FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Bachelorprüfung Data Science: Statistik (21761), Sommersemester 2021

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, SoSe 2021

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.

1.1 Drei Städte A, B und C eines Landes werden zufällig als „Teststädte“ für ein neues Produkt ausgewählt, das anschließend in diesem Land auf den Markt kommen soll. Alle Einwohner der drei Städte sind zum Testen registriert. Wie nennt man das hier gewählte Stichprobenverfahren?

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

1.2 Gegeben sind drei Zufallsvariablen mit den folgenden Verteilungen:

$X \sim N(\mu = 5, \sigma^2 = 3)$
$Y \sim t_5$
$Z \sim t_{10}$

Welche der folgenden Aussagen bezüglich der Überschusskurtosis κ 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 $\kappa_X > \kappa_Y = \kappa_Z$

1.3 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. Welche Behauptung über ein Konfidenzinterval der Form $\left[\overline{X} - t_{n-1;1-\alpha/2} \cdot \hat{\sigma}/\sqrt{n}, \overline{X} + t_{n-1;1-\alpha/2} \cdot \hat{\sigma}/\sqrt{n}\right]$ ist korrekt?

A Die Breite des Konfidenzintervalls nimmt mit zunehmenden Freiheitsgraden der t-Verteilung ab.
B Je größer die geschätzte Standardabweichung, desto schmaler ist das Konfidenzintervall.
C Das Quantil der t-Verteilung wird verwendet, weil μ unbekannt ist.
D Je höher das Konfidenzniveau gewählt wird, desto schmaler ist das Konfidenzintervall.
E Die Breite des Konfidenzintervalls nimmt mit zunehmendem Stichprobenumfang zu.

1.4 Die Zufallsvariable X sei normalverteilt mit unbekanntem Erwartungswert μ und bekannter Varianz $\sigma^2 = 25$. Sie erheben eine *i.i.d.*-Stichprobe vom Umfang $n = 100$ aus der Grundgesamtheit und berechnen als Stichprobenmittel $\overline{x} = 5$. Wie lautet das zentrale Konfidenzintervall für den Erwartungswert der Grundgesamtheit zum 95%-Niveau?

A $[4.020, 5.980]$
B $[0.100, 9.900]$
C $[4.510, 5.490]$
D $[4.178, 5.822]$
E $[3.712, 6.288]$

1.5 Auf der Verpackung eines Orangensaftes wird behauptet, dass der Fruchtsaftgehalt exakt 50% sei. Sie wissen, dass der Fruchtsaftgehalt durch eine normalverteilte Zufallsvariable X mit unbekanntem Erwartungswert und bekannter Standardabweichung $\sigma = 6\%$ beschrieben werden kann. Sie bezweifeln die Behauptung auf der Verpackung und möchten diese mit einem statistischen Hypothesentest überprüfen. Auf Basis einer *i.i.d.*-Zufallsstichprobe berechnen Sie dazu das folgende zentrale 99%-Konfidenzintervall für μ:

$$KI_{0.99} = [44.81\%, 49.19\%]$$

Wie lautet die korrekte Testentscheidung auf Basis des realisierten Konfidenzintervalls?

A Da $50\% \notin KI_{0.99}$, wird H_0 auf dem 99%-Niveau verworfen. Die Behauptung auf der Verpackung steht somit im Widerspruch zu den Daten.
B Da $50\% \notin KI_{0.99}$, wird H_0 auf dem 99%-Niveau nicht verworfen. Die Behauptung auf der Verpackung steht somit nicht im Widerspruch zu den Daten.
C Da $50\% \notin KI_{0.99}$, wird H_0 auf dem 99%-Niveau verworfen. Die Behauptung auf der Verpackung steht somit nicht im Widerspruch zu den Daten.
D Da $50\% \in KI_{0.99}$, wird H_0 auf dem 99%-Niveau nicht verworfen. Die Behauptung auf der Verpackung steht somit nicht im Widerspruch zu den Daten.
E Da $50\% \in KI_{0.99}$, wird H_0 auf dem 99%-Niveau verworfen. Die Behauptung auf der Verpackung steht somit im Widerspruch zu den Daten.

1.6 Die Vorbereitungszeit in Stunden auf eine Statistik-Klausur sei normalverteilt mit einem Erwartungswert von $45h$ und einer Varianz von $15h^2$. Wie hoch ist die Wahrscheinlichkeit dafür, dass die Vorbereitungszeit mindestens $42h$, aber höchstens $48h$ beträgt?

A 0.5588
B 0.1586
C 0.2381
D 0.6127
E 0.4269

1.7 Eine Urne enthält 20 schwarze und 80 weiße Kugeln. Es wird nun eine Stichprobe vom Umfang $n = 15$ ohne Zurücklegen gezogen. Wie groß ist die Wahrscheinlichkeit dafür, dass sich exakt 5 schwarze Kugeln in der Stichprobe befinden?

