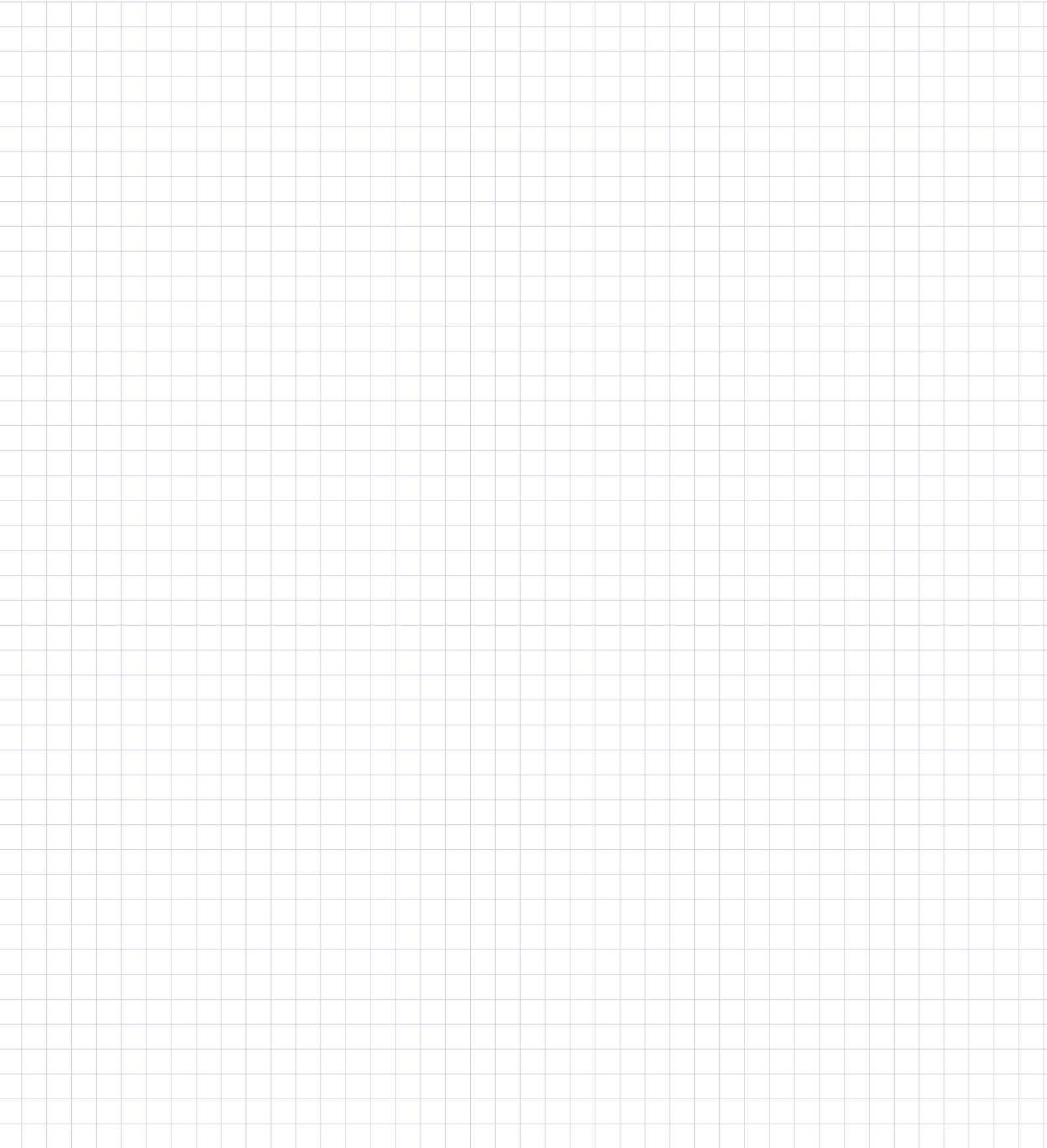


Für 10 Unternehmen ist jeweils eine Kreditwürdigkeitsbewertung

$$x = (x_1, \dots, x_{10}) = (A, A, B, B, A, A, C, A, C, D)$$

mit der ordinaler Skala:  $A$  = sehr kreditwürdig über  $B, C$  zu  $D$  = nicht kreditwürdig gegeben.

- (i) Bestimmen Sie die möglichen Lagemaße für die Mitte von  $x$ .
- (ii) Bestimmen Sie das untere und obere Quartil von  $x$ .



## Aufgabe 4.3: Lagemaße 2 (3 Punkte)



Um den Zeitaufwand für einen bestimmten Arbeitsvorgang einzuschätzen, wurde die Dauer  $x_i, i = 1, \dots, 100$ , des Arbeitsvorgangs in einer Stichprobe  $x = (x_1, \dots, x_{100})$  bei  $n = 100$  verschiedenen Mitarbeitern erfasst. Die Stichprobe ist in der folgenden Häufigkeitstabelle zusammengefasst:

|                              |    |    |    |    |
|------------------------------|----|----|----|----|
| Zeitdauer $a_i$ (in Minuten) | 4  | 6  | 9  | 12 |
| absolute Häufigkeit $h(a_i)$ | 30 | 20 | 40 | 10 |

- (i) Bestimmen Sie die möglichen Lagemaße für die Mitte von  $x$ .
- (ii) Bestimmen Sie das untere und obere Quartil von  $x$ .



Betrachten Sie den Datenvektor:

$$x = (1, 1, 5, 2, 5, 2, 0)$$

und den erweiterten Datenvektor

$$y = (1, 1, 5, 2, 5, 2, 0, x_0)$$

mit einer Variable  $x_0 \in \mathbb{R}$ .

- (i) Berechnen Sie den Median  $x_{\frac{1}{2}}$ , das arithmetische Mittel (salopp: Durchschnitt)  $\bar{x}$ , und den Modalwert  $x_{mod}$  des Datenvektors  $x$ .
- (ii) Berechnen Sie den Median  $y_{\frac{1}{2}}$ , das arithmetische Mittel  $\bar{y}$  und den Modalwert  $y_{mod}$  als Funktion von  $x_0$ .
- (iii) Zeichnen Sie die Graphen der Funktionen  $y_{\frac{1}{2}}(x_0)$  und  $\bar{y}(x_0)$ .
- (iv) Erklären Sie mit Hilfe der Zeichnung den wesentlichen Unterschied zwischen arithmetischem Mittel und Median.

#### Aufgabe 4.5: Eigenschaften von arithmetischem Mittel und Median (8 Punkte)



Sei  $x = (x_1, \dots, x_n)$ ,  $n \in \mathbb{N}$ , eine Stichprobe eines metrischen Merkmals mit arithmetischem Mittel  $\bar{x}$  und einem empirischen 50%-Quantil  $x_{0,5}$ . Zeigen Sie die Minimierungseigenschaften:

(i)  $\arg \min_{z \in \mathbb{R}} \sum_{i=1}^n (x_i - z)^2 = \bar{x}$

(ii)  $x_{0,5} \in \arg \min_{z \in \mathbb{R}} \sum_{i=1}^n |x_i - z|$

*Hinweise:*

- Die Funktion  $\arg \min$  liefert alle Minimalstelle(n) einer Funktion. Wenn es nur eine Minimalstelle gibt, ist das ein Wert, sonst eine Menge. Beispielsweise gilt  $\arg \min_{x \in \mathbb{R}} (x - 2)^2 = 2$  (das Minimum dieser Funktion beträgt 0) und  $\arg \min_{x \in \mathbb{R}} \cos(x) = \{2k\pi | k \in \mathbb{Z}\}$  (das Minimum dieser Funktion beträgt -1).
- Teil (ii) der Aufgabe ist relativ schwierig. Wenn Sie den Beweis nicht allgemein hinkriegen, versuchen Sie es bitte für die Spezialfälle  $n = 2$  und  $n = 3$ . Und dann für  $n = 4$ .

#### Aufgabe 4.6: Verschiebungsformel für Stichprobenvarianz (4 Punkte)



Sei  $x = (x_1, \dots, x_n)$  eine Stichprobe reeller Zahlen mit arithmetischem Mittel  $\bar{x}$  und Stichproben-Varianz  $s^2$ . Zeigen Sie, dass für alle  $c \in \mathbb{R}$  die Verschiebungsformel

$$(n-1)s^2 = \sum_{i=1}^n (x_i - c)^2 - n(\bar{x} - c)^2$$

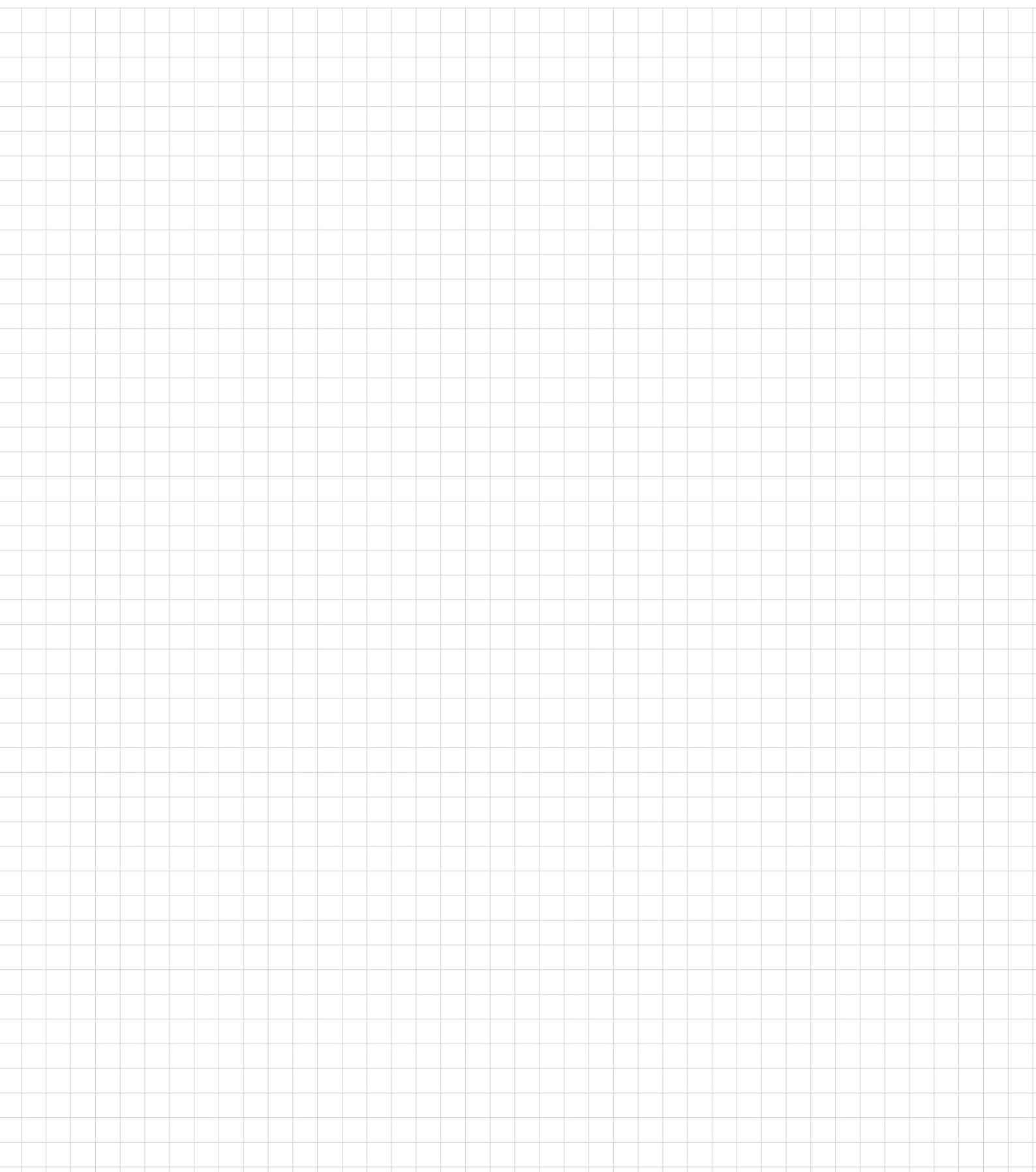
gilt.



#### Aufgabe 4.7: Histogramme, Säulen- und Streudiagramme mit R (5 Punkte)



- (i) Zeichnen Sie Histogramme für die Merkmale `mpg`, `cyl`, `hp`, `wt` der R-internen Datentabelle *mtcars*.
- (ii) Finden Sie heraus, für welche der Merkmale `mpg`, `cyl`, `hp`, `wt` ein Säulendiagramm sinnvoll ist.
- (iii) Zeichnen Sie für die in (ii) ermittelten Merkmale Säulendiagramme.
- (iv) Zeichnen Sie jeweils ein Streudiagramm für die Merkmalspaare `mpg ~ hp`, `mpg ~ wt` und interpretieren Sie die Diagramme. Berechnen Sie die zugehörigen Pearson-Korrelationskoeffizienten.



