FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACHBEREICH WIRTSCHAFTS-WISSENSCHAFTEN

**Bachelorprüfung Statistik (RUW-2172), Sommersemester 2021**

Liebe Studierende,

markieren Sie bitte bei den Single-Choice-Fragen Ihre Antwort auf dem Antwortbogen am Ende des Gehefts in der folgenden Weise: ○⊗○ .
Wenn Sie eine Antwort korrigieren möchten, füllen Sie bitte die **falsch** markierte Antwort vollständig aus, ungefähr so: ○●⊗ .
Die Fünfecke beziehen sich auf den Freitextteil und werden nur von der Korrektorin bzw. dem Korrektor ausgefüllt; wenn Sie ein Fünfeck selbst markieren, erhalten Sie für die betreffende Frage 0 Punkte.

Bitte füllen Sie folgende Angaben deutlich lesbar aus:

**Nachname** : _____

**Vorname** : _____

**Matrikelnummer** : _____

**Studiengang** : _____

**Raum, Platz** : _____

**Prüfer** : Prof. Dovern



**WICHTIG: Bitte kreuzen Sie Ihre Matrikelnummer auch auf dem Antwortbogen an!**

_____

_____

Nachfolgende Angaben sind nur vom Prüfer auszufüllen:

Aufgaben 1+2: Teilnote:

_____

Aufgabe 3: Teilnote:

_____

Gesamtnote:

_____

Unterschrift Prüfer:

**Bitte beachten Sie folgende Hinweise**:

- Das Geheft **muss** zusammen bleiben!
- Die Klausur besteht aus einem **Single-Choice** und einem **Freitextteil**.
- **Single-Choice-Teil (Aufgaben 1 und 2)**
  - Der Single-Choice-Teil umfasst 27 Single-Choice-Fragen.
  - Verwenden Sie für Ihre Antworten zu den Single-Choice-Fragen ausschließlich den Single-Choice-Antwortbogen am Ende des Gehefts. **Einträge in der Aufgabenstellung werden nicht gewertet!**
  - Beschriften Sie den Antwortbogen deutlich lesbar mit Ihrem Namen und Ihrer Matrikelnummer, und kreuzen Sie Ihre Matrikelnummer zusätzlich an!
  - Verwenden Sie auf dem Antwortbogen bitte einen **dunklen Kugelschreiber**!
- **Freitextteil (Aufgabe 3)**
  - Der Freitextteil umfasst 12 offene Aufgaben, die in den Lösungsfeldern in **diesem Geheft** zu beantworten sind.
  - Schreiben Sie Ihre Freitextantworten **lesbar**.
- Bearbeitungszeit: 120 Minuten
- **Erlaubte Hilfsmittel:**
  - Nicht-programmierbarer Taschenrechner
  - Die vom Lehrstuhl offiziell herausgegebene Formelsammlung, 2. bis 4. Auflage, ohne weitere Eintragungen oder Markierungen, mit Ausnahme von farblichen Hinterlegungen von Textpassagen und/oder Formeln bzw. unbeschriebenen Post-Its
  - Cheat Sheet für Basics in R, das über StudOn bereitgestellt wurde, ohne weitere Eintragungen oder Markierungen, mit Ausnahme von farblichen Hinterlegungen von Textpassagen und/oder Befehlen



**Viel Erfolg!**

_____

_____

# Bachelorprüfung Statistik, SoSe 2021

---

## Aufgabe 1: Single-Choice-Fragen

---

**Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.**

**Hinweis**: Aufgabe 1 besteht aus 18 Teilaufgaben, bei denen jeweils 3 Punkte erreicht werden können. Jede Frage bietet mehrere Antwortmöglichkeiten, von denen **jeweils nur eine korrekt ist**. Kreuzen Sie jeweils die korrekte Antwort **auf dem Antwortbogen** an. Beachten Sie, dass es **keinen Punktabzug für falsch beantwortete Fragen** gibt.



**1.1** Die Zufallsvariable $X$ sei normalverteilt mit unbekanntem Erwartungswert $\mu$ und bekannter Varianz $\sigma^2$. Man erhebt eine *i.i.d.*-Stichprobe um $\mu$ zu schätzen. Welche Behauptung über Konfidenzintervalle der Form $\left[\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ ist korrekt?

**A** Je höher das Konfidenzniveau gewählt wird, desto breiter ist das Konfidenzintervall.
**B** Je größer die Standardabweichung, desto schmaler ist das Konfidenzintervall.
**C** Die t-Verteilung wird nicht eingesetzt, weil $\mu$ unbekannt ist.
**D** Das Konfidenzintervall ist asymmetrisch bezüglich $\mu$.
**E** Die Breite des Konfidenzintervalls nimmt mit steigendem Stichprobenumfang zu.

Für eine gegebene Stichprobe haben Sie die durchschnittliche Tagesrendite $\bar{x}$ der Kryptowährung Bitcoin im Jahr 2020 berechnet. Sie gehen davon aus, dass die Tagesrendite durch eine unabhängig identisch normalverteilte Zufallsvariable mit unbekanntem Erwartungswert $\mu$ und bekannter Varianz $\sigma^2$ beschrieben werden kann. Ein Bekannter behauptet, dass die durchschnittliche Tagesrendite negativ sei. Sie sind jedoch vom Gegenteil überzeugt. Auf dem Signifikanzniveau $\alpha = 0.1$ testen Sie daher die Behauptung des Bekannten.

