Friedrich-Alexander-Universität
Rechts- und Wirtschafts-
wissenschaftliche Fakultät

Bachelorprüfung Data Science: Statistik (21761), Wintersemester 2021/22

Liebe Studierende,

markieren Sie bitte Ihre Antworten 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: ○●⊗ .

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:

_____

Note:

_____

Unterschrift Prüfer:

Bitte beachten Sie folgende Hinweise:

- Das Geheft **muss** zusammen bleiben!
- Die Klausur besteht aus insgesamt 20 **Single-Choice-Fragen**, von denen 4 R-Bezug haben.
- Verwenden Sie für Ihre Antworten ausschließlich den 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 dort zusätzlich an!
- Verwenden Sie auf dem Antwortbogen bitte einen **dunklen Kugelschreiber**!
- Bearbeitungszeit: 60 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 Data Science: Statistik, WiSe 2021/22

Aufgabe 1

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 16 Teilaufgaben, bei denen jeweils ein Punkt erreicht werden kann. 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.

Die Zufallsvariable X sei normalverteilt mit unbekanntem Erwartungswert μ und bekannter Varianz $\sigma^2 = 56.25$. Sie erheben eine *i.i.d.*-Stichprobe vom Umfang $n = 625$ und berechnen als Stichprobenmittel $\bar{x} = 7$.

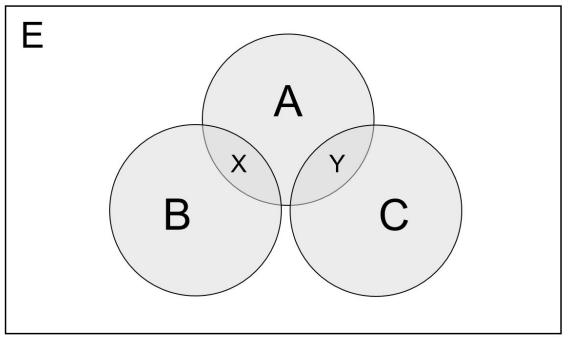
1.1 Welches der folgenden Intervalle entspricht dem zentralen Konfidenzintervall für μ zum 99%-Niveau?

A [6.2273, 7.7727]
B [6.4120, 7.5880]
C [1.2045, 12.7956]
D [2.5900, 11.4100]
E [6.7691, 7.2309]

1.2 In einem Restaurant werden an einem Tisch mit acht Personen dreimal Pasta, viermal Pizza und einmal Lasagne bestellt. Wie viele Möglichkeiten gibt es, die Gerichte auf die acht Personen am Tisch zu verteilen?

A 280
B 34
C 1680
D 4616
E 96

Betrachten Sie das folgende Venn-Diagramm, das die Mengen A, B und C darstellt:



1.3 Welche der folgenden Aussagen über das gegebene Venn-Diagramm ist **nicht** korrekt?

A $(A \cup B) \cap C = \{Y\}$
B $A \cup (B \cap C) = \{\varnothing\}$
C $(A \cap B) \cup (A \cap C) = \{X, Y\}$
D $A \cap B \cap C = \{\varnothing\}$
E $A \setminus \overline{B} = \{X\}$

Die Zufallsvariable X beschreibt die Wartezeit in Minuten zwischen zwei S-Bahnen der Berliner Ringbahn. X sei exponentialverteilt mit dem Parameter $\lambda = \frac{1}{12}$.

1.4 Wie hoch ist die Wahrscheinlichkeit, dass die Wartezeit zwischen zwei S-Bahnen mehr als 15 Minuten beträgt?

A 0.2865
B 0.7135
C 0.0239
D 0.0724
E 0.3490

Die Zufallsvariable X beschreibt den Stundenlohn eines Fabrikarbeiters in US Dollar (USD). Der Erwartungswert von X sei $E(X) = 21$, die Varianz sei $V(X) = 8$. X soll mit dem Faktor 0.85 von USD in Euro umgerechnet werden.