#### Aufgabe 4.8: Streuungszerlegung (keine Punkte)



Gegeben seien  $m \in \mathbb{N}$  (Teil-)Stichproben (man sagt auch Schichten)  $x_1, \dots, x_m$  reeller Zahlen mit den Umfängen  $n_1, \dots, n_m$ ,  $n := \sum_{i=1}^m n_i$ , arithmetischen Mitteln  $\bar{x}_1, \dots, \bar{x}_m$  und Stichproben-Varianzen  $s_1^2, \dots, s_m^2$ . Zeigen Sie, dass für das arithmetische Mittel  $\bar{z}$  und die Stichproben-Varianz  $s_z^2$  der gemeinsamen Stichprobe  $z = (z_1, \dots, z_n)$ , die aus allen Stichprobenwerten der  $m$  Teilstichproben besteht, folgende Formeln gelten.

(i)  $\bar{z} = \frac{1}{n} \sum_{i=1}^m n_i \bar{x}_i$  (gewichtetes Mittel)

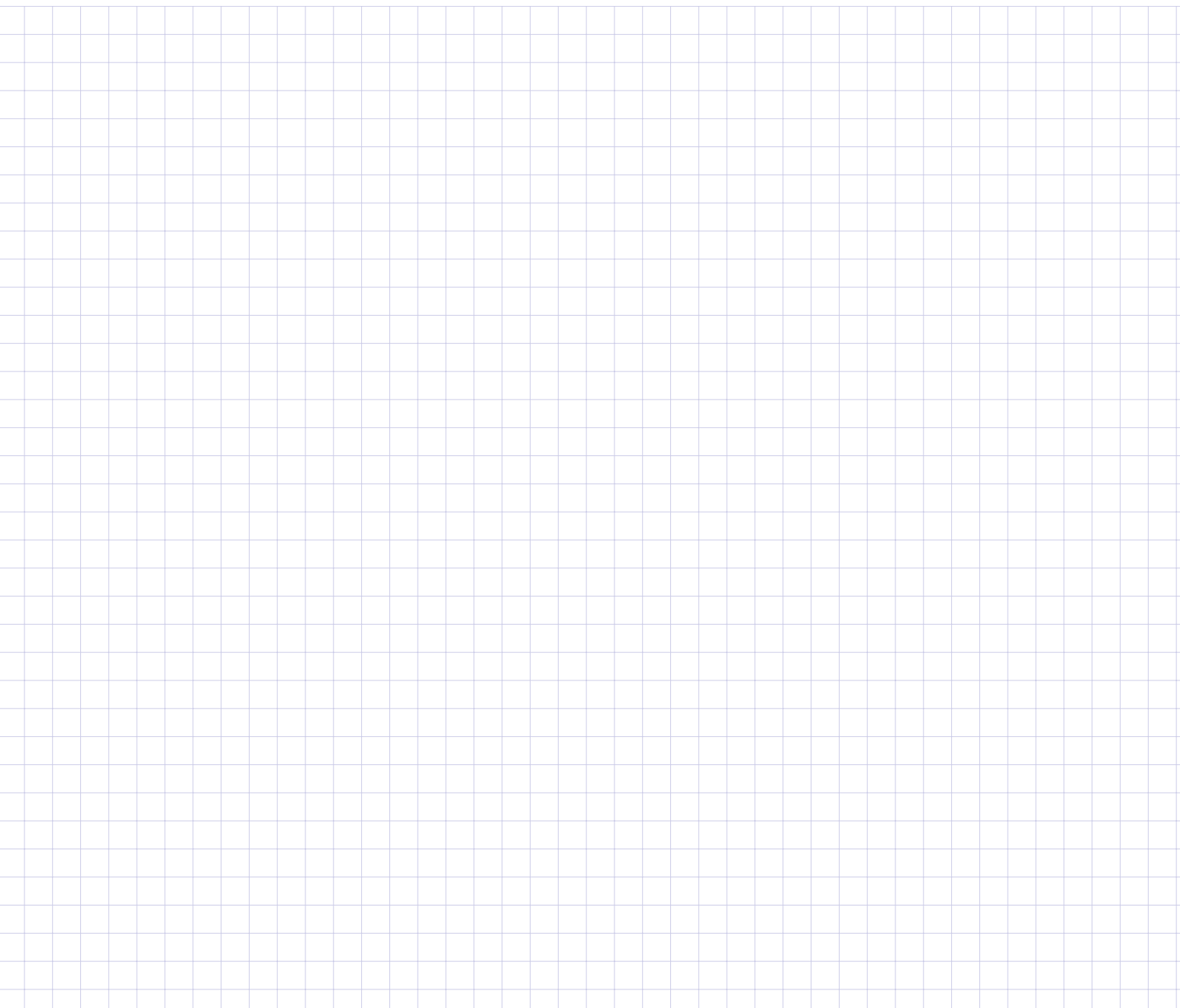
(ii)  $s_z^2 = \frac{1}{n-1} \sum_{i=1}^m (n_i - 1) s_i^2 + \frac{1}{n-1} \sum_{i=1}^m n_i (\bar{x}_i - \bar{z})^2$  (Streuungszerlegung)

(iii) Die Streuungszerlegung, vgl. (ii)), wird oft in der Form:

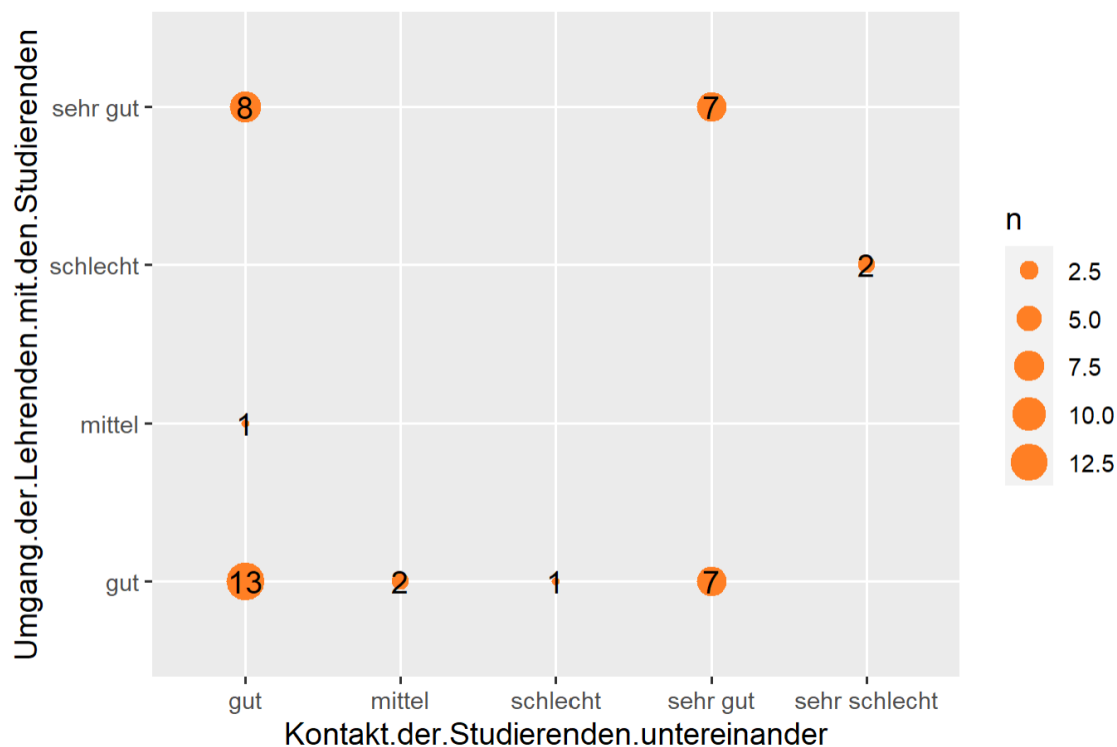
Gesamtstreuung = Streuung in den Schichten + Streuung zwischen den Schichten

ausgedrückt. Interpretieren Sie diese Aussage, z.B. grafisch, anhand der Formel in (ii).

*Hinweis: Ohne Einschränkung der Allgemeinheit kann man davon ausgehen, dass die ersten  $n_1$  Stichprobenwerte  $z_1, \dots, z_{n_1}$  der gemeinsamen Stichprobe  $z$  die Werte der Stichprobe  $x_1$  sind, die folgenden  $n_2$  Werte  $z_{n_1+1}, \dots, z_{n_1+n_2}$  die Werte der Stichprobe  $x_2$  sind usw. bis zu den  $n_m$  Werten  $z_{n_1+\dots+n_{m-1}+1}, \dots, z_n$ , die die Stichprobenwerte von  $x_m$  sind.*



- (i) Lesen Sie die Datei „AbsolventenDat.csv“ wie in Aufgabe 2.4 in den tibble `dat.Absolventen` ein.
- (ii) Wandeln Sie die beiden Merkmale *Kontakt der Studierenden untereinander* und *Umgang der Lehrenden mit den Studierenden* in den Datentyp `factor` um und kodieren Sie die Stufen der beiden Merkmale mit dem Befehl `levels` in Schulnoten (1 bis 5) um.
- (iii) Berechnen Sie mit dem Befehl `table` eine Kontingenztabelle der beiden oben genannten Merkmale.
- (iv) Zeichnen Sie einen Häufigkeitsplot der beiden Merkmale.  
*Hinweis: Verwenden Sie den `geom_count()`-Befehl.*
- (v) Schreiben Sie zusätzlich die absolute Häufigkeit in die Kreise des Häufigkeitsplots.  
Beispielsweise könnte die Grafik so aussehen:



*Hinweis: Sehen Sie sich die Internetseite:*

[www.r-graph-gallery.com/5-correlation-of-discrete-variables](http://www.r-graph-gallery.com/5-correlation-of-discrete-variables)

*an. Auf der Seite [www.r-graph-gallery.com](http://www.r-graph-gallery.com) finden Sie eine Menge schöner Grafiken und den zugehörigen R-code. Außerdem ist [stackoverflow.com](http://stackoverflow.com) immer eine Suche wert.*

## Aufgabe 4.10: Korrelationen 2 (keine Punkte)



Sei  $(x_i, y_i), i = 1, \dots, n$ , eine bivariate Stichprobe zweier metrischer Merkmale mit  $s_x > 0, s_y > 0$  und Pearson-Korrelationskoeffizient  $r_{x,y}$ , wobei  $x = (x_1, \dots, x_n)^T$  und  $y = (y_1, \dots, y_n)^T$ . Zeigen Sie, dass die folgenden Aussagen gelten.

(i)  $-1 \leq r_{x,y} \leq 1$

(ii)  $\forall i = 1, \dots, n : y_i = a + bx_i, a, b \in \mathbb{R}, b \neq 0 \Rightarrow r_{x,y} = \text{sgn}(b)$ , wobei  $\text{sgn} : \mathbb{R} \rightarrow \{-1, 0, 1\}$  die Signum- oder Vorzeichenfunktion mit der Abbildungsvorschrift

$$\text{sgn}(x) = \begin{cases} -1 & \text{für } x < 0 \\ 0 & \text{für } x = 0 \\ 1 & \text{für } x > 0 \end{cases}$$

bezeichnet.

(iii)  $r_{x,y} = \pm 1 \Rightarrow \exists a, b \in \mathbb{R}, \text{sgn}(b) = \text{sgn}(r_{x,y}) : y_i = a + bx_i, \forall i = 1, \dots, n$ .

*Hinweis: Verwenden Sie die Ungleichung von Cauchy-Schwarz (Schwarzsche Ungleichung): Für alle  $u, v \in \mathbb{R}^n, n \in \mathbb{N}$ , gilt, dass*

$$|u \cdot v| \leq \|u\| \|v\|.$$

*Das Gleichheitszeichen gilt genau dann, falls  $u$  und  $v$  linear abhängig sind. Der Malpunkt steht dabei für das Skalarprodukt.*

## Aufgabe 5.1: Mengenlehre (5 Punkte)



Seien  $A$ ,  $B$  und  $C$  Ereignisse in einem Grundraum  $\Omega$ . Beschreiben Sie folgende Aussagen durch  $A$ ,  $B$  und  $C$  sowie geeignete Mengenoperationen:

- (i) Es tritt nur  $A$  ein.
- (ii) Mindestens zwei der Ereignisse treten ein.
- (iii) Genau eines der Ereignisse tritt ein.
- (iv)  $A$  und  $B$  treten ein, aber nicht  $C$ .
- (v) Höchstens zwei Ereignisse treten ein.

Versuchen Sie nun, folgende Ereignisse in Worten zu fassen:

$$(A \cup B) \cap \overline{A \cap B}, \quad A \triangle B := (A \setminus B) \cup (B \setminus A), \quad \overline{A \cap B}, \quad \overline{A \cup B}, \\ A \cap \overline{B \cup C}.$$

Beim „Mäxle“-Spielen<sup>1</sup> mit Ihren Freunden taucht folgendes Problem auf: Ihr Freund behauptet, er habe „65“ gewürfelt, und gibt den Becher an Sie weiter. Sie glauben ihm und würfeln. Wie groß ist die Wahrscheinlichkeit, dass Sie einen Pasch oder Mäxle werfen? Hierbei tritt Mäxle auf, falls ein Würfel „2“ zeigt und der andere „1“. Geben Sie einen geeigneten Grundraum mit Wahrscheinlichkeitsverteilung an, und berechnen Sie die gesuchte Wahrscheinlichkeit!

Welche Annahmen sind vernünftigerweise zu treffen?