**1.2** Ihre Berechnung ergibt die Teststatistik: $t = 1.813$. Welche Behauptung über die Testentscheidung auf Basis dieser realisierten Prüfgröße ist korrekt?

**A** Da $t > z_{1-\alpha}$, wird $H_0$ auf dem 90%-Niveau verworfen. Die Behauptung Ihres Bekannten steht somit im Widerspruch zu den Daten.

**B** Da $t < -z_{1-\alpha}$, wird $H_0$ auf dem 90%-Niveau verworfen. Die Behauptung Ihres Bekannten steht somit im Widerspruch zu den Daten.

**C** Da $t > z_{1-\alpha}$, wird $H_0$ auf dem 90%-Niveau nicht verworfen. Die Behauptung Ihres Bekannten steht somit im Widerspruch zu den Daten.

**D** Da $t < z_{1-\alpha}$, wird $H_0$ auf dem 90%-Niveau nicht verworfen. Die Behauptung Ihres Bekannten steht somit nicht im Widerspruch zu den Daten.

**E** Da $t > z_{1-\alpha}$, wird $H_0$ auf dem 90%-Niveau verworfen. Die Behauptung Ihres Bekannten steht somit nicht im Widerspruch zu den Daten.



**1.3** Gegeben sind drei Zufallsvariablen mit den folgenden Verteilungen:

$X \sim N(\mu = 5, \sigma^2 = 3)$
$Y \sim t_{10}$
$Z \sim t_{100}$

Welche der folgenden Aussagen bezüglich der Überschusskurtosis $\kappa$ ist korrekt?

**A** $\kappa_Y > \kappa_Z > \kappa_X$

**B** $\kappa_X > \kappa_Z > \kappa_Y$

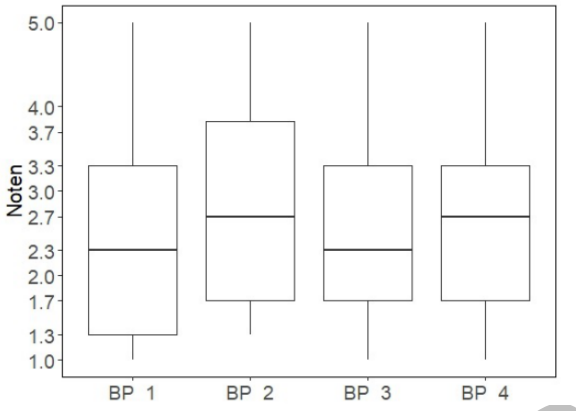
**C** $\kappa_Y > \kappa_X > \kappa_Z$

**D** $\kappa_X = \kappa_Y = \kappa_Z$

**E** $\kappa_X > \kappa_Y = \kappa_Z$

Das Merkmal $X$ gibt die erzielte Note in einer Klausur wieder. Die folgende Urliste beschreibt die Notenverteilung einer Klausur für 16 Studierende, $i = 1, \ldots, 16$. Außerdem sind Boxplots (BP) für verschiedene Notenverteilungen gegeben.

| $i:$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i:$ | 1.0 | 1.0 | 1.3 | 1.7 | 1.7 | 2.0 | 2.0 | 2.3 | 2.3 | 2.7 | 3.0 | 3.3 | 3.3 | 3.7 | 4.0 | 5.0 |




**1.4** Welcher Boxplot passt zu der gegebenen Urliste?

**A** BP 3
**B** BP 2
**C** BP 1
**D** BP 4
**E** Keiner



Die folgende Tabelle zeigt den Pro-Kopf-Wasserverbrauch in einer Gemeinde in Litern pro Tag für drei Verbrauchsklassen, $j = 1, 2, 3$. Dabei gibt $n_j$ die Anzahl der Beobachtungen, $\overline{x}_j$ den Durchschnitt und $s_j^2$ die Varianz der jeweiligen Klasse an. In der letzten Spalte der Tabelle stellt $\overline{x}$ das arithmetische Mittel des gepoolten Datensatzes dar.

| $j$ | Klassengrenzen | $n_j$ | $\overline{x}_j$ | ${s_j}^2$ | $(\overline{x}_j - \overline{x})^2$ |
|---|---|---|---|---|---|
| 1 | $[0, 100)$ | 4 | 90 | 25 | 2862.25 |
| 2 | $[100, 200)$ | 10 | 125 | 15 | 342.25 |
| 3 | $[200, 300)$ | 6 | 210 | 50 | 4422.25 |

**1.5** Welche der folgenden Aussagen ist korrekt?

**A** Die Varianz innerhalb der Verbrauchsklassen ist kleiner als die Varianz zwischen den Verbrauchsklassen.
**B** Die Gesamtvarianz $s^2$ entspricht 45.8.
**C** Die Varianz innerhalb der Verbrauchsklassen entspricht 5.24.
**D** Die Varianz zwischen den Verbrauchsklassen entspricht 45.5.
**E** Die Varianz innerhalb der Verbrauchsklassen ist größer als die Gesamtvarianz $s^2$.