1.5 Was sind Erwartungswert und Varianz der Zufallsvariable Y "Stundenlohn eines Fabrikarbeiters in Euro"?

A $E(Y) = 17.85,\ V(Y) = 5.78$
B $E(Y) = 24.71,\ V(Y) = 9.41$
C $E(Y) = 17.85,\ V(Y) = 6.8$
D $E(Y) = 21.25,\ V(Y) = 6.8$
E $E(Y) = 14.45,\ V(Y) = 3.40$

Das Bundesamt für Verbraucherschutz möchte analysieren, wie stark die Böden in Deutschland mit Chemikalien belastet sind. Ein Forscherteam teilt dazu die Bodenfläche in folgende Kategorien ein: Landwirtschaft, Wald, Siedlung und Verkehr, Wasser und Sonstiges. In einem zweiten Schritt werden aus allen fünf Kategorien zufällige Bodenstichproben gezogen.

1.6 Um welches Stichprobenverfahren handelt es sich im beschriebenen Szenario?

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

Ein Kosmetikhersteller verkauft Cremes in Tuben. Die Zufallsvariable X beschreibt die Abfüllmenge in einer Tube. X sei normalverteilt mit Erwarungswert $\mu = 101$ und Standardabweichung $\sigma = 2$. Bei einer internen Analyse wird eine *i.i.d.*-Stichprobe vom Umfang $n = 250$ erhoben.

1.7 Welcher Verteilung folgt das Stichprobenmittel $\overline{X}$?

A $\overline{X} \sim N(101, 0.016)$
B $\overline{X} \sim N(0.404, 0.016)$
C $\overline{X} \sim N(101, 4)$
D $\overline{X} \sim N(101, 2)$
E $\overline{X} \sim N(101, 0.253)$

In einem Statistik-Kurs befinden sich $n = 100$ Studierende, von denen 50 weiblich und 50 männlich sind. Sie interessieren sich für den Zusammenhang zwischen dem Geschlecht der Studierenden und der Tatsache, ob diese die Statistik-Klausur bestanden haben. Die nachfolgenden Tabellen zeigen die empirisch beobachteten Häufigkeiten (links) sowie die bei Unabhängigkeit zu erwartenden Häufigkeiten (rechts):

Kontingenztabelle

	Mann	Frau
nicht bestanden	15	10
bestanden	35	40

Indifferenztabelle

	Mann	Frau
nicht bestanden	12.5	12.5
bestanden	37.5	37.5

1.8 Wie lautet der korrekte Wert der Teststatistik für den Chi-Quadrat-Unabhängigkeitstest auf Basis der in den Tabellen gegebenen Werte?

A $X^2 = 1.3333$
B $X^2 = 0$
C $X^2 = 1.3765$
D $X^2 = 2000$
E $X^2 = 133$

1.9 Welche Aussage zur Maximum-Likelihood-Methode ist bei der Schätzung von Parametern allgemein korrekt?

A Unter der Unabhängigkeitsannahme ergibt sich die Likelihoodfunktion als Produkt der individuellen Dichte- bzw. Wahrscheinlichkeitsfunktionen der einzelnen Stichprobenvariablen.
B Die Maximum-Likelihood-Methode ist der Intervallschätzung zuzordnen.
C Die Maximum-Likelihood-Methode und die Momentenmethode führen immer zu denselben Schätzern.
D Durch das Logarithmieren der Likelihoodfunktion verschiebt sich die Lage des Maximums.
E Die Maximum-Likelihood-Methode ist nur bei stetigen Verteilungsmodellen anwendbar.

Die monatlichen Mietausgaben von Studierenden in Nürnberg können durch eine normalverteilte Zufallsvariable Y mit unbekanntem Erwartungswert μ_Y und bekannter Varianz $\sigma_Y^2 = 1600$ beschrieben werden. Auf einem Immobilienportal wird behauptet, dass die mittleren monatlichen Mietausgaben 350 EUR betragen. Sie bezweifeln die Behauptung des Immobilienportals und möchten diese mit einem Hypothesentest überprüfen. Eine *i.i.d.*-Stichprobe unter $n = 150$ Studierenden ergab ein Stichprobenmittel von $\overline{y} = 338.50$.

1.10 Wie lautet die korrekte Teststatistik für diesen Hypothesentest?

A $t = -3.5211$
B $t = -0.0880$
C $t = -1.0781$
D $t = -43.1250$
E $t = -4.6734$

1.11 Auf Basis der Stichprobe berechnen Sie nun das folgende 95%-Konfidenzintervall für μ_Y:

$KI_{0.95} = [332.10, 344.90]$

Wie lautet die korrekte Testentscheidung auf Basis des realisierten Konfidenzintervalls?