---

<sup>1</sup>Bei „Mäxle“ wird gleichzeitig mit zwei Laplace-Würfeln geworfen. Ziel ist es, eine „höhere“ Kombination als der vorherige Spieler zu erreichen. Die genauen Spielregeln von „Mäxle“ können Sie bei mir erfragen. Die Aufgabe ist aber auch ohne Kenntnis der Regeln lösbar.

### Aufgabe 5.3: Glücksspirale (7 Punkte)



Bei der ersten Ziehung der *Glücksspirale* 1971 wurden für die Ermittlung einer 7-stelligen Gewinnzahl aus einer Trommel, die Kugeln mit den Ziffern  $0, 1, \dots, 9$  je 7mal enthält, nacheinander rein zufällig 7 Kugeln ohne Zurücklegen gezogen.

- (i) Geben Sie einen geeigneten Wahrscheinlichkeitsraum an.
- (ii) Geben Sie die Ziehungswahrscheinlichkeit für jede 7-stellige Gewinnzahl  $(a_1, \dots, a_7) \in \{0, \dots, 9\}^7$  an. Welche 7-stelligen Gewinnzahlen hatten hierbei die größte und welche die kleinste Ziehungswahrscheinlichkeit?
- (iii) Bestimmen Sie die Gewinnwahrscheinlichkeit für die Zahl 3 143 643.
- (iv) Wie würden Sie den Ziehungsmodus abändern, um allen Gewinnzahlen die gleiche Ziehungswahrscheinlichkeit zu sichern?

#### Aufgabe 5.4: Ereignissysteme (5 Punkte)



- (i) Sei  $\Omega = \mathbb{R}$  und  $\mathcal{F} = \{M \subseteq \mathbb{R} : M \text{ offen}\}$ . Begründen Sie, dass  $\mathcal{F}$  kein Ereignissystem über  $\mathbb{R}$  ist.
- (ii) Man betrachte das Zufallsexperiment eines 3-fachen Münzwurfs einer Münze mit den Seiten 0 und 1. Bestimmen Sie einen Grundraum  $\Omega$  und ein Ereignissystem  $\mathcal{F}$  dieses Zufallsexperiments, geben Sie exemplarisch zwei Ereignisse an und bestimmen Sie die Mächtigkeiten von  $\Omega$  und  $\mathcal{F}$ .
- (iii) Sei  $\Omega = \{0; 1\}^3$  und  $\mathcal{F} = \mathcal{P}(\Omega)$ . Zeigen oder widerlegen Sie, dass

$$f(\omega) = \frac{1}{7}, \omega \in \Omega$$

eine Zähldichte eines Wahrscheinlichkeitsmaßes  $P$  auf  $\{\Omega, \mathcal{F}\}$  darstellt.

Beweisen Sie Satz II.3.2:

Sei  $(\Omega, \mathcal{F}, P)$  ein W-Raum und  $A, B, A_1, A_2, \dots \in \mathcal{F}$  beliebige Ereignisse. Dann gilt:

- (i)  $P(\emptyset) = 0$ .
- (ii) Monotonie:  $A \subseteq B \Rightarrow P(A) \leq P(B)$ .
- (iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .
- (iv)  $P(A) + P(A^C) = 1$ .
- (v)  $\sigma$ -Subadditivität:  $P\left(\bigcup_{i \geq 1} A_i\right) \leq \sum_{i \geq 1} P(A_i)$ .
- (vi)  $P(A \setminus B) = P(A) - P(A \cap B)$ .

*Hinweis: Beweisen Sie (vi) vor (iii).*

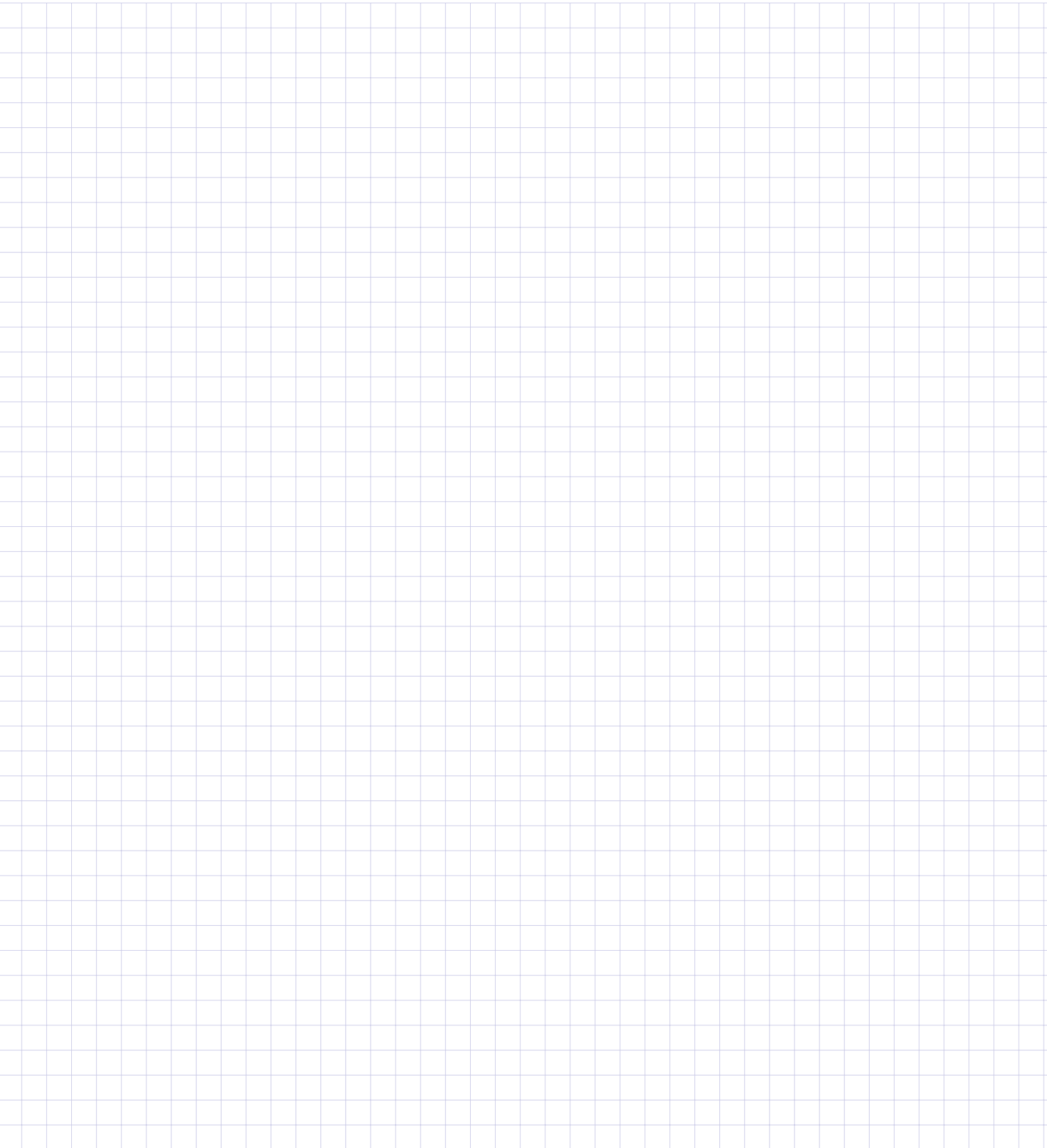
### Aufgabe 5.6: Hypergeometrische Verteilung (R) (3 Punkte)



In Ihrem Dorfweiher leben 200 Forellen und 100 Karpfen. Sie gehen fischen und haben 10 Fische gleichzeitig in Ihrem Netz gefangen, ausschließlich Forellen und Karpfen.

- (i) Wie groß ist die Wahrscheinlichkeit dafür, dass Sie genau 3 Forellen im Netz haben?
- (ii) Wie groß ist die Wahrscheinlichkeit dafür, dass Sie mehr als 7 Karpfen im Netz haben?
- (iii) Wie groß ist die Wahrscheinlichkeit dafür, dass Sie mindestens 9 Fische der gleichen Art im Netz haben?

*Hinweis:* Machen Sie sich mit den R-Befehlen **dhyper** und **phyper** vertraut.



### Aufgabe 5.7: Siebformel (keine Punkte)



Beweisen Sie den Ein- und Ausschlussatz von Sylvester-Poincare (Siebformel):

$\forall n \in \mathbb{N} \forall A_1, \dots, A_n \in \mathcal{F}$ :

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n (-1)^{i-1} \sum_{T \in \left\{ \binom{n}{i} \right\}} P\left(\bigcap_{j \in T} A_j\right)$$

wobei  $\left\{ \binom{n}{i} \right\} := \{T \subseteq \{1, \dots, n\} : \#T = i\}$  „alle  $i$ -elementigen Teilmengen aus  $\{1, \dots, n\}$ “ sind.  
(Beachte  $\# \left\{ \binom{n}{i} \right\} = \binom{n}{i}$ )

*Hinweis: Folgern Sie die Aussage aus Satz II.3.2 mithilfe der vollständigen Induktion.*

## Aufgabe 6.1: Mengenfolgen (7 Punkte)



Seien  $\Omega$  eine nichtleere Menge und  $\{A_n\}_{n \in \mathbb{N}}$  eine Folge in  $\mathcal{P}(\Omega)$ . Dann definiert man den *Limes inferior* und den *Limes superior* der Mengenfolge  $\{A_n\}_{n \in \mathbb{N}}$  als

$$\liminf_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \quad \text{bzw.} \quad \limsup_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Zeigen Sie:

- (i)  $\liminf_{n \rightarrow \infty} A_n = \{\omega \in \Omega \mid \omega \in A_n \text{ für schließlich alle } n \in \mathbb{N}\}$
- (ii)  $\limsup_{n \rightarrow \infty} A_n = \{\omega \in \Omega \mid \omega \in A_n \text{ für unendlich viele } n \in \mathbb{N}\}$

Berechnen Sie nun für die Beispiele

- (iii)  $A_n := (-\frac{1}{n}, \frac{1}{n})$ ,
- (iv)  $A_n := \{(-1)^n\}$ ,
- (v) 
$$\begin{aligned} A_n &:= \{i \in \mathbb{N} : i \text{ ist durch drei teilbar}\} && \text{für } n \text{ ungerade,} \\ A_n &:= \{i \in \mathbb{N} : i \text{ ist durch zwei teilbar}\} && \text{für } n \text{ gerade} \end{aligned}$$

den *Limes inferior* und den *Limes superior* der Mengenfolge  $\{A_n\}_{n \in \mathbb{N}}$ .

Für jede Menge  $C \in \mathcal{P}(\Omega)$  definiert man die *Indikatorfunktion von C* als

$$\mathbb{1}_C : \Omega \rightarrow \{0,1\}, \quad \omega \mapsto \begin{cases} 1 & \text{für } \omega \in C \\ 0 & \text{für } \omega \notin C \end{cases}.$$

Beweisen Sie:

- (vi)  $\mathbb{1}_{\liminf_{n \rightarrow \infty} A_n} = \liminf_{n \rightarrow \infty} \mathbb{1}_{A_n},$

Bemerkung: Auch die Aussage  $\mathbb{1}_{\limsup_{n \rightarrow \infty} A_n} = \limsup_{n \rightarrow \infty} \mathbb{1}_{A_n}$  gilt und kann analog zu (vi) bewiesen werden.

### Aufgabe 6.2: Eigenschaften von Ereignissystemen (6 Punkte)



Sei  $\mathcal{F}$  eine  $\sigma$ -Algebra über einer Menge  $\Omega$ ,  $A, B \in \mathcal{F}$  und  $\{A_n\}_{n \in \mathbb{N}}$  eine Mengenfolge in  $\mathcal{F}$ . Zeigen Sie:

- (i)  $\emptyset \in \mathcal{F}$ ,
- (ii)  $A \cup B \in \mathcal{F}$  und  $A \cap B \in \mathcal{F}$ ,
- (iii)  $A \setminus B = A \cap B^C \in \mathcal{F}$ ,  $A \triangle B := A \setminus B \cup B \setminus A \in \mathcal{F}$   
( $A_1 \triangle A_2$  heißt „symmetrische Differenz“.),
- (iv)  $\bigcap_{i \geq 1} A_i \in \mathcal{F}$  (Durchschnittsstabilität abzählbar vieler Mengen).
- (v)  $\limsup_{n \rightarrow \infty} A_n \in \mathcal{F}$ ,
- (vi)  $\liminf_{n \rightarrow \infty} A_n \in \mathcal{F}$ .

*Hinweis: Die Teile (i) bis (iv) beweisen Lemma II.1.*

### Aufgabe 6.3: Eigenschaften von Ereignissystemen (2 Punkte)



Seien  $\mathcal{E}_1$  und  $\mathcal{E}_2$  Mengensysteme über einer Menge  $\Omega$ .