**1.6** Welche der folgenden Aussagen bezüglich der relativen Konzentration und der Lorenzkuve ist korrekt?

**A** Bei vollständiger Disparität mit $n$ Merkmalsträgern liegt die Lorenzkurve auf der x-Achse an der Nulllinie bis zur Stelle $(n-1)/n$.

**B** Bei vollständiger Disparität fällt die Lorenzkurve mit der Diagonalen zusammen.

**C** Man spricht vom Vorliegen einer niedrigen relativen Konzentration, wenn ein großer Anteil der Merkmalssumme auf eine kleine Anzahl von Merkmalsträgern verteilt ist.

**D** Die Fläche unter der Lorenzkurve nennt man Konzentrationsfläche.

**E** Die Lorenzkurve ergibt sich als Darstellung der einzelnen Merkmalsanteile gegen die kumulierten Anteile der Merkmalsträger.

Sie interessieren sich dafür, ob Studierende, die online für eine Prüfung lernen, bessere Prüfungsergebnisse erzielen als Studierende, die sich offline für die Prüfung vorbereiten. Sie betrachten die folgenden Ereignisse:

*O:* "Die Person lernt für die Prüfung online"
*B:* "Die Person besteht die Prüfung"

Zudem seien folgende Wahrscheinlichkeiten bekannt:

| | $O$ | $\overline{O}$ | $\Sigma$ |
|---|---|---|---|
| $B$ | 0.13 | 0.27 | 0.40 |
| $\overline{B}$ | 0.20 | 0.40 | 0.60 |
| $\Sigma$ | 0.33 | 0.67 | 1 |



**1.7** Durch welchen Ausdruck und durch welchen Wert ist die Wahrscheinlichkeit gegeben, dass eine Person nicht online lernt und die Prüfung nicht besteht?

**A** $P(\overline{O \cup B}) = 0.40$

**B** $P(\overline{O} \setminus \overline{B}) = 0.07$

**C** $P(\overline{O \cap B}) = 0.40$

**D** $P(\overline{O} \cup \overline{B}) = 0.87$

**E** $P(\overline{O \cup B}) = 0.77$

**1.8** Durch welchen Ausdruck und durch welchen Wert ist die Wahrscheinlichkeit gegeben, dass eine Person die Prüfung besteht, unter der Bedingung, dass sie nicht online lernt?

**A** $P(B|\overline{O}) = 0.40$

**B** $P(\overline{O}|B) = 0.40$

**C** $P(B|\overline{O}) = 0.66$

**D** $P(\overline{O}|B) = 0.66$

**E** $P(B \cap \overline{O}) = 0.27$

Die Zufallsvariable $X$ gibt die Anzahl von Toren in einem Fußballspiel an. Die folgende Tabelle stellt die Wahrscheinlichkeitsfunktion von $X$ dar:

| Anzahl von Toren $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | $\geq 7$ |
|---|---|---|---|---|---|---|---|---|
| Wahrscheinlichkeit $P(X = x)$ | 0.1 | 0.15 | 0.2 | 0.22 | 0.15 | 0.07 | 0.05 | 0.06 |

**1.9** Wie groß ist die Wahrscheinlichkeit dafür, dass mindestens 1 Tor und höchstens 5 Tore fallen?

**A** 0.79
**B** 0.89
**C** 0.82
**D** 0.21
**E** 0.11

**1.10** Auf einem Jahrmarkt werden Verzehrgutscheine verlost. Im Lostopf befinden sich 4800 Nieten und 200 Gewinnlose. Der erste Teilnehmer zieht 10 Lose. Wie groß ist die Wahrscheinlichkeit genau 2 Gewinnlose zu ziehen?

**A** 0.0518
**B** 0.1931
**C** 0.0138
**D** 0.0056
**E** 0.0259



Betrachten Sie die Zufallsvariable $Y$ mit folgender Dichtefunktion:

$$f(y) = \begin{cases} -0.5y + 5 & \text{für } 8 \leq y \leq 10 \\ 0 & \text{sonst} \end{cases}$$

**1.11** Wie lautet die zugehörige Verteilungsfunktion?

**A** $F(y) = -0.25y^2 + 5y - 24$
**B** $F(y) = -\infty$
**C** $F(y) = 0.25y^2 + 5y + 47.5$
**D** $F(y) = -0.5y^2 + 5y$
**E** $F(y) = -0.25y^3 + 2.5y^2$

**1.12** Der Alkoholgehalt einer Weinflasche ist normalverteilt mit einem Erwartungswert von 13% und einer Standardabweichung von 2.5%. Wie hoch ist die Wahrscheinlichkeit dafür, dass der Alkoholgehalt einer Weinflasche 15% übersteigt?

**A** 0.2119
**B** 0.8000
**C** 0.7881
**D** 0.3745
**E** 0.6255

**1.13** Eine Gemeinde erhebt die Merkmale "Anzahl an heißen Tagen pro Monat" ($T$) und "Sterbefälle aufgrund von Herz-Kreislauf-Versagen pro Monat" ($S$). Die Kovarianz der Merkmale betrage 55.00. Es ist weiterhin bekannt, dass die Varianz von $T$ einen Wert von 38.44 hat, während die Varianz von $S$ einen Wert von 222.01 hat. Welcher Wert entspricht dem Korrelationskoeffizienten nach Pearson?