A 0.1008
B 0.0046
C 0.2163
D 0.1451
E 0.0151

1.8 Gegeben sei der Maximum-Likelihood-Schätzer für den Parameter λ einer Poisson-verteilten Zufallsvariable X:

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

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

A Der Schätzer ist erwartungstreu.
B Der Bias des Schätzers beträgt 0.
C Der Schätzer ist konsistent.
D Der Stichprobenumfang hat Einfluss auf die Varianz des Schätzers.
E Weil der Schätzer unverzerrt ist, entspricht der realisierte Schätzwert stets dem wahren Parameterwert.

Ein 6-seitiger Würfel wird ein Mal geworfen. Der Würfel ist allerdings nicht fair, sondern gezinkt, sodass die 6 möglichen Ergebnisse nicht gleich wahrscheinlich sind. Die Zufallsvariable X gibt die gewürfelte Augenzahl an. Die folgende Tabelle stellt Ihnen die Wahrscheinlichkeitsfunktion von X dar:

Geworfene Augenzahl x	1	2	3	4	5	6
Wahrscheinlichkeit $P(X = x)$	0.05	0.15	0.20	0.10	0.15	0.35

1.9 Wie groß ist die Wahrscheinlichkeit dafür, dass die Augenzahl mindestens 2, aber höchstens 5 beträgt?

A 0.60
B 0.45
C 0.30
D 0.35
E 0.65

1.10 Sie interessieren sich für den Zusammenhang zwischen Fußball- und Bierkonsum und betrachten dazu die folgenden Ereignisse:

A: "Die Person schaut häufig Fußball."
B: "Die Person liebt Bier."

Zudem seien folgende Wahrscheinlichkeiten bekannt:

$P(A) = 0.18$, $P(B) = 0.31$, $P(\overline{A \cup B}) = 0.21$

Durch welchen Ausdruck und durch welchen Wert ist die Wahrscheinlichkeit gegeben, dass eine Person nur selten Fußball schaut, unter der Bedingung, dass sie kein Bier mag?

A $P(\overline{A}|\overline{B}) = 0.3043$
B $P(\overline{B}|\overline{A}) = 0.2561$
C $P(\overline{B}|\overline{A}) = 0.1816$
D $P(\overline{A}|\overline{B}) = 0.6774$
E $P(A|\overline{B}) = 0.2125$

Betrachten Sie die Zufallsvariable Y mit folgender Dichtefunktion:

$$f(y) = \begin{cases} y - 0.5 & \text{für } 1 \leq y \leq 2 \\ 0 & \text{sonst} \end{cases}$$

1.11 Welcher Wert entspricht dem Erwartungswert von Y?

A 1.500
B 1.000
C 2.500
D 1.583
E 0.833

X sei eine geometrisch verteilte Zufallsvariable mit unbekanntem Parameter p. Gegeben sei die zugehörige Loglikelihoodfunktion:

$$\ln(L(p)) = n \cdot \ln(p) + \ln(1-p)\left(\sum_{i=1}^{n} X_i - n\right)$$

1.12 Welcher Ausdruck entspricht dem Maximum-Likelihood Schätzer $\hat{p}_{ML}$?

A $1/\overline{X}$
B $\overline{X}$
C $n\overline{X}$
D $e^{\overline{X}}$
E $\sum_{i=1}^{n} X_i$

1.13 Gegeben sei eine unabhängig und identisch verteilte Zufallsstichprobe $X_1, ..., X_n$, auf deren Basis Sie das Stichprobenmittel $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ berechnen. Welche der folgenden Aussagen ist **nicht** korrekt?

A Für sehr große n sind die einzelnen Stichprobenvariablen normalverteilt.
B $\overline{X}$ ist für sehr große n symmetrisch verteilt.
C Die Standardabweichung von $\overline{X}$ nimmt mit zunehmendem n ab.
D Die Varianz von $\overline{X}$ konvergiert für $n \to \infty$ gegen Null.
E Für sehr große n konvergiert die empirische Verteilungsfunktion gegen die theoretische Verteilungsfunktion.

1.14 Am Notruftelefon einer Nürnberger Polizeiinspektion gehen im Durchschnitt 3 Notrufe pro Stunde ein. Die Beamten betreuen das Notruftelefon im Schichtbetrieb, wobei eine Telefonschicht 4 Stunden lang ist. Welchem Verteilungsmodell folgt die Zufallsvariable N "Anzahl der eingehenden Notrufe innerhalb einer Telefonschicht"?

A Poisson-Verteilung: $N \sim P(\lambda = 12)$
B Poisson-Verteilung: $N \sim P(\lambda = 3)$
C Binomial-Verteilung: $N \sim B(n = 4, p = 0.3)$
D Poisson-Verteilung: $N \sim P(\lambda = 4)$
E Binomial-Verteilung: $N \sim B(n = 12, p = 0.25)$

Im Rahmen des Qualitätsmanagements werden $n = 40$ zufällig ausgewählte Gäste eines Restaurants am Ende ihres Besuchs zu ihrer Zufriedenheit auf einer Skala von 1 (unzufrieden) bis 5 (sehr zufrieden) befragt (Zufallsvariable X).