Beweisen Sie:

$$\mathcal{E}_1 \subseteq \mathcal{E}_2 \Rightarrow \sigma(\mathcal{E}_1) \subseteq \sigma(\mathcal{E}_2).$$

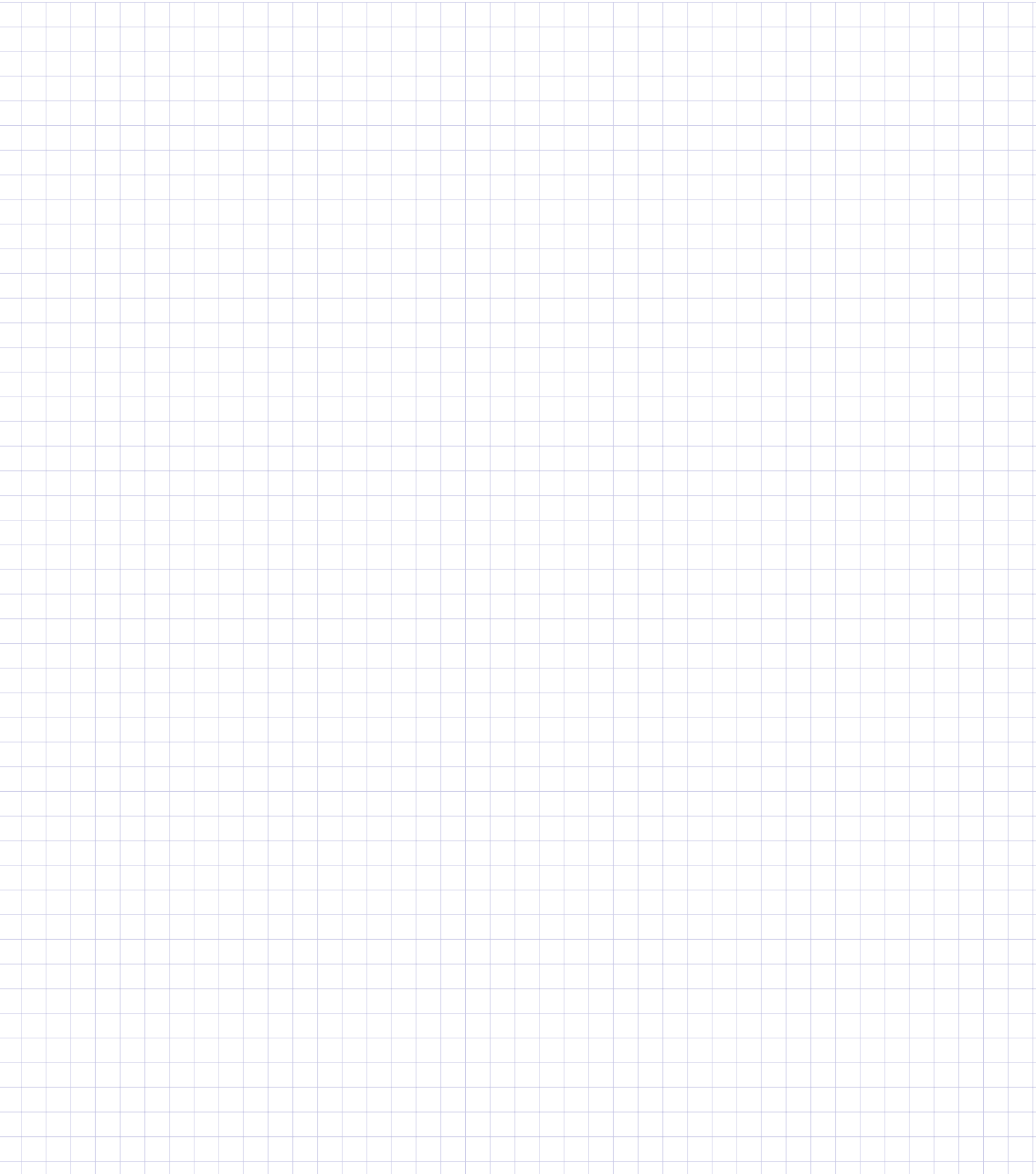
#### Aufgabe 6.4: Klassisches Maßproblem (3 Punkte)



In der Vorlesung wurde erwähnt, dass es auf  $\mathcal{P}([0,1])$  keine nichtnegative, normierte, sigma-additive und kongruenztreue Abbildung gibt.

- (i) Existiert eine nichtnegative, normierte und sigma-additive Abbildung auf  $\mathcal{P}([0,1])$ ?
- (ii) Existiert eine nichtnegative, sigma-additive und kongruenztreue Abbildung auf  $\mathcal{P}([0,1])$ ?

Begründen Sie Ihre Antworten!



## Aufgabe 7.1: Ringsysteme 2 (6 Punkte)



Sei  $X$  eine abzählbar unendliche Menge. Zeigen Sie:

- (i)  $\mathcal{A} := \{A \subseteq X : A \text{ endlich oder } A^C \text{ endlich}\}$  ist ein Ringsystem.

*Hinweis: Wenn Ihnen der Beweis für ein allgemeine Menge  $X$  zu schwierig ist, zeigen Sie die Aussage bitte für den Spezialfall  $X = \mathbb{N}$ .*

- (ii) Die Abbildung

$$\begin{aligned} \mu : \mathcal{A} &\rightarrow \bar{\mathbb{R}} \\ A &\mapsto \begin{cases} 0, & \text{falls } A \text{ endlich ist.} \\ \infty, & \text{falls } A^c \text{ endlich ist.} \end{cases} \end{aligned}$$

ist ein Inhalt, aber kein Prämaß.

*Hinweis: Zeigen Sie zunächst, dass  $\mu$  wohldefiniert ist, d.h. dass es sich bei  $\mu$  auch wirklich um eine Abbildung handelt, die jedem  $A \in \mathcal{A}$  nur einen Wert zuordnet.*

Lernen Sie Lesen und Schreiben des griechischen Alphabets:

| Groß      | Klein                   | Name    | Groß     | Klein               | Name    |
|-----------|-------------------------|---------|----------|---------------------|---------|
| A         | $\alpha$                | Alpha   | N        | $\nu$               | Ny      |
| B         | $\beta$                 | Beta    | $\Xi$    | $\xi$               | Xi      |
| $\Gamma$  | $\gamma$                | Gamma   | O        | o                   | Omikron |
| $\Delta$  | $\delta$                | Delta   | $\Pi$    | $\pi, \varpi$       | Pi      |
| E         | $\epsilon, \varepsilon$ | Epsilon | P        | $\rho, \varrho$     | Rho     |
| Z         | $\zeta$                 | Zeta    | $\Sigma$ | $\sigma, \varsigma$ | Sigma   |
| H         | $\eta$                  | Eta     | T        | $\tau$              | Tau     |
| $\Theta$  | $\theta, \vartheta$     | Theta   | Y        | $\upsilon$          | Ypsilon |
| I         | $\iota$                 | Iota    | $\Phi$   | $\phi, \varphi$     | Phi     |
| K         | $\kappa, \varkappa$     | Kappa   | X        | $\chi$              | Chi     |
| $\Lambda$ | $\lambda$               | Lambda  | $\Psi$   | $\psi$              | Psi     |
| M         | $\mu$                   | My      | $\Omega$ | $\omega$            | Omega   |

### Aufgabe 7.3: Stetigkeitssatz (9 Punkte)



Sei  $\mathcal{R}$  ein Ringsystem und  $\mu : \mathcal{R} \rightarrow \bar{\mathbb{R}}$  ein Inhalt.

Beweisen Sie:

- a)  $\mu$  ist Prämaß.
- $\Updownarrow$
- b)  $\forall A_1, A_2, \dots \in \mathcal{R}, \text{isoton} : \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{R} \Rightarrow \lim_{n \rightarrow \infty} \mu(A_n) = \mu \left( \bigcup_{n \in \mathbb{N}} A_n \right).$
- $\Downarrow$
- c)  $\forall A_1, A_2, \dots \in \mathcal{R}, \text{antiton} : \bigcap_{n \in \mathbb{N}} A_n \in \mathcal{R} \text{ und } \mu(A_1) < \infty \Rightarrow \lim_{n \rightarrow \infty} \mu(A_n) = \mu \left( \bigcap_{n \in \mathbb{N}} A_n \right).$
- $\Updownarrow$
- d)  $\forall A_1, A_2, \dots \in \mathcal{R}, \text{antiton} : \bigcap_{n \in \mathbb{N}} A_n = \emptyset \text{ und } \mu(A_1) < \infty \Rightarrow \lim_{n \rightarrow \infty} \mu(A_n) = 0.$

*Hinweis: Zeigen Sie zunächst (a)  $\Leftrightarrow$  (b), dann (a)  $\Rightarrow$  (c) und schließlich (c)  $\Leftrightarrow$  (d).*

#### Aufgabe 7.4: Vereinigung von Ereignissystemen (2 Punkte)



Man zeige durch Gegenbeispiel, dass im Allgemeinen die Vereinigung  $\mathcal{F} \cup \mathcal{F}'$  zweier  $\sigma$ -Algebren (über demselben  $\Omega$ ) keine  $\sigma$ -Algebra ist.

## Aufgabe 8.1: Stochastische Unabhängigkeit (4 Punkte)



Seien  $(\Omega, \mathcal{F}, P)$  ein Wahrscheinlichkeitsraum und  $A_1, \dots, A_n$  ( $n \in \mathbb{N}$ ) Ereignisse. Zeigen oder widerlegen Sie:

- (i) Wenn  $A_1, \dots, A_n$  paarweise stochastisch unabhängig sind, dann sind  $A_1, \dots, A_n$  stochastisch unabhängig.
- (ii)  $P(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$  impliziert stochastische Unabhängigkeit von  $A_1, \dots, A_n$ .

### Aufgabe 8.2: Coupon collector's problem (Teil (iii): R) (7 Punkte)



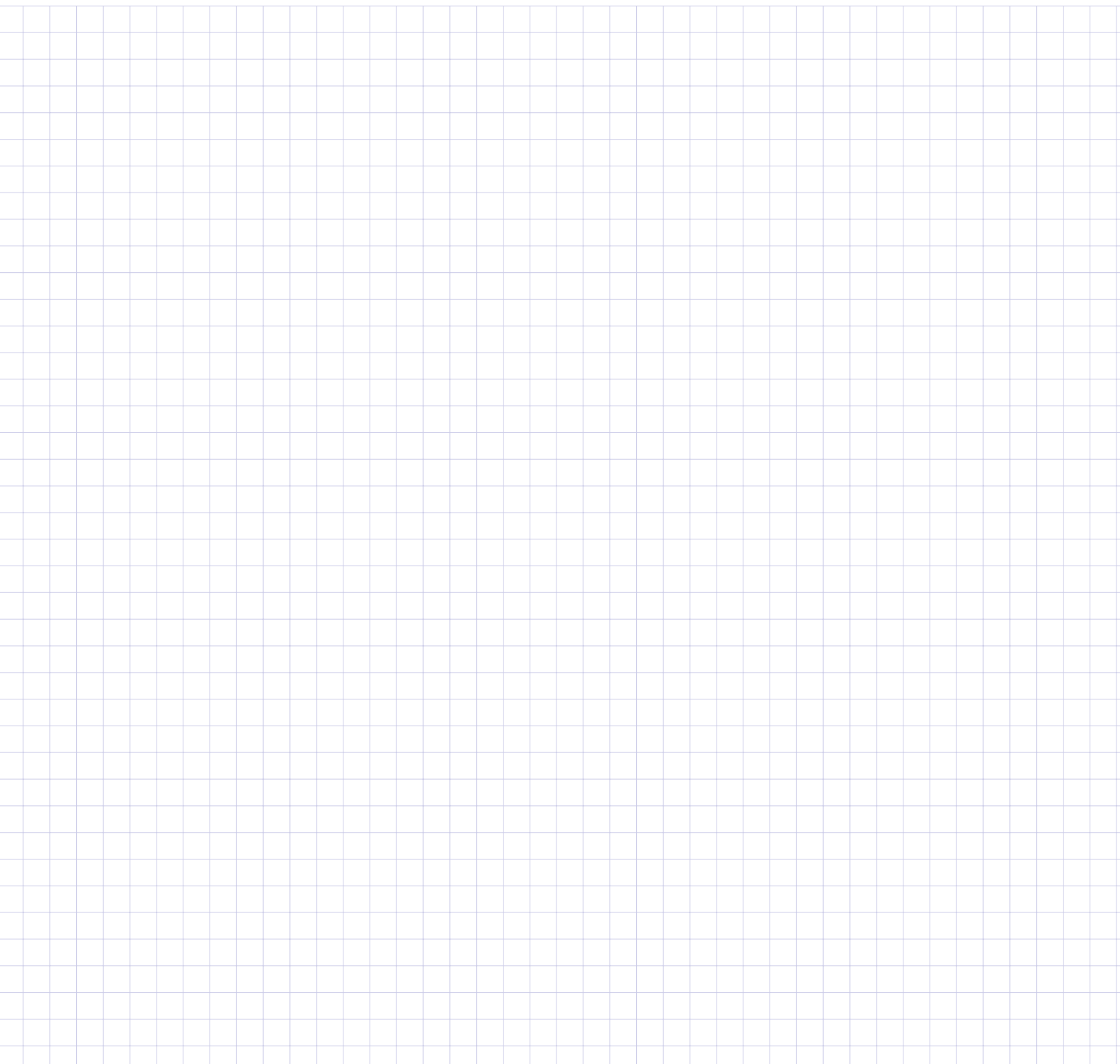
In jedem Schokoriegel einer bekannten Firma ist ein Aufkleber enthalten. Wie viele Schokoriegel muss man kaufen, um mit 95%-iger Wahrscheinlichkeit eine komplette Sammlung zu erhalten, wenn es 30 verschiedene Aufkleber gibt?