**A** 0.5954
**B** 0.0064
**C** 0.3545
**D** 0.0803
**E** 0.0550

**1.14** Ein Büromitarbeiter empfängt durchschnittlich 9 Emails pro Stunde. Das Eintreffen der Emails pro Stunde ist Poisson-verteilt. Welchem Verteilungsmodell folgt die Zufallvariable $D$: "Wartezeit zwischen dem Eintreffen zweier aufeinanderfolgender Emails"?

**A** Poisson-Verteilung: $D \sim P(\lambda = 9)$
**B** Exponentialverteilung: $D \sim Exp(\lambda = 9)$
**C** Normalverteilung: $D \sim N(9, 60)$
**D** Exponentialverteilung: $D \sim Exp(\lambda = 9/60)$
**E** Poisson-Verteilung: $D \sim P(\lambda = 3)$



**1.15** Ein Student schreibt seine Bachelorarbeit über die Akzeptanz von Finanz-Apps und erstellt zu diesem Zweck eine Umfrage. Zur Datenerhebung postet er die Umfrage in einem großen Onlineforum. Wie nennt man das hier gewählte Stichprobenverfahren?

**A** Cluster-Stichprobe
**B** Einfache Zufallsstichprobe
**C** Auswahl auf's Geratewohl
**D** Geschichtete Stichprobe
**E** Typische Stichprobe

**1.16** Gegeben sei der Maximum-Likelihood-Schätzer für den Parameter $p$ einer binomialverteilten Zufallsvariable $X$:

$$\hat{p}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Welche der folgenden Aussagen ist bezüglich der Eigenschaften des dargestellten Schätzers **nicht** korrekt?

**A** Der Schätzer ist erwartungstreu.
**B** Der Schätzer ist konsistent.
**C** Ein kleiner mittlerer quadratischer Abstand garantiert, dass jeder Schätzwert $\hat{p}_{ML}$ sehr nahe am tatsächlichen Parameterwert liegt.
**D** Die Differenz $E(\hat{p}_{ML}) - p$ ist Null.
**E** Für sehr große Stichproben geht die Varianz von $\hat{p}_{ML}$ gegen Null.

Sie wollen eine Zielvariable $Y$ durch einen Entscheidungsbaum erklären. Sie betrachten dazu die Merkmalsausprägungen der Merkmale $A$, $B$ und $C$ für $n = 20$ Beobachtungen.

Die Entropie des Merkmals $Y$ betrage $E(Y) = 0.72$.

Die folgende Tabelle gibt Ihnen die Entropien an, die sich bei der Partitionierung des Datensatzes anhand der angegebenen Merkmalsausprägungen ergeben:

| | $n(\bullet)$ | $E(Y\|\bullet)$ |
|---|---|---|
| $A =$ "Ja" | 10 | 0.6 |
| $A =$ "Nein" | 10 | 0.5 |
| $B =$ "<50" | 2 | 0 |
| $B =$ ">50 und <85" | 12 | 0.2 |
| $B =$ ">85" | 6 | 0.5 |
| $C = 11.3$ | 5 | 0.7 |
| $C = 11.8$ | 7 | 0.2 |
| $C = 12.5$ | 4 | 0.4 |
| $C = 13.1$ | 4 | 0.1 |

**1.17** Welche der folgenden Aussagen zum Entscheidungsbaum ist korrekt?

**A** Merkmal $A$ ist gemäß der Entropieänderung das informativste Merkmal.
**B** Merkmal $B$ ist gemäß der Entropieänderung das informativste Merkmal.
**C** Die Merkmale $B$ und $C$ sind gemäß der Entropieänderung die informativsten Merkmale. Der Wurzelknoten ist daher nicht eindeutig.
**D** Merkmal $C$ ist gemäß der Entropieänderung das informativste Merkmal.
**E** Da es sich bei $C$ um ein quantitatives Merkmal handelt, sollte es nicht als Wurzelknoten verwendet werden.

**1.18** Bei einem Fußballspiel werden für alle Spielerinnen folgende Daten erhoben:

1. gelaufene Distanz
2. Schüsse aufs Tor
3. Performance anhand einer Notenskala von 1 bis 6

Ordnen Sie den drei erhobenen Merkmalen die korrekten Skalenniveaus zu.
**A** 1. Absolutskala, 2. Intervallskala, 3. Ordinalskala
**B** 1. Verhältnisskala, 2. Absolutskala, 3. Ordinalskala
**C** 1. Intervallskala, 2. Ordinalskala, 3. Ordinalskala
**D** 1. Intervallskala, 2. Ordinalskala, 3. Nominalskala
**E** 1. Verhältnisskala, 2. Verhältnisskala, 3. Absolutskala

**Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.**

## Aufgabe 2: Single-Choice-Fragen zu R

**Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.**

**Hinweis**: Aufgabe 2 besteht aus 9 Teilaufgaben, bei denen jeweils 3 Punkte erreicht werden können. Jede Frage bietet mehrere Antwortmöglichkeiten, von denen **jeweils nur eine korrekt ist**. Kreuzen Sie jeweils die korrekte Antwort **auf dem Antwortbogen** an. Beachten Sie, dass es **keinen Punktabzug für falsch beantwortete Fragen** gibt.