A Da $350 \in KI_{0.95}$, kann H_0 zum 5%-Niveau nicht abgelehnt werden. Die Behauptung des Immobilienportals steht somit nicht im Widerspruch zu den Daten.
B Da $350 \in KI_{0.95}$, kann H_0 zum 5%-Niveau abgelehnt werden. Die Behauptung des Immobilienportals steht somit im Widerspruch zu den Daten.
C Da $350 \notin KI_{0.95}$, kann H_0 zum 5%-Niveau nicht abgelehnt werden. Die Behauptung des Immobilienportals steht somit im Widerspruch zu den Daten.
D Da $350 \notin KI_{0.95}$, kann H_0 zum 5%-Niveau abgelehnt werden. Die Behauptung des Immobilienportals steht somit nicht im Widerspruch zu den Daten.
E Da $350 \notin KI_{0.95}$, kann H_0 zum 5%-Niveau abgelehnt werden. Die Behauptung des Immobilienportals steht somit im Widerspruch zu den Daten.

Die Zufallsvariable X sei normalverteilt mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2. Man erhebt eine *i.i.d.*-Stichprobe, um μ zu schätzen.

1.12 Welche Behauptung über Konfidenzintervalle der Form $\left[\overline{X} \pm t_{n-1;1-\alpha/2}\hat{\sigma}/\sqrt{n}\right]$ ist allgemein korrekt?

A Während der zu schätzende Parameter μ eine Konstante darstellt, sind die Grenzen des Konfidenzintervalls ex ante Zufallsvariablen.
B Je größer die Stichprobenstandardabweichung, desto schmaler ist das Konfidenzintervall.
C Die Breite des Konfidenzintervalls nimmt mit zunehmenden Freiheitsgraden der t-Verteilung zu.
D Der Stichprobenumfang hat keinen Einfluss auf die Breite des Konfidenzintervalls.
E Je geringer das Konfidenzniveau gewählt wird, desto breiter ist das Konfidenzintervall.

Eine faire Münze mit den Seiten „Kopf“ (K) und „Zahl“ (Z) wird viermal hintereinander geworfen. Die Ergebnismenge E kann daher wie folgt geschrieben werden:

$$E = \{(ZZZZ), (ZZZK), ..., (KKKK)\}.$$

Betrachten Sie nun das folgende Ereignis A: „Es wird mindestens dreimal Zahl geworfen.“

1.13 Wie groß ist die Wahrscheinlichkeit $P(A)$?

A 5/16
B 4/16
C 1/2
D 4/32
E 5/32

Betrachten Sie die stetige Zufallsvariable X mit folgender Dichtefunktion:

$$f(x) = \begin{cases} 2x & \text{für } 0 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

1.14 Was ist der Erwartungswert von X?

A 2/3
B 3/2
C 1/2
D 1/3
E 3/4

Ein Automobilhersteller weiß, dass die jährliche Anzahl an Maschinenausfällen in der Produktion als Poisson-verteilte Zufallsvariable X mit $\lambda = 5$ beschrieben werden kann.

1.15 Wie groß ist die Wahrscheinlichkeit, dass innerhalb eines Jahres genau sieben Maschinen ausfallen?

A 0.1044
B 0.0595
C 0.1977
D 0.1462
E 0.3461

1.16 Gegeben sei eine unabhängig und identisch verteilte Zufallsstichprobe. Welche der folgenden Aussagen ist **nicht** korrekt?

A Der Zentrale Grenzwertsatz besagt, dass das Stichprobenmittel für große n approximativ normalverteilt ist.
B Für sehr große n ist die Verteilung des Stichprobenmittels gemäß des Zentralen Grenzwertsatzes symmetrisch.
C Das schwache Gesetz der großen Zahlen besagt, dass die Varianz des Mittelwertschätzers für $n \to \infty$ gegen die Populationsvarianz konvergiert.
D Der Zentrale Grenzwertsatz besagt, dass die Summe der Stichprobenvariablen asymptotisch normalverteilt ist.
E Gemäß des Hauptsatzes der Statistik konvergiert die empirische Verteilungsfunktion bei einem steigenden Stichprobenumfang immer näher zur theoretischen Verteilungsfunktion.