- (i) Geben Sie für den Kauf von  $n$  ( $n \in \mathbb{N}$ ) Aufklebern einen geeigneten Wahrscheinlichkeitsraum an. Welche Annahmen sind vernünftigerweise zu treffen?
- (ii) Berechnen Sie in dem in (i) gefundenen Modell die Wahrscheinlichkeit dafür, dass eine zufällige Sammlung von  $n$  ( $n \in \mathbb{N}$ ) Aufklebern komplett ist!

*Hinweis: Betrachten Sie die Ereignisse „Sammlung enthält Bild  $i$  ( $i \in \{1, \dots, 30\}$ )“, bilden Sie den Schnitt über diese Ereignisse und gehen Sie auf das Komplement über. Verwenden Sie dann den Ein- und Ausschlussatz von Sylvester und Poincaré.*

- (iii) Lösen Sie obiges Problem!

*Hinweis: Hier kann ein Computerprogramm wie EXCEL (oder noch viel besser: R) hilfreich sein.*



Eine bestimmte Krebsart kommt bei ca. 0,5 % der Bevölkerung vor. Ein Test zur Auffindung der Krebserkrankung führt bei 99 % der Kranken zu einer Reaktion, aber auch bei 2 % der Gesunden.

- (i) Wie groß ist die Wahrscheinlichkeit, dass eine Person, bei der die Reaktion eintritt, die Krebserkrankung wirklich hat?
- (ii) Wie groß sind die Sensitivität und die Spezifität des Tests?

*Hinweis: Die Definitionen der beiden Begriffe finden Sie im Internet oder im Buch, z.B.: „Sachs, L. und Hedderich, J. (2016): Angewandte Statistik, Methodensammlung mit R. Berlin, Heidelberg: Springer Verlag.“, das über die Bibliothek als E-Book ausleihbar ist.*

- (iii) Weisen Sie allgemein nach, dass die Chancen (engl.: odds) für eine Erkrankung bei positivem Test um den Faktor

$$\frac{\text{Sensitivität}}{1 - \text{Spezifität}} = \frac{\text{richtig-positiv Rate}}{\text{falsch-positiv Rate}}$$

größer sind als vor dem Test.

*Hinweis: Mit „odds“ bezeichnet man das Verhältnis zwischen der Wahrscheinlichkeit für „krank“ (=„Treffer“) und der Wahrscheinlichkeit für „gesund“ (=„kein Treffer“). Der Begriff wird auch bei Sportwetten verwendet.*

- (iv) Prüfen Sie die Gültigkeit der Formel aus Teil (iii) anhand der Zahlen aus der Aufgabe und dem Ergebnis aus Teil (i).

#### Aufgabe 8.4: Formel von Bayes (4 Punkte)

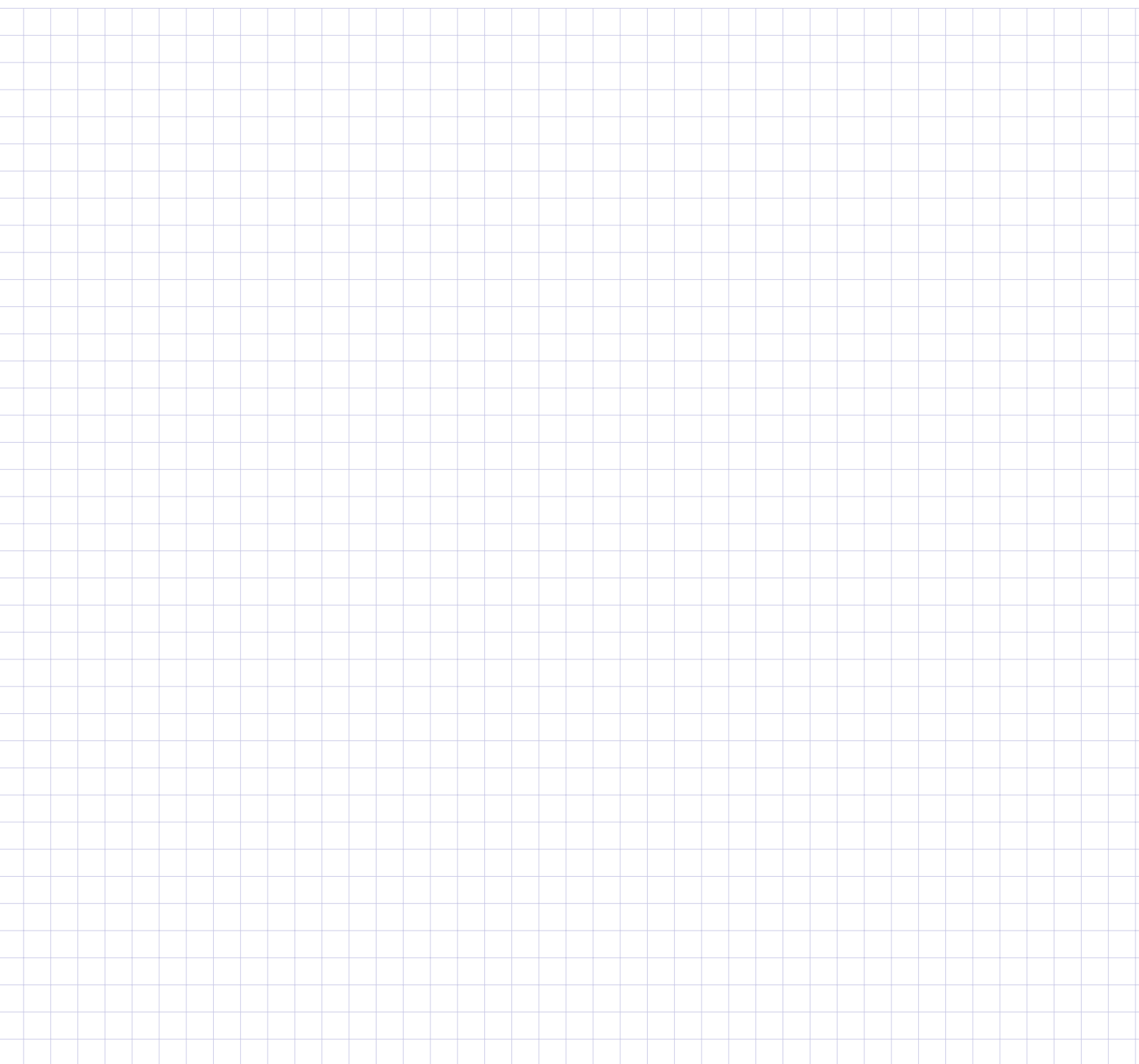


In der Augsburger Straßenbahn sind 10% der Fahrgäste Schwarzfahrer. 70% der Schwarzfahrer haben keine Fahrkarte, während die anderen Schwarzfahrer gefälschte oder illegal besorgte Karten besitzen. Von den ehrlichen Fahrgästen haben 5% ihre Fahrkarte vergessen. In der Aufgabe soll nun mit zwei verschiedenen Methoden die Wahrscheinlichkeit dafür, dass ein kontrollierter Fahrgast, der keine Karte vorzeigen kann, ein Schwarzfahrer ist, berechnet werden.

- (i) Lösen Sie die Aufgabe „sauber“ mit einem geeigneten Wahrscheinlichkeitsraum?
- (ii) Lösen Sie die Aufgabe - wie vom Psychologen Gigerenzer vorgeschlagen - mit einer Vierfeldertafel und absoluten Zahlen.

*Hinweis: Stellen Sie sich dazu 1000 Fahrgäste vor und überlegen Sie, wie viele Schwarzfahrer mit bzw. ohne Fahrkarte es gibt und wie viele ehrliche Kunden mit bzw. ohne Fahrkarte. Setzen Sie die geeigneten Zahlen ins Verhältnis, um die gesuchten Wahrscheinlichkeiten zu berechnen. Am übersichtlichsten geht das in einer Vierfeldertafel.*

- (iii) Wann stößt die Methode von Herrn Gigerenzer an ihre Grenzen?



## Aufgabe 9.1: Poisson-Verteilung (4 Punkte)



Die sog. Poisson-Verteilung (nach Siméon-Denis Poisson, 1781-1840) ist für  $\lambda > 0$  durch die Zähldichte

$$P_\lambda(\{k\}) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{für } k \in \mathbb{N}_0$$

gegeben. Sie ist ein natürliches Modell für die Anzahl von rein zufälligen Ereignissen, die in einem Zeitintervall auftreten, wenn im Schnitt in diesem Zeitintervall  $\lambda$  Ereignisse eintreten.

- (i) Zeigen Sie, dass  $f : \mathbb{N}_0 \rightarrow \mathbb{R}$  mit  $f(k) := P_\lambda(\{k\})$  eine Zähldichte ist.

*Hinweis: Zur Definition der Zähldichte siehe Satz II.3.1*

- (ii) Die Binomialverteilung kann für große  $n$  und kleine  $p$  durch die Poissonverteilung approximiert werden, d.h.:

$$B_{n,p}(\{k\}) \approx P_\lambda(k) \quad \text{für } \lambda := np, \quad n \text{ groß, } p \text{ klein.}$$

(grobe Merkregel:  $n > 10; p < \frac{1}{10}$ ).

Rechtfertigen Sie diese Approximation, in dem Sie folgende Grenzwertbeziehung zeigen:

Sei  $p_n$  eine Folge mit  $\lim_{n \rightarrow \infty} np_n = \lambda$ . Dann gilt:

$$\forall k \in \mathbb{N}_0 : \lim_{n \rightarrow \infty} B_{n,p_n}(\{k\}) = P_\lambda(\{k\})$$

*Hinweis: Verwenden Sie die aus der Analysis bekannte Beziehung:  $\lim_{n \rightarrow \infty} (1 + \frac{x_n}{n})^n = e^x$ , falls  $\lim_{n \rightarrow \infty} x_n = x$ .*

- (iii) In einer Telefonzentrale rufen im Schnitt fünf Personen je Minute an. Wie groß ist die Wahrscheinlichkeit, dass in einer Minute gar keiner anruft. Und wie groß ist die Wahrscheinlichkeit, dass mehr als sechs Personen anrufen?

## Aufgabe 9.2: Faltung (5 Punkte)



Sei  $(\Omega, \mathcal{F}, P)$  ein diskreter Wahrscheinlichkeitsraum.

- (i) Die Zufallsvariablen  $X, Y : \Omega \rightarrow \mathbb{N}_0$  seien stochastisch unabhängig. Die Verteilungen  $P^X$  und  $P^Y$  besitzen die Zähldichten  $f$  und  $g$ . Bestimmen Sie die Verteilung der Zufallsvariablen  $Z := X + Y$ .
- (ii) Seien  $X, Y$  stochastisch unabhängige Zufallsvariablen, wobei  $X$   $P_\lambda$ -verteilt und  $Y$   $P_\mu$ -verteilt sei. Zeigen Sie unter Verwendung von (i), dass  $Z := X + Y$   $P_{\lambda+\mu}$ -verteilt ist.
- (iii) Seien  $X_1, \dots, X_n$  ( $n \in \mathbb{N}$ ) u.i.  $P_\lambda$ -verteilte Zufallsvariablen. Bestimmen Sie unter Verwendung von (ii) die Zähldichte von  $Z := \sum_{i=1}^n X_i$ .

- (i) Erzeugen Sie einen Vektor mit zehntausend  $N_{170,100}$ -verteilten Zufallszahlen.

*Hinweis:* Normalverteilte Zufallszahlen können in R über die Funktion `rnorm` erzeugt werden.

- (ii) Wie groß ist der Anteil der Komponenten, die kleiner als 160 sind?
- (iii) Wie groß ist der Anteil der Komponenten, die zwischen 140 und 200 liegen?
- (iv) Welchen Wert hat die größte Komponente?
- (v) An welcher Indexposition tritt der größte Wert auf?

*Hinweis:* Verwenden Sie die Funktion `which`.

- (vi) Wie groß ist der Anteil der Komponenten, deren betragsmäßiger Abstand zur nächsten Komponente größer als  $\sqrt{200} \approx 14,14$  ist?
- (vii) Ermitteln Sie die Indizes der Komponenten mit den zehn größten absoluten Abständen zum Nachfolger.