**2.1** Es sei $X$ eine standardnormalverteilte Zufallsvariable. Vervollständigen Sie den Befehl

```
ggplot(data=data.frame(x=-4:4), aes(x=x)) +
   stat_function(fun = Y, args=list(mean=0, sd=1), color="blue") +
   stat_function(fun = Z, args=list(mean=0, sd=1), color="red")
```

so, dass die **Dichtefunktion** von $X$ **in blau** sowie die **Verteilungsfunktion** von $X$ **in rot** in einem gemeinsamen Schaubild dargestellt werden.

**A** **Y**: dnorm, **Z**: pnorm
**B** **Y**: pnorm, **Z**: dnorm
**C** **Y**: dnorm, **Z**: qnorm
**D** **Y**: qnorm, **Z**: pnorm
**E** **Y**: rnorm, **Z**: pnorm



**2.2** Es sei $X$ eine binomialverteilte Zufallsvariable mit $n = 5$ und $p = 0.3$. Mit welchem Befehl können Sie die Wahrscheinlichkeit $P(X > 2)$ berechnen?

**A** 1 - pbinom(2, size=5, prob=0.3)
**B** pbinom(2, size=5, prob=0.3)
**C** 1 - pbinom(2, size=5, prob=0.7)
**D** dbinom(2, size=5, prob=0.3)
**E** 1 - dbinom(2, size=5, prob=0.3)

**2.3** Welchen Output zeigt die nachfolgende for-Schleife in der R-Konsole an?

```
for(i in 1:5){
   j <- i^2
   k <- (j/i)+1
   print(k)
}
```

**A** 2 4 6 8 10
**B** 1 2 3 4 5
**C** 2 3 4 5 6
**D** 2 3 5 8 13
**E** 1 2 3 5 8

Gehen Sie für die nächsten Fragen von dem folgenden Workspace in R aus. Der Dataframe df enthält Informationen aus dem Current Population Survey, einer Umfrage unter U.S.-amerikanischen Haushalten. Ihre Daten umfassen die Antworten von $n = 7986$ befragten Personen. Für jede Person enthält der Dataframe Angaben zu den folgenden Merkmalen:

Spalte 1: Der Stundenlohn der befragten Person in U.S.-Dollar. (wage)

Spalte 2: Eine Indikatorvariable für das Geschlecht, die den Wert 1 für Frauen annimmt und 0 für Männer. (female)

Spalte 3: Eine Indikatorvariable für den Bildungsstand, die den Wert 1 annimmt, falls die befragte Person einen Bachelorabschluss hat und 0 wenn nicht. (bachelor)

Spalte 4: Das Alter der befragten Person in Jahren. (age)

Für jede befragte Person liegen vollständige Informationen zu allen Merkmalen vor (d.h. es gibt keine NAs). Es gibt keine weiteren Spalten im Dataframe und Sie haben auch sonst keine Datenobjekte (z.B. Values oder Funktionen) abgespeichert. Sie haben das Paket tidyverse in Ihrer aktuellen Session bereits aktiviert.

Im Rahmen Ihrer Analyse haben Sie die Befehlssequenz

```
df %>%
  group_by(female, bachelor) %>%
  summarize(info1 = mean(wage), info2 = sqrt( (NROW(wage)-1)/NROW(wage) * var(wage) ))
```

ausgeführt und die folgende Tabelle als Output erhalten:

| | female | bachelor | info1 | info2 |
|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 0 | 0 | 14.9 | 7.16 |
| 2 | 0 | 1 | 22.0 | 10.4 |
| 3 | 1 | 0 | 11.9 | 5.39 |
| 4 | 1 | 1 | 18.5 | 8.16 |



**2.4** Betrachten Sie die oben angezeigte Tabelle. Welche der folgenden Aussagen zu den durchschnittlichen Stundenlöhnen der verschiedenen Personengruppen ist **nicht** korrekt?

**A** Männer mit Bachelorabschluss haben den höchsten Durchschnittslohn.
**B** Frauen ohne Bachelorabschluss haben den niedrigsten Durchschnittslohn.
**C** Frauen mit Bachelorabschluss verdienen im Schnitt mehr als Männer ohne Bachelorabschluss.
**D** Frauen mit Bachelorabschluss verdienen im Schnitt mehr als Männer mit Bachelorabschluss.
**E** Frauen ohne Bachelorabschluss verdienen im Schnitt weniger als Männer ohne Bachelorabschluss.

**2.5** Betrachten Sie die oben angezeigte Tabelle. Welche der folgenden Aussagen zur Streuung der Stundenlöhne der verschiedenen Personengruppen ist korrekt?

**A** Die Löhne der Männer ohne Bachelorabschluss streuen am stärksten.
**B** Die Löhne der Männer mit Bachelorabschluss streuen am stärksten.
**C** Die Löhne der Frauen mit Bachelorabschluss streuen am wenigsten.
**D** Die Löhne der Frauen ohne Bachelorabschluss streuen stärker als die der Frauen mit Bachelorabschluss.
**E** Die Löhne der Frauen mit Bachelorabschluss streuen stärker als die der Männer mit Bachelorabschluss.