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

Aufgabe 2

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 4 Teilaufgaben, bei denen jeweils ein Punkt erreicht werden kann. 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 Poisson-verteilte Zufallsvariable mit $\lambda = 4$. Mit welchem Befehl können Sie die Wahrscheinlichkeit $P(3 \leq X \leq 6)$ **nicht** berechnen?

A dpois(3, lambda=4) + dpois(4, lambda=4) + dpois(5, lambda=4) + dpois(6, lambda=4)
B ppois(7, lambda=4) - ppois(2, lambda=4) - dpois(7, lambda=4)
C ppois(6, lambda=4) - ppois(3, lambda=4) + dpois(3, lambda=4)
D ppois(6, lambda=4) - ppois(2, lambda=4)
E ppois(6, lambda=4) - ppois(3, lambda=4)

2.2 Betrachten Sie den Vektor G unter dem Sie eine Grundgesamtheit bestehend aus den Buchstaben "A" bis "D" gespeichert haben:

```
G <- c("A", "B", "C", "D")
```

Sie möchten nun anhand des sample-Befehls zufällige Stichproben aus dieser Grundgesamtheit ziehen. Welche der nachfolgenden Aussagen ist **nicht** korrekt?

A Für sample(x = G, size = 5, replace = TRUE) wird eine *i.i.d.*-Stichprobe gezogen.
B Für sample(x = G, size = 3) wird eine einfache Zufallsstichprobe gezogen.
C Für sample(x = G, size = 5) erhalten wir eine Fehlermeldung in der Konsole.
D Für sample(x = G) erhalten wir eine Vollerhebung.
E Für sample(x = G, size = 3) kann derselbe Buchstabe auch mehrfach gezogen werden.

Gehen Sie für die nächsten Fragen von dem folgenden Workspace in R aus. Der Dataframe df enthält $n = 1000$ Realisation für die folgende Zufallsvariable:

Spalte 1: Realisationen einer normalverteilten Zufallsvariable mit **unbekanntem** Erwartungswert μ und bekannter Varianz $\sigma^2 = 4$. (var1)

Es liegen alle Realisationen der Zufallsvariable 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.

2.3 Vervollständigen Sie den Befehl

```
ggplot(data = df, aes(x = var1)) +
  geom_histogram(aes(y = ..density..), X = 1, boundary = 0) +
  stat_function(fun = Y, args = list(mean = 3, sd = Z))
```

so, dass das Histogramm der Realisationen von var1 sowie die Dichtefunktion einer Normalverteilung mit $\mu = 3$ und $\sigma^2 = 4$ in einem gemeinsamen Schaubild dargestellt werden. Die Klassen des Histogramms sollen dabei eine Klassenbreite von 1 haben und eine der Klassengrenzen soll auf dem Wert 0 liegen.

A **X**: binwidth, **Y**: dnorm, **Z**: 2
B **X**: bins, **Y**: dnorm, **Z**: 2
C **X**: binwidth, **Y**: pnorm, **Z**: 2
D **X**: bins, **Y**: pnorm, **Z**: 4
E **X**: binwidth, **Y**: dnorm, **Z**: 4

2.4 Sie möchten auf Basis der Stichprobenrealisation in der Spalte var1 ein zweiseitiges 95%-Konfidenzintervall für μ berechnen. Vervollständigen Sie den Befehl

```
c(mean(df$var1) - X(p = Y, mean = 0, sd = 1) * 2 / sqrt(Z),
  mean(df$var1) + X(p = Y, mean = 0, sd = 1) * 2 / sqrt(Z))
```

so, dass die Grenzen des Konfidenzintervalls korrekt bestimmt werden.

A **X**: qnorm, **Y**: 0.975, **Z**: 1000
B **X**: pnorm, **Y**: 0.975, **Z**: 1000
C **X**: qnorm, **Y**: 0.95, **Z**: 1000
D **X**: pnorm, **Y**: 0.95, **Z**: 999
E **X**: qnorm, **Y**: 0.975, **Z**: 999

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

Musterlösung

Bachelorprüfung Data Science: Statistik, WiSe 2021/22

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
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