#### Aufgabe 9.4: Dichteplots I mit R (6 Punkte)



Aus Satz II.3.1 kennen Sie den Begriff der Zähldichte. Verwenden Sie das Statistikpaket R, um die Zähldichten folgender Verteilungen zu plotten:

- (i)  $H_{60;40;10}$
- (ii)  $B_{10;0.5}$
- (iii)  $U_{\{1;2;3;4;5;6\}}$

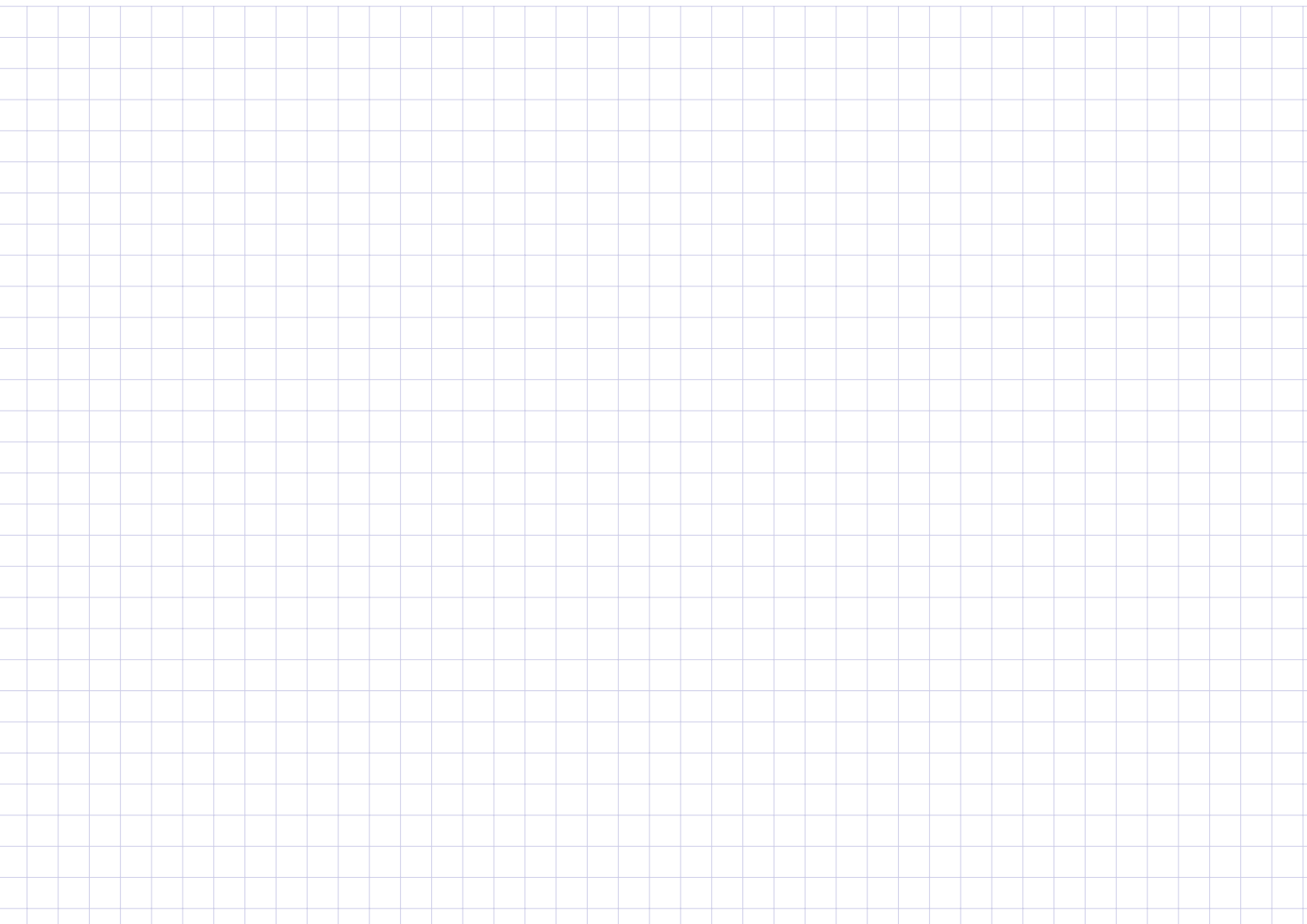
Verwenden Sie diese Plots, um grafisch (!) Näherungswerte für die folgende Wahrscheinlichkeiten zu bestimmen:

- (i)  $H_{60;40;10}(\{3; 4; 5\})$
- (ii)  $B_{10;0.5}(\{1; 3; 7; 9\})$
- (iii)  $U_{\{1;2;3;4;5;6\}}(\{2; 6\})$

*Hinweise:*

- Bei allen R-Aufgaben sollte Ihre Lösung aus einem kommentierten Ausdruck der Ergebnisse und einer ausführlichen Beschreibung Ihrer Vorgehensweise bestehen.
- Als Vorlage für Ihren R-Code kann folgendes Beispiel dienen:  
Einen Plot der Zähldichte von der Bernoulliverteilung mit  $p = 0,1$  erhält man mit folgenden beiden R-Kommandos:

```
x<-c(0,1) # Menge, auf der die Vtlg. eine positive Wsk. hat
plot(x,dbinom(x,1,0.1), xlim = c(0, 1), ylim = c(0, 0.9),
 font.axis=2, cex.axis=1.5, las=1, lwd=2,col="black",
 xlab=, ylab=,main="B 1;0.1")
```



## Aufgabe 10.1: Messbarkeit (5 Punkte)



Seien  $(\Omega, \mathcal{F})$  ein Messraum, d.h.  $\mathcal{F}$  ist ein Ereignissystem auf  $\Omega$ , und  $f : \Omega \rightarrow \bar{\mathbb{R}}$  eine Abbildung. Zeigen Sie die Äquivalenz der folgenden Aussagen.

- (i)  $f$  ist  $\mathcal{F} - \bar{\mathbb{B}}$ -messbar.
- (ii)  $\{f > a\} \in \mathcal{F} \quad \forall a \in \mathbb{R}.$
- (iii)  $\{f \geq a\} \in \mathcal{F} \quad \forall a \in \mathbb{R}.$
- (iv)  $\{f < a\} \in \mathcal{F} \quad \forall a \in \mathbb{R}.$
- (v)  $\{f \leq a\} \in \mathcal{F} \quad \forall a \in \mathbb{R}.$

Zeigen Sie ferner für den Spezialfall  $\mathcal{F} = \sigma(\{\Omega\})$ : (i)  $\Leftrightarrow f \equiv a$  für ein  $a \in \bar{\mathbb{R}}$ .

*Hinweis:* Verwenden Sie, dass

$$\bar{\mathbb{B}} := \{B, B \cup \{+\infty\}, B \cup \{-\infty\}, B \cup \{-\infty, \infty\} \mid B \in \mathbb{B}\}$$

eine  $\sigma$ -Algebra ist, die vom Mengensystem  $\mathcal{D} := \{[-\infty, a] : a \in \mathbb{R}\}$  erzeugt wird.

Zeigen Sie zunächst (i)  $\Leftrightarrow$  (v), in dem Sie Lemma A aus Kapitel VIII.7 verwenden. Zeigen Sie anschließend den Ringschluss: (ii)  $\Rightarrow$  (iii)  $\Rightarrow$  (iv)  $\Rightarrow$  (v)  $\Rightarrow$  (ii).

### Aufgabe 10.2: Messbarkeit II (3 Punkte)



Seien  $(\Omega, \mathcal{F})$  ein Messraum und  $f_i : \Omega \rightarrow \mathbb{R}$ ,  $i \in \mathbb{N}_0$  eine Folge  $\mathcal{F} - \bar{\mathbb{B}}$ -messbarer Abbildungen. Zudem sei  $N : \Omega \rightarrow \mathbb{N}_0$  eine  $\mathcal{F} - \mathcal{P}(\mathbb{N}_0)$ -messbare Abbildung. Zeigen Sie, dass

$$S(\omega) := \sum_{i=0}^{N(\omega)} f_i(\omega) \quad \text{für alle } \omega \in \Omega$$

eine  $\mathcal{F} - \bar{\mathbb{B}}$ -messbare Abbildung auf  $\Omega$  ist.

*Hinweis: Verwenden Sie die Äquivalenz (i)  $\Leftrightarrow$  (v) aus Aufgabe 8.1. und zerlegen Sie die auftretende Menge nach den Werten von  $N$ .*

### Aufgabe 10.3: Extremwertverteilungen (4 Punkte)



Seien  $X_1, \dots, X_n$  ( $n \in \mathbb{N}$ ) unabhängige, reellwertige Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{F}, P)$  mit zugehörigen Verteilungsfunktionen  $F_1, \dots, F_n$ . Man stelle die Verteilungsfunktionen von

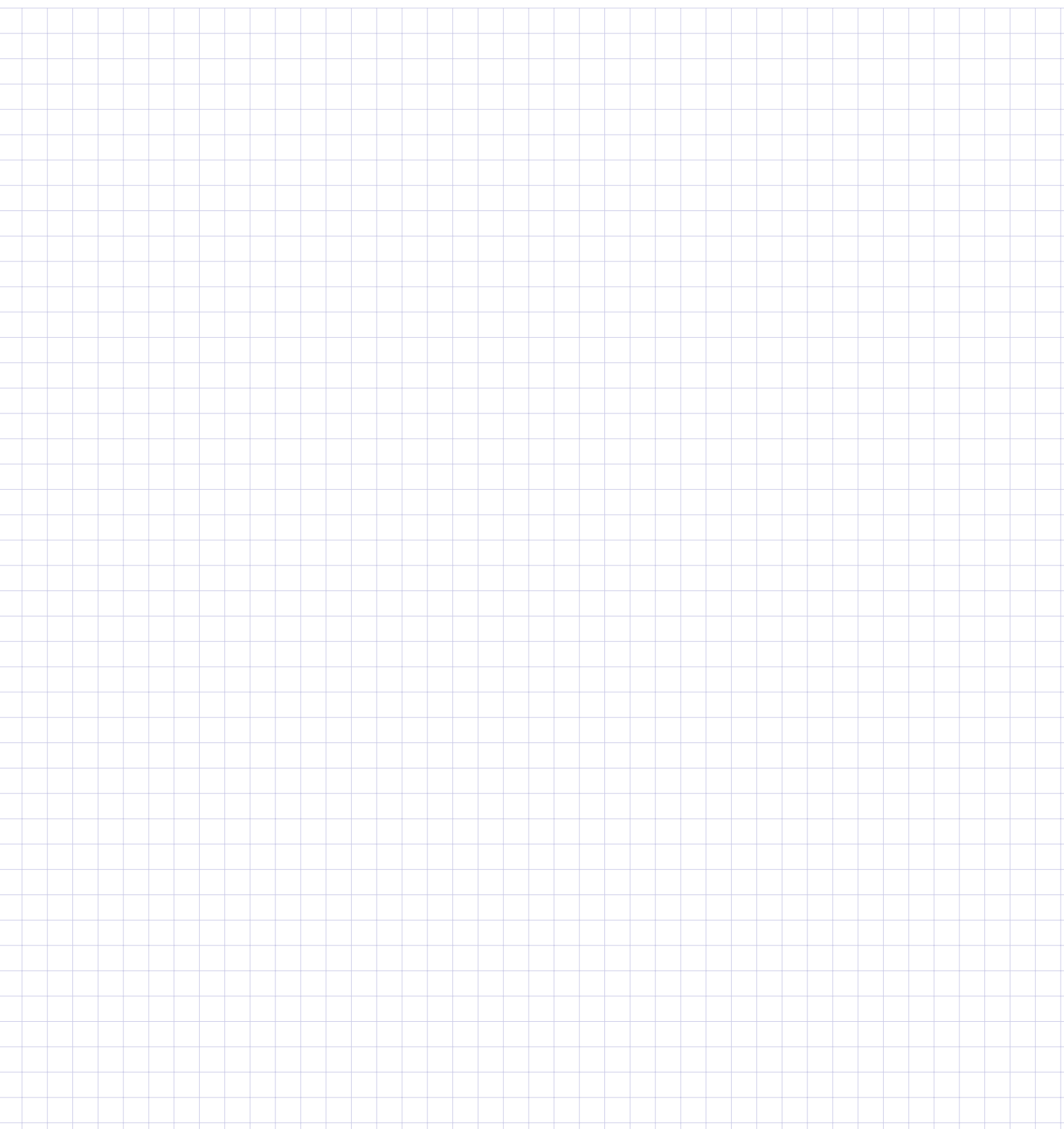
(i)  $\max(X_1, \dots, X_n)$  und

(ii)  $\min(X_1, \dots, X_n)$

durch die Verteilungsfunktionen  $F_1, \dots, F_n$  dar.

Betrachten Sie nun unabhängige, exponential-verteilte Zufallsvariablen, d.h.

$F_i(t) = (1 - e^{-\lambda_i t}) \mathbb{1}_{[0, \infty)}(t) \forall t \in \mathbb{R} \quad (\lambda_i \geq 0)$ . Wie ist  $\min(X_1, \dots, X_n)$  verteilt?



#### Aufgabe 10.4: Monte-Carlo-Methode zur Schätzung von $\pi$ (R) (4 Punkte)



Erzeugen Sie mit R zweimal je  $n$  ( $n \in \mathbb{N}$ )  $U_{[0,1]}$ -verteilte Zufallszahlen (also  $n$  zufällige Punkte in der Menge  $[0; 1] \times [0; 1]$ ) und schätzen Sie mit diesen Zahlen die Zahl  $\pi$ . Wie hängt die „Genauigkeit“ der Schätzung von der Anzahl  $n$  ab?

*Hinweis: Betrachten Sie einen geeigneten Viertelkreis im Quadrat  $[0; 1] \times [0; 1]$  und zählen Sie, wie viele der zufälligen Punkte im Viertelkreis landen. Das Verhältnis der Treffer zur Gesamtzahl aller Punkte hat etwas mit  $\pi$  zu tun.*



Aufgabe 10.5: Positiv-/Negativteil (1 Punkte)



Zerlegen Sie die Funktion  $f : \mathbb{R} \rightarrow \mathbb{R}; x \mapsto (x - 3)^3$  in Positiv- und Negativteil.