**2.6** Stellen Sie sich ein fiktives Szenario vor, in dem die Stundenlöhne aller Personen steigen. Die Löhne aller Frauen steigen um exakt 2 US-Dollar, während die der Männer um 1 US-Dollar steigen. Vervollständigen Sie den Befehl

mean(ifelse(df$female==**X**, df$wage+**Y**, df$wage+**Z**))

so, dass der neue Durchschnittslohn der befragten Personen in diesem fiktiven Szenario bestimmt wird.

**A** **X**: 1, **Y**: 2, **Z**: 1
**B** **X**: 0, **Y**: 2, **Z**: 1
**C** **X**: 1, **Y**: 1, **Z**: 2
**D** **X**: 2, **Y**: 1, **Z**: 1
**E** **X**: 0, **Y**: 1, **Z**: 1



**2.7** Welcher der nachfolgenden Befehle liefert Ihnen **nicht** den Pearson-Korrelationskoeffizient zwischen den Stundenlöhnen und dem Alter der befragten Personen.

**A** cor(df$wage, df$age)
**B** cov(df$wage, df$age) / (sd(df$wage) * sd(df$age))
**C** cov(df$wage, df$age) / (sqrt(var(df$wage)) * sqrt(var(df$age)))
**D** cor(df$wage, df$age, method="pearson")
**E** cov(df$wage, df$age) / (var(df$wage) * var(df$age))

**2.8** Welcher der nachfolgenden Befehle bzw. Befehlssequenzen liefert Ihnen **keine** Informationen zur Streuung der Stundenlöhne?

**A** var(df$wage)
**B** IQR(df$wage)
**C** summary(df$wage)
**D** max(df$wage) - min(df$wage)
**E** length(df$wage)

**2.9** Vervollständigen Sie den Befehl

```
df %>%
  filter(age<26 & female==X & bachelor==Y) %>%
  Z(wage)
```

so, dass Ihnen die Stundenlöhne aller unter 26-jährigen weiblichen Personen ohne Bachelorabschluss anzeigt werden.

**A** **X**: 1, **Y**: 0, **Z**: select
**B** **X**: 0, **Y**: 1, **Z**: select
**C** **X**: 1, **Y**: 0, **Z**: filter
**D** **X**: 0, **Y**: 1, **Z**: filter
**E** **X**: 1, **Y**: 1, **Z**: summarize

**Bitte vergessen Sie nicht, Ihre Antworten auf den Antwortbogen zu übertragen und dort auch Ihren Namen, Vornamen sowie Ihre Matrikelnummer anzugeben.**

## Aufgabe 3: Freitextaufgaben

**Hinweis**: Aufgabe 3 besteht aus 12 Teilaufgaben, bei denen insgesamt 39 Punkte erreicht werden können. Verwenden Sie für die Lösung der Aufgaben die durch die Linien begrenzten Lösungsfelder direkt unter dem jeweiligen Aufgabentext. **Nehmen Sie für diese Aufgabe keine Markierungen auf dem Antwortbogen vor.** Falls nötig, runden Sie Ihre Ergebnisse auf **vier Nachkommastellen**.

Sie veranstalten zu ausgewählten Spielen der Fußball-Europameisterschaft Grillfeiern im Garten und schenken dabei auch Bier aus. Sie möchten den Bierkonsum Ihrer Gäste analysieren und betrachten hierfür folgende Zufallsvariable $X$:

$X$: „Ausgeschenkte Biermenge während eines Spiels in Litern“

Die Zufallsvariable $X$ sei normalverteilt mit unbekanntem Erwartungswert $\mu$ und unbekannter Varianz $\sigma^2$.

Sie vermuten, dass die Varianz der Zufallsvariable $X$ kleiner als 9.5 ist und möchten dies mittels eines Hypothesentests überprüfen. Die dafür ermittelte *i.i.d.*-Stichprobe vom Umfang $n = 10$ ergibt eine Stichprobenstandardabweichung von $\hat{\sigma} = 3$.

**3.1** Stellen Sie für Ihre Vermutung die korrekte Null- und Alternativhypothese auf. (2 Punkte)

**3.2** Geben Sie für den oben beschriebenen Hypothesentest die Prüfgröße inklusive asymptotischer Verteilung unter der Nullhypothese an. (3 Punkte)

**3.3** Geben Sie für eine Irrtumswahrscheinlichkeit von $\alpha = 0.01$ die kritische Schranke des obigen Tests an. (2 Punkte)

**3.4** Berechnen Sie die realisierte Prüfgröße, treffen Sie die korrekte Testentscheidung und begründen Sie Ihre Entscheidung. Nehmen Sie dabei – unabhängig von Ihren Ergebnissen aus den vorherigen Teilaufgaben – an, dass die kritische Schranke den Wert 4 annimmt. (4 Punkte)



**3.5** Nehmen Sie nun an, dass die Verteilung der Zufallsvariable $X$ gegeben ist durch $X \sim N(\mu = 30, \sigma^2 = 16)$.