Die Geschäftsführerin vermutet, dass die Zufriedenheit der befragten Gäste nicht unabhängig davon ist, ob diese zuvor ein Dessert gegessen haben oder nicht (Zufallsvariable Y).

Die Geschäftsführerin möchte daher die Unabhängigkeit der Zufallsvariablen X und Y durch einen Unabhängigkeitstest mit einem Signifikanzniveau $\alpha = 0.05$ überprüfen.

1.15 Welcher Verteilung folgt die Prüfgröße unter der Nullhypothese?

A χ^2_4
B χ^2_{10}
C t_{39}
D χ^2_{40}
E t_{40}

1.16 Die Berechnung der Geschäftsführerin ergibt eine Prüfgröße von 17.68. Die kritische Schranke liegt bei 14.86. Treffen Sie die korrekte Testentscheidung auf Basis dieser Informationen.

A Da die Prüfgröße im kritischen Bereich liegt, wird die Nullhypothese abgelehnt.
B Da die kritische Schranke die Prüfgröße übersteigt, kann die Nullhypothese nicht abgelehnt werden.
C Da die Prüfgröße außerhalb des kritischen Bereichs liegt, wird die Nullhypothese abgelehnt.
D Da 17.68 > 14.86, kann die Nullhypothese nicht abgelehnt werden.
E Um eine Testentscheidung treffen zu können, müsste noch die obere Grenze des kritischen Bereichs berechnet werden.

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.

Abgesehen von den in den einzelnen Aufgabenteilen explizit erwähnten Datenobjekten haben Sie keine weiteren Datenobjekte (z.B. Dataframes, Values oder Funktionen) abgespeichert. Sie haben das Paket *tidyverse* in Ihrer aktuellen Session bereits aktiviert.

2.1 Es sei X eine binomialverteilte Zufallsvariable mit Stichprobenumfang $n = 4$ und Erfolgswahrscheinlichkeit $p = 0.2$. Mit welchem Befehl können Sie die Wahrscheinlichkeit $P(X > 2)$ **nicht** berechnen?

A dbinom(3, size=4, prob=0.2) + dbinom(4, size=4, prob=0.2)
B pbinom(4, size=4, prob=0.2) - pbinom(2, size=4, prob=0.2)
C pbinom(2, size=4, prob=0.2)
D 1 - pbinom(2, size=4, prob=0.2)
E 1 - dbinom(0, size=4, prob=0.2) - dbinom(1, size=4, prob=0.2) - dbinom(2, size=4, prob=0.2)

2.2 Es sei X eine exponentialverteilte Zufallsvariable mit $\lambda = 2$. Vervollständigen Sie den Befehl

```
ggplot(data = data.frame(x = 0:4), aes(x = x)) +
   stat_function(fun = X, args = list(rate = 2), color = "blue") +
   stat_function(fun = Y, args = list(rate = 2), color = "red")
```

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

A **X**: dexp, **Y**: pexp
B **X**: pexp, **Y**: dexp
C **X**: dexp, **Y**: qexp
D **X**: qexp, **Y**: pexp
E **X**: rexp, **Y**: qexp

2.3 Sie ziehen eine Zufallsstichprobe der Größe $n = 50$ aus einer standardnormalverteilten Grundgesamtheit. Anschließend möchten Sie mit einem Hypothesentest auf dem Niveau $\alpha = 5\%$ überprüfen, ob die Nullhypothese $H_0 : \mu = 0$ für die vorliegende Stichprobe verworfen wird oder nicht. Welcher der folgenden Befehle liefert Ihnen eine Fehlermeldung, da ein zwingend notwendiger Befehlsinput nicht spezifiziert wurde?

A t.test(x = rnorm(n = 50, mean = 0, sd = 1), mu = 0, conf.level = 0.95)
B t.test(x = rnorm(n = 50, mean = 0, sd = 1), mu = 0)
C t.test(x = rnorm(n = 50, mean = 0, sd = 1), conf.level = 0.95)
D t.test(x = rnorm(n = 50, mean = 0, sd = 1))
E t.test(x = rnorm(mean = 0, sd = 1), mu = 0, conf.level = 0.95)

2.4 Sie ziehen eine Zufallsstichprobe der Größe $n = 100$ aus einer normalverteilten Grundgesamtheit mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2 und speichern diese in Ihrem Workspace unter dem Namen sample ab. Sie möchten nun auf Basis dieser Stichprobe ein zweiseitiges 99%-Konfidenzintervall für μ berechnen. Vervollständigen Sie den Befehl

c(mean(sample) - qt(p = **X**, df = **Y**) * sd(sample) / sqrt(**Z**),
mean(sample) + qt(p = **X**, df = **Y**) * sd(sample) / sqrt(**Z**))

so, dass die Grenzen des Konfidenzintervalls korrekt bestimmt werden.

A **X**: 0.995, **Y**: 99, **Z**: 100
B **X**: 0.995, **Y**: 100, **Z**: 100
C **X**: 0.995, **Y**: 99, **Z**: 99
D **X**: 0.975, **Y**: 99, **Z**: 100
E **X**: 0.975, **Y**: 100, **Z**: 100

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, 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
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