### Aufgabe 10.6: Lebesgue vs. Riemann (3 Punkte)



In einer Kirchengemeinde wirft jede Besucherin nach der Messe genau ein Geldstück in einen Sammelkorb. Mit der Funktion:

$$f : \{1; 2; \dots; 20\} \rightarrow \{0,5; 1; 2\}$$

wird chronologisch festgehalten, wie viel die  $i$ -te Besucherin gegeben hat. Es ergaben sich folgende Werte:

|        |   |     |   |   |   |     |   |   |   |    |    |    |    |    |     |    |    |     |    |    |
|--------|---|-----|---|---|---|-----|---|---|---|----|----|----|----|----|-----|----|----|-----|----|----|
| $i$    | 1 | 2   | 3 | 4 | 5 | 6   | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15  | 16 | 17 | 18  | 19 | 20 |
| $f(i)$ | 2 | 0,5 | 1 | 2 | 2 | 0,5 | 1 | 2 | 1 | 1  | 1  | 2  | 1  | 1  | 0,5 | 2  | 1  | 0,5 | 2  | 2  |

Nun soll die Gesamtsumme der Spenden, das ist  $\int_{\{1;2;\dots;20\}} f(n) d\zeta_{\mathbb{N}}(n)$  berechnet werden.

Beschreiben Sie, wie man dieses Integral nach Riemann bzw. Lebesgue berechnet. Warum ist die Methode von Lebesgue geschickter?



## Aufgabe 11.1: Satz von der monotonen Konvergenz (4 Punkte)



Sei  $\mathbb{Q} = \{q_1, q_2, q_3, \dots\}$  eine Abzählung der rationalen Zahlen. Wir betrachten die Funktionenfolge

$$f_n : [0,1] \rightarrow \mathbb{R}, \quad n \in \mathbb{N}, \quad x \mapsto \begin{cases} 1 & \text{für } x \in \{q_1, \dots, q_n\}, \\ 0 & \text{für } x \notin \{q_1, \dots, q_n\}. \end{cases}$$

Zeigen Sie:

(i) Für das Lebesgue-Integral gilt:

$$\lim_{n \rightarrow \infty} \int f_n d\lambda = \int \lim_{n \rightarrow \infty} f_n d\lambda,$$

(ii) für das Riemann-Integral erhält man dagegen:

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx = 0, \quad \text{aber:} \quad \int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx \quad \text{existiert nicht.}$$

*Bemerkung: Dieses Beispiel zeigt, dass die Konvergenzsätze für das Riemann-Integral nicht anwendbar sind. Tatsächlich wird in den entsprechenden Aussagen der Analysis gleichmäßige Konvergenz vorausgesetzt.*

### Aufgabe 11.2: Kettenregel (5 Punkte)



Sei  $(\Omega, \mathcal{F})$  ein Messraum und  $\mu, \nu$  zwei sigmaendliche Maße auf  $\mathcal{F}$ , wobei  $\nu$  eine  $\mu$ -Dichte  $f$  besitze. Zeigen Sie, dass für alle  $\nu$ -integrierbaren Funktionen  $g : \Omega \rightarrow \mathbb{R}$  gilt:

$$\int g \, d\nu = \int g f \, d\mu \quad (\text{Kettenregel}).$$

*Hinweis: Ochsentour.*



Aufgabe 11.3: Fast überall verschwindende Funktionen (6 Punkte)



Beweisen Sie Satz VIII.19 (b), (c) und (d) aus der Vorlesung.



Berechnen Sie folgende Integrale:

(i)  $\int_{\mathbb{R}} e^{x^2} \cos(7x) d\delta_{\{\pi\}}$ , wobei  $\delta_{\{\pi\}}$  das Einpunktmaß in  $\pi$  ist.

(ii)  $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} d\lambda(x)$

(iii)  $\int_{\mathbb{R}} \frac{1}{3} x^2 dU_{[2;4]}(x)$

(iv)  $\int_{\mathbb{R}} e^{-x^6} \sin(7x) d(\delta_{\{\pi\}} + \delta_{\{\frac{\pi}{2}\}})$

(v)  $\int_{-1}^5 e^{\frac{x^2}{2}} x^3 dN_{0;1}(x)$

## Aufgabe 12.1: Erwartungswerte (5 Punkte)



Berechnen Sie folgende Erwartungswerte:

- (i)  $E[X^5]$ , wobei  $X \sim B_{1,p}$ .
- (ii)  $E[\sqrt{X}]$ , wobei  $X \sim B_{2,p}$ .
- (iii)  $E[\cos(X)]$ , wobei  $X \sim U_{\{0; \frac{\pi}{2}; \pi\}}$ .
- (iv)  $E[e^X]$ , wobei  $X \sim \text{Exp}_\lambda$  und  $\lambda > 1$ .
- (v)  $E[\frac{1}{3}X^2]$ , wobei  $X \sim U_{[2;4]}$ .

### Aufgabe 12.2: Poisson-Verteilung (5 Punkte)



Sei  $X$  eine Zufallsvariable mit  $X \sim P_\lambda$  und Wahrscheinlichkeitsfunktion  $p_\lambda$ . Zeigen Sie:

(i)  $\sum_{i=0}^{\infty} p_\lambda(i) = 1$

(ii)  $E[X] = \lambda$

(iii)  $\text{Var}[X] = \lambda$



### Aufgabe 12.3: Wahrscheinlichkeiten berechnen (ab Teil (iii) R) (4 Punkte)



Berechnen Sie folgende Wahrscheinlichkeiten:

- |                                          |                                                |
|------------------------------------------|------------------------------------------------|
| (i) $P(X = 1)$ mit $X \sim N_{0,1}$      | (v) $P(X = 3)$ mit $X \sim B_{5,0,5}$          |
| (ii) $P(X = 1)$ mit $X \sim U_{\{0,1\}}$ | (vi) $P(X < 6)$ mit $X \sim P_2$               |
| (iii) $P(X \geq 1)$ mit $X \sim Exp_2$   | (vii) $P(X \in (-2; 3))$ mit $X \sim N_{0,4}$  |
| (iv) $P(X = 3)$ mit $X \sim H_{10;20;5}$ | (viii) $P(X \in (6; 7))$ mit $X \sim \chi_4^2$ |



Aufgabe 12.4: Varianz der Gleichverteilung (3 Punkte)



Berechnen Sie die Varianz der  $U_{(a,b)}$ -Verteilung.



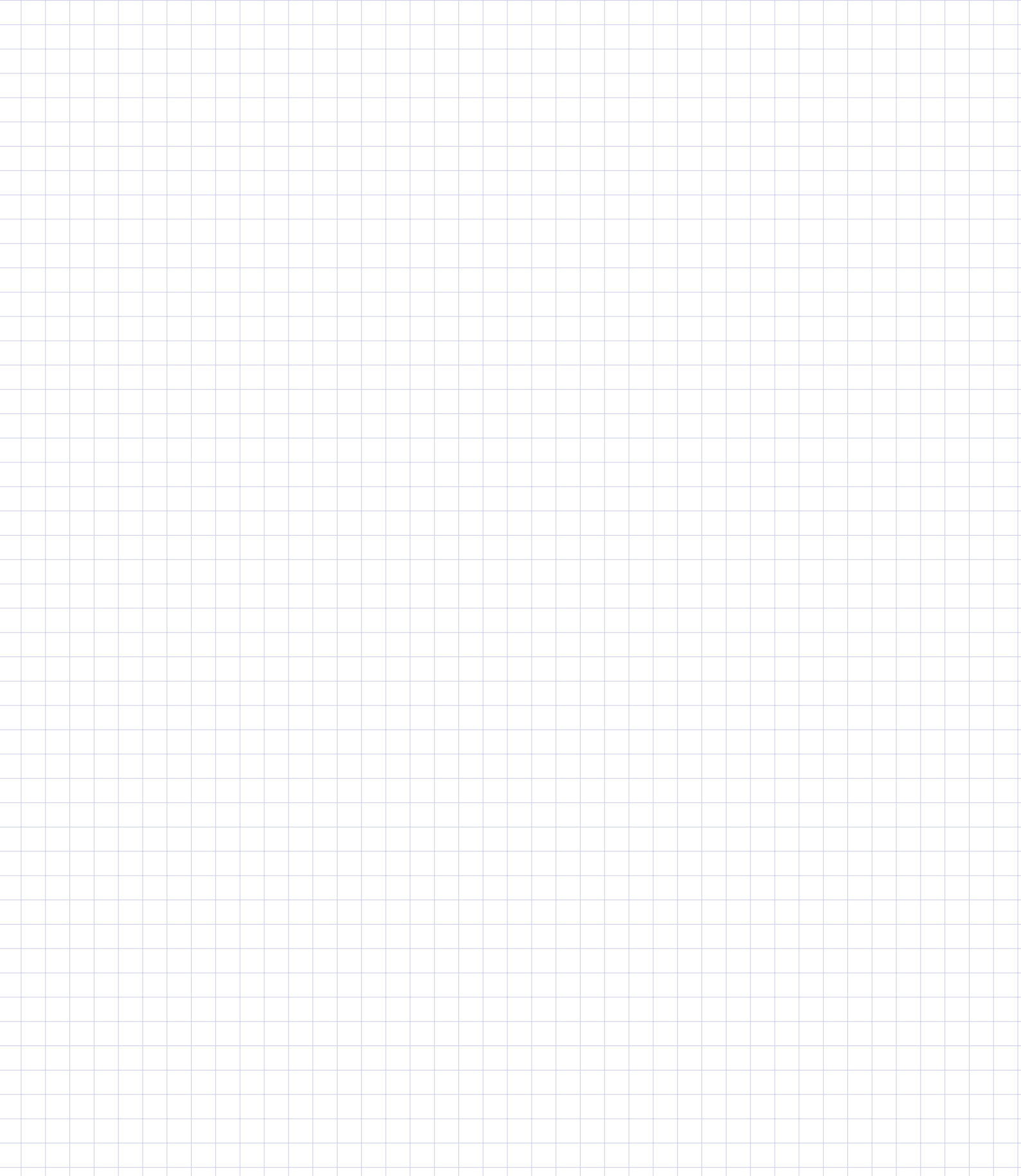
### Aufgabe 12.5: Gedächtnislosigkeit der Exponentialverteilung (2 Punkte)



Zeigen Sie, dass die Exponentialverteilung gedächtnislos ist, d.h. für alle  $\lambda, s, t > 0$  gilt:

$$X \sim \text{Exp}_\lambda \Rightarrow \mathbb{P}(X > t + s \mid X > t) = \mathbb{P}(X > s)$$

Salopp:  $X$  altert nicht. Dies gilt näherungsweise für die Lebenszeit von Transistoren, nicht aber von Lebewesen.



## Aufgabe 13.1: Substitutionsregel (7 Punkte)



Sei  $X \sim \text{Exp}_2$ .

- (i) Berechnen Sie die Lebesgue-Dichte der Zufallsvariable  $Y := 5X + 2$ .
- (ii) Berechnen Sie die Wahrscheinlichkeit  $P(Y > 11)$ .
- (iii) Berechnen Sie den Erwartungswert  $E[Y \mathbf{1}_{[5, \infty)}(Y)]$ .
- (iv) Mit welcher Befehlskombination können Sie die in den letzten beiden Teilaufgaben berechneten Größen, nämlich

$$P(Y > 11) \text{ und } E[Y \mathbf{1}_{[5, \infty)}(Y)]$$

mit der Statistiksoftware R durch Simulation abschätzen?

Sei  $X \sim N_{\mu;\sigma^2}$ .

- (i) Berechnen Sie die Lebesgue-Dichte der Zufallsvariable  $Y := e^X$ .  
*Hinweis: Die Verteilung von  $Y := e^X$  heißt Log-Normalverteilung, i.Z.:  $LN_{\mu;\sigma^2}$ , weil  $\ln(Y)$  normalverteilt ist.*
- (ii) Berechnen Sie den Erwartungswert von  $Y$ .
- (iii) Berechnen Sie die Varianz von  $Y$ .
- (iv) Berechnen Sie näherungsweise die Wahrscheinlichkeit:

$$P(Y > 4)$$

für  $\mu = 0$  und  $\sigma^2 = 1$  unter Verwendung von R:

- i.) mit Hilfe der Log-Normalverteilung (**lnorm**-Befehlsfamilie),
- ii.) mit Hilfe der Normalverteilung (**norm**-Befehlsfamilie),
- iii.) durch numerische Integration der in (i) ermittelten Dichte mit dem R-Befehl **integrate**.

Sei  $X \sim \chi_2^2$ .

(i) Berechnen Sie die Lebesgue-Dichte der Zufallsvariable  $Y := \sqrt{X}$ .

(ii) Berechnen Sie die Wahrscheinlichkeit:

$$\mathbb{P}\left(Y > \frac{1}{2}\right)$$

i.) exakt ohne Verwendung von R,

ii.) unter Verwendung der Chiquadratverteilung (`chisq`-Befehlsfamilie).

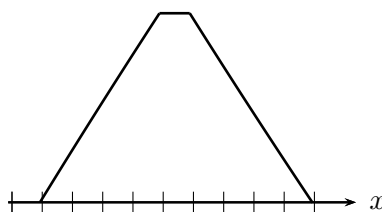
(iii) Berechnen Sie den Erwartungswert von  $Y$  numerisch mit dem R-Befehl `integrate`.

Die Verteilungen der stochastisch unabhängigen, reellwertigen Zufallsvariablen  $X$  und  $Y$  besitzen die Lebesgue-Dichten  $f$  bzw.  $g$ .