Berechnen Sie die Wahrscheinlichkeit dafür, dass die ausgeschenkte Menge an Bier während eines Spiels mindestens 33 Liter beträgt. (3 Punkte)

**3.6** Die Wahrscheinlichkeit für ein zentrales Schwankungsintervall von $X$ der Form $\mu - k \cdot \sigma \leq X \leq \mu + k \cdot \sigma$ betrage 90%.

Ermitteln Sie die Werte für die Ober- und Untergrenze des hier dargestellten Intervalls. (5 Punkte)

---

---



**3.7** Während der Halbzeitpausen wollen sehr viele Gäste Bratwürste essen, sodass es regelmäßig zu Warteschlangen am Grill kommen kann. Nehmen Sie an, dass die Zufallsvariable $Y$ die Wartezeit am Grill während der Halbzeitpause eines Spiels in Minuten angibt.

Es gilt: $Y \sim Exp(\lambda)$.

Der Parameter $\lambda$ sei unbekannt und soll nun auf Basis einer *i.i.d.*-Stichprobe $Y_1, \ldots, Y_n$ geschätzt werden.

Nennen Sie zwei geeignete Verfahren zur Schätzung des Parameters $\lambda$ mittels eines spezifischen Werts. (2 Punkte)

---

---

**3.8** Leiten Sie die Schätzfunktion für den Paramter $\lambda$ mittels einer Methode Ihrer Wahl her.

Bestimmen Sie anschließend für das vorliegende Beispiel den konkreten Schätzwert für den Fall, dass das Stichprobenmittel den Wert $\bar{y} = 2$ annimmt. (4 Punkte)



**3.9** Sie wollen zusätzlich eine Intervallschätzung durchführen und ermitteln dazu auf Basis Ihrer Stichprobe ein zweiseitiges Konfidenzintervall für den Parameter $\lambda$ zum 99%-Niveau. Sie erhalten folgendes Ergebnis: $KI_{0.99} = [0.315, 0.685]$.

Interpretieren Sie das realisierte Konfidenzintervall. (2 Punkte)

**3.10** Nehmen Sie nun für die Verteilung von $Y$ an: $Y \sim Exp(\lambda = 0.8)$.

Berechnen Sie die Wahrscheinlichkeit dafür, dass die Wartezeit am Grill mindestens zwei, aber höchstens drei Minuten beträgt. (4 Punkte)

**3.11** Ihnen fällt auf, dass eine Würstchensorte besonders gefragt ist. Betrachten Sie folgende relative Häufigkeiten der nachgefragten Menge aller angebotenen Bratwurstsorten:

Nürnberger Rostbratwurst: $h_1 = 0.63$
Thüringer Bratwurst: $h_2 = 0.23$
Vegane Bratwurst: $h_3 = 0.14$

Berechnen Sie die Gini-Simpson-Entropie für die genannten Wurstsorten und geben Sie zudem den Minimal- sowie Maximalwert des Streuungsmaßes im dargestellten Kontext an. (4 Punkte)

**3.12** Ihr Freund behauptet, dass der Herfindahl-Index *H* als Konzentrationsmaß ansteigen würde, wenn man die Merkmalsbeiträge von „Thüringer Bratwurst“ und „Vegane Bratwurst“ zu „Sonstige Bratwürste“ zusammenfassen würde. Ermitteln Sie den Herfindahl-Index *H* für das in der vorherigen Aufgabe dargestellte Beispiel und nehmen Sie zur Aussage Ihres Freundes begründet Stellung. (4 Punkte)

**Musterlösung**

# Bachelorprüfung Statistik, SoSe 2021

| | |
|---|---|
| **1.1** | ■ A □ B □ C □ D □ E |
| **1.2** | ■ A □ B □ C □ D □ E |
| **1.3** | ■ A □ B □ C □ D □ E |
| **1.4** | ■ A □ B □ C □ D □ E |
| **1.5** | ■ A □ B □ C □ D □ E |
| **1.6** | ■ A □ B □ C □ D □ E |
| **1.7** | ■ A □ B □ C □ D □ E |
| **1.8** | ■ A □ B □ C □ D □ E |
| **1.9** | ■ A □ B □ C □ D □ E |
| **1.10** | ■ A □ B □ C □ D □ E |
| **1.11** | ■ A □ B □ C □ D □ E |
| **1.12** | ■ A □ B □ C □ D □ E |
| **1.13** | ■ A □ B □ C □ D □ E |
| **1.14** | □ A ■ B □ C □ D □ E |
| **1.15** | □ A □ B ■ C □ D □ E |
| **1.16** | □ A □ B ■ C □ D □ E |
| **1.17** | □ A ■ B □ C □ D □ E |
| **1.18** | □ A ■ B □ C □ D □ E |
| **2.1** | ■ A □ B □ C □ D □ E |
| **2.2** | ■ A □ B □ C □ D □ E |
| **2.3** | □ A □ B ■ C □ D □ E |
| **2.4** | □ A □ B □ C ■ D □ E |
| **2.5** | □ A ■ B □ C □ D □ E |
| **2.6** | ■ A □ B □ C □ D □ E |
| **2.7** | □ A □ B □ C □ D ■ E |
| **2.8** | □ A □ B □ C □ D ■ E |
| **2.9** | ■ A □ B □ C □ D □ E |