- (i) Bestimmen Sie die Lebesgue-Dichte der Verteilung der Zufallsvariablen  $Z = \frac{X}{Y}$  für den Fall, dass  $Y$  nur positive Werte annimmt.
- (ii) Betrachten Sie nun einen Autohersteller, der für ein bestimmtes Modell den Verbrauch pro 100 km ermitteln möchte. Dazu lässt der Hersteller mehrmals Autos dieses Modells eine 100 km lange Teststrecke fahren. Im Messprotokoll wird der Verbrauch pro 100 km, der als Quotient des gemessenen Verbrauchs und der gemessenen Strecke ermittelt wird, notiert.

- i.) Wie ist die Dichte der Verteilung des Verbrauchs pro 100 km, wenn Sie annehmen, dass der Verbrauch mit der  $U_{(7,9;8,1)}$ -Verteilung und die gemessene Strecke (in 100km) davon unabhängig mit der  $U_{(0,99;1,01)}$ -Verteilung modelliert werden kann?

*Hinweis: Zur moralischen Unterstützung verrate ich Ihnen, dass der Graph der gesuchten Dichte so aussieht (bei welchen  $x$ -Werten sind die Knickpunkte?):*



- ii.) Berechnen Sie die Wahrscheinlichkeit dafür, dass der gemessene Verbrauch pro 100 km (also  $\frac{X}{Y}$ ) zwischen 7,9 und 8,1 Litern liegt.
- iii.) Ermitteln Sie die in (b) gesuchte Wahrscheinlichkeit per Simulation.

Simulieren Sie 20 000 normalverteilte Zufallszahlen mit Erwartungswert 4 und Varianz 10 und berechnen Sie die Folge der arithmetischen Mittel:

$$\begin{aligned}\bar{x}_1 &= x_1 \\ \bar{x}_2 &= \frac{1}{2}(x_1 + x_2) \\ \bar{x}_3 &= \frac{1}{3}(x_1 + x_2 + x_3) \\ &\dots \\ \bar{x}_{20\,000} &= \frac{1}{20\,000}(x_1 + x_2 + \dots + x_{20\,000})\end{aligned}$$

Tragen Sie die Folge der arithmetischen Mittel über dem Index auf. Zeichnen Sie in das gleiche Diagramm noch zwei weitere Versuche mit jeweils 20 000 normalverteilten Zufallszahlen mit Erwartungswert 4 und Varianz 10 mit verschiedenen Farben ein.

Was ist zu beobachten?

*Hinweise:*

- (i) Einen Vektor  $S$  mit den kumulierten Summen seiner Einträge erhalten Sie mit dem Befehl `cumsum()`. Durch welchen Vektor müssen Sie  $S$  teilen, um die gewünschten arithmetischen Mittel zu erhalten?
- (ii) Um die starken Ausreißer zu Beginn nicht die Grafik dominieren zu lassen, ist es sinnvoll die y-Achse mit der `plot`-Option `ylim = c(3,5)` zu beschränken.

## Aufgabe 14.1: Charakt. Fkt. der Exponential-Verteilung (6 Punkte)



Die Exponentialverteilung (i.Z.  $Exp_\lambda$ ) ist durch die Lebesgue-Dichte

$$f_{Exp_\lambda}(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0;\infty)}(x)$$

gegeben ( $\lambda > 0$ ). Sie wird beispielsweise zur Modellierung der Ausfallzeit von Bauteilen oder zur Modellierung der Wartezeit auf den nächsten Kunden verwendet.

- (i) Berechnen Sie die charakteristische Funktion der Exponentialverteilung.
- (ii) Berechnen Sie Erwartungswert und Varianz der Exponentialverteilung mit Hilfe der charakteristischen Funktion.
- (iii) Berechnen Sie die charakteristische Funktion der  $n$ -fachen Faltung<sup>2</sup> der  $Exp_\lambda$ -Verteilung und vergleichen Sie diese Funktion mit der charakteristischen Funktion der  $\Gamma_{\lambda,r}$ -Verteilung:

$$\varphi_{\Gamma_{\lambda,r}}(t) = \left(1 - \frac{it}{\lambda}\right)^{-r}.$$

Was fällt Ihnen auf?

<sup>2</sup>Mit der  $n$ -fachen Faltung einer Verteilung  $P^X$  ist die Verteilung der Zufallsvariable  $\sum_{i=1}^n X_i$  gemeint, wobei alle  $X_i$  u.i. gemäß  $P^X$ -verteilt sind.

- (i) Berechnen Sie die charakteristische Funktion der Binomialverteilung  $B_{n,p}$ .

*Hinweis: Verwenden Sie den binomischen Lehrsatz.*

- (ii) Wie ist die Summe einer  $B_{n;p}$ - und einer  $B_{m;p}$ -verteilten Zufallsvariable verteilt, wenn die beiden Zufallsvariablen stochastisch unabhängig sind?

- (i) Generieren Sie einen Vektor aus zehn Millionen zufälligen Ziffern, wobei jede Ziffer 0,...,9 gleich wahrscheinlich sein soll.

*Hinweis: Verwenden Sie den Befehl `sample`.*

- (ii) Wie viele Zwillingspaare treten in dem Vektor auf, d.h. wie oft tritt am Index  $n$  die gleiche Ziffer wie am Index  $(n - 1)$  auf? ( $n = 2, 3, \dots, 10\,000\,000$ )

- (iii) Bei welchem Index wird erstmals eine Folge von sechs gleichen Ziffern vollendet? Geben Sie den Index sowie die Folge der sechs Ziffer aus.

*Hinweis: Diese und die beiden folgenden Teilaufgaben können entweder mit einer Schleife oder (auf Stringebene) mit der Funktion `regexpr` gelöst werden. Wenn Sie Zeit und Lust haben, können Sie das Paket `microbenchmark` verwenden um herauszufinden, welche Methode die effizienter ist.*

- (iv) Bei welchem Index wird erstmals die Folge 1,2,3,4,5,6 vollendet?

- (v) Bei welchem Index wird erstmals Ihr Geburtsdatum in der Form TTMMJJ (also z.B. die Folge 2,3,0,3,9,8 wenn Sie am 23.3.1998 geboren worden sind) vollendet?

- (vi) Welche Zahlen erwarten Sie bei (ii) bis (iv)? *Hinweis: Machen Sie sich für (iii) und (iv) mit der geometrischen Verteilung vertraut.*

- (vii) Mit welchem Index überschreitet die Summe der bisherigen Zufallszahlen den Wert 1 000 000?

- (viii) Laden Sie die ersten zehn Millionen Nachkommastellen von  $\pi$  mit der Funktion `scan` aus der Textdatei „pi-10-million.txt“ in den Vektor `pi.str`.

Die Datei finden Sie im Moodle-Kursraum.

- (ix) Lösen Sie die Aufgabenteile (ii) bis (v) für `pi.str`.

*Bemerkung: Beim Lösen von (iii) für `pi.str` werden Sie den Feynman-Punkt finden. Wer Genaueres wissen will, der lese: <https://de.wikipedia.org/wiki/Feynman-Punkt>*

#### Aufgabe 14.4: Verschiedene Konvergenzen (6 Punkte)



Sind folgende Aussagen wahr oder falsch? Begründen Sie Ihre Antworten!

- (i) Sei  $X_n : \mathbb{R} \rightarrow \{0,1\}$ ;  $X_n = \mathbb{1}_{\{1,\dots,n\}}$  für alle  $n \in \mathbb{N}$ . Es gilt:  $X_n \rightarrow 1$   $N_{0,1}$ -fast sicher.
- (ii) Sei  $X_n : \mathbb{R} \rightarrow \{0,1\}$ ;  $X_n = \mathbb{1}_{\{1,\dots,n\}}$  für alle  $n \in \mathbb{N}$ . Es gilt:  $X_n \rightarrow 1$  in Ws. (mit  $P = N_{0,1}$ ).
- (iii) Sei  $X_n : \mathbb{R} \rightarrow \{0,1\}$ ;  $X_n = \mathbb{1}_{\{1,\dots,n\}}$  für alle  $n \in \mathbb{N}$ . Es gilt:  $X_n \rightarrow 1$   $U_{\{1,2\}}$ -fast sicher.

## Aufgabe 15.1: Abschätzen von Wahrscheinlichkeiten (Teil (iii): R) (7 Punkte)



Nehmen Sie an, dass die Verteilung der Körpergröße der deutschen Männer mit einer  $N_{180[cm];90[cm^2]}$ -Verteilung modelliert werden kann. Gesucht ist die Wahrscheinlichkeit dafür, dass das arithmetische Mittel aus den Größen zehn zufällig ausgewählter Männer um weniger als 9 [cm] von 180 [cm] abweicht.

- (i) Verwenden Sie die Chebyshev-Ungleichung, um die Wahrscheinlichkeit abzuschätzen.
- (ii) Berechnen Sie den genauen Wert obiger Wahrscheinlichkeit. Warum bleibt der berechnete Wert ungefähr richtig, wenn die Verteilung der Körpergröße der deutschen Männer von der Normalverteilung abweicht?
- (iii) Nehmen Sie nun an, dass die Verteilung des natürlichen Logarithmus der Körpergröße der deutschen Männer mit einer  $N_{5.19;0.0527}$ -Verteilung modelliert werden kann. Die Körpergröße genügt dann einer Log-Normalverteilung mit Parametern  $\mu = 5.19$  und  $\sigma = 0.0527$ , i.Z.  $K \sim LN_{5.19,0.0527}$  (vgl. Übungsaufgabe 13.2), wobei  $K$  die Körpergröße bezeichne. Der Erwartungswert der Körpergröße ist dann wie oben (gerundet) 180 [cm]<sup>3</sup> und die Varianz wie oben (gerundet) 90 [cm<sup>2</sup>]<sup>4</sup>. Schätzen Sie nun per Simulation die obige Wahrscheinlichkeit<sup>5</sup>.

---

<sup>3</sup> $E[LN_{\mu,\sigma}] = e^{\mu + \frac{\sigma^2}{2}}$

<sup>4</sup> $\text{Var}[LN_{\mu,\sigma}] = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$

<sup>5</sup>Die Berechnung der exakten Wahrscheinlichkeit gestaltet sich in diesem Fall sehr schwierig, weil die Faltung von Lognormal-verteilten Zufallsvariable eine sehr komplizierte Dichte besitzt.

Sie arbeiten in der Qualitätskontrolle einer Chipfabrik. Ihr Chef fordert, dass die Chips im Schnitt drei Jahre lang fehlerfrei funktionieren müssen. Sie planen daher, den Chef zu verständigen, wenn die durchschnittliche Ausfallzeit innerhalb einer Stichprobe weniger als 32 Monate beträgt. Sie wollen ihn aber ungern umsonst stören, weswegen die Wahrscheinlichkeit ihn zu stören, obwohl die erwartete Lebensdauer der Chips drei Jahre beträgt, höchstens 1% sein soll.

- (i) Welche Verteilung bietet sich zur Modellierung der Ausfallzeit an?
- (ii) Wie groß muss der Stichprobenumfang sein, wenn Sie die Chebychev-Ungleichung verwenden?
- (iii) Wie groß muss der Stichprobenumfang sein, wenn Sie den Zentralen Grenzwertsatz als Hilfsmittel für eine Näherung verwenden?
- (iv) Wie groß muss der Stichprobenumfang sein, wenn Sie mit der wahren Verteilung (welcher?) rechnen?

In Baden-Württemberg gab es in den Jahren 1996-1999 231.432 männliche und 218.674 weibliche Geburten. Es stellt sich die Frage, ob dies mit der Vermutung konsistent ist, dass beide Geburtsarten gleich wahrscheinlich sind (aus Meintrup/Schäffler, Bsp. 7.44, S.208f). Um diese Vermutung zu prüfen, berechnen wir die Wahrscheinlichkeit dafür, dass man obige Geburtenzahlen oder noch mehr männliche Geburten beobachtet, wenn beide Geburtsarten gleich wahrscheinlich sind.

- (i) Schätzen Sie die gesuchte Wahrscheinlichkeit mit der Chebychev-Ungleichung ab.
- (ii) Berechnen Sie die gesuchte Wahrscheinlichkeit unter Verwendung des Zentralen Grenzwertsatzes.
- (iii) Berechnen Sie die gesuchte Wahrscheinlichkeit unter Verwendung der wahren Verteilung (welcher?) der Geschlechter.

#### Aufgabe 15.4: Simulation des Zentralen Grenzwertsatzes mit R (4 Punkte)



Seien  $(\Omega, \mathcal{F}, P)$  ein Wahrscheinlichkeitsraum und  $(X_n)_{n \in \mathbb{N}}$  eine Folge von u.i.  $B_{1;0,9}$  verteilten Zufallsvariablen auf  $(\Omega, \mathcal{F}, P)$ . Realisieren Sie mit einem Statistikprogramm jeweils 50000 Werte der Zufallsvariable

$$\sqrt{k} \frac{\frac{1}{k} \sum_{i=1}^k X_i - 0,9}{0,3}$$

für  $k = 10, 30, 100, 1000, 10000$  und zeichnen Sie die zugehörigen empirischen Verteilungsfunktionen. Zeichnen Sie zum Vergleich die Verteilungsfunktion der Standardnormalverteilung. Was fällt Ihnen auf?