**3.1**

*Nullhypothese* $H_0: \sigma^2 \geq 9.5$
*Alternativhypothese* $H_1: \sigma^2 < 9.5$

*2 Punkte. Je 1 TP pro richtige Hypothese.*

**3.2**

$$T = (n-1)\frac{\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

*Alternative Darstellung:* $T = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2}$

*3 Punkte. 2 TP für korrekte Prüfgröße (richtiger Test und Formel) und 1 TP für Verteilung.*

**3.3**

$$\chi^2_{n-1;\alpha} = \chi^2_{9;0.01} = 2.09$$

*2 Punkte. Je 1 TP für korrekte Angabe des Quantils und richtigen Wert.*

**3.4**

$$t = 9 \cdot \frac{3^2}{9.5} = \frac{81}{9.5} = 8.5263$$

*Da* $t = 8.5263 > 4$*, liegt die Prüfgröße nicht im kritischen Bereich und* $H_0$ *kann somit am 1%-Niveau nicht abgelehnt werden.*

*4 Punkte. 2 TP für korrekte Berechnung der Prüfgröße (pro Fehler 1 TP Abzug), 1 TP für richtige Testentscheidung und 1 TP für Begründung.*

**3.5**

$$P(X \geq 33) = 1 - P(X \leq 33) = 1 - \Phi\left(\frac{33-30}{\sqrt{16}}\right) =$$
$$1 - \Phi(0.75) = 1 - 0.7734 = 0.2266$$

*3 Punkte. Richtiger Ansatz (1 TP), Standardisierung und korrektes Ablesen aus Verteilungstabelle (1 TP), Endergebnis (1 TP).*

**3.6**

$P(\mu - k \cdot \sigma \leq X \leq \mu + k \cdot \sigma) = 0.9$
$\Rightarrow 2 \cdot \Phi(k) - 1 = 0.9$

*Bestimmung von k:*
$\Phi(k) = 0.95 \rightarrow k = z_{0.95} = 1.6448$

*Untergrenze* $= 30 - 1.6448 \cdot 4 = 23.4208$
*Obergrenze* $= 30 + 1.6448 \cdot 4 = 36.5792$

*5 Punkte. Richtiger Ansatz (1 TP), Bestimmung von k (2 TP), Berechnung Untergrenze (1TP) und Berechnung Obergrenze (1 TP).*

**3.7**

*Maximum-Likelihood-Methode, Momentenmethode*

*2 Punkte. Je ein TP pro richtiges Schätzverfahren*

**3.8**

*Über MM: Gleichsetzen der theoretischen und empirischen Momente:* $E(Y) = \overline{Y}$

*Schätzfunktion:* $\overline{Y} = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_{MM} = \frac{1}{\overline{Y}}$

*Konkreter Schätzwert:* $\hat{\lambda}_{MM} = \frac{1}{2} = 0.5$

*Alternativ über ML: Aufstellen der Log-Likelihood-Funktion, ableiten und Nullsetzen liefert die Schätzfunktion:*

$\ln(L(\lambda)) = n \cdot \ln(\lambda) - \lambda \cdot \sum_{i=1}^{n} Y_i$

$$\frac{\partial \ln(L(\lambda))}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} Y_i \stackrel{!}{=} 0 \Rightarrow \hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^{n} Y_i} = \frac{1}{\overline{Y}}$$

*4 Punkte. Herleitung Schätzfunktion (3 TP, pro Fehler bei richtigem Ansatz 1 TP Abzug), richtiger Schätzwert 1 TP.*

**3.9**

*Bei wiederholten Stichproben enthalten 99% der auf diese Weise berechneten Konfidenzintervalle den wahren, aber unbekannten Parameterwert.*

*2 Punkte.*

**3.10**

$P(2 \leq Y \leq 3) = P(Y \leq 3) - P(Y \leq 2) =$
$F_{exp}(3) - F_{exp}(2) =$
$1 - exp(-0.8 \cdot 3) - (1 - exp(-0.8 \cdot 2)) = 0.1112$

*4 Punkte. 1 TP für Ansatz, 2 TP für Verteilungsfunktion und richtiges Einsetzen, 1 TP für Endergebnis*

**3.11**

$GE = 1 - \sum_{i=1}^{n} h_i^2 = 1 - \left(0.63^2 + 0.23^2 + 0.14^2\right) =$
$= 0.5306$

*Wertebereich:* $0 \leq GE \leq 1 - \frac{1}{3} = 0.6667$

*4 Punkte. Berechnung GE (2 TP), Minimalwert und Maximalwert jeweils 1 TP.*

**3.12**

$H = 1 - GE = 1 - 0.5306 = 0.4694$ *(Alternativ über andere Formel)*

*Die Aussage ist korrekt, da die Konzentration so zunimmt (Transfereigenschaft von Konzentrationsmaßen).*
*Alternativ über Berechnung zeigen:*
$H_{neu} = 0.63^2 + 0.37^2 = 0.5338 > H$

*4 Punkte. Berechnung von H (2 TP), Stellungnahme und Begründung jeweils 1 TP*