Klausur im SS 21, 15.07.2021

Stochastik
Angewandte Informatik (SPO3)

Prof. Dr. Barbara Staehle, HTWG Konstanz

Bearbeitungszeit: 60 Minuten

Hinweise:

- **Falls Sie für die Aufgaben alle Punkte haben wollen, begründen Sie Ihre Antworten, bzw. stellen Sie den Lösungs- / Rechenweg mittels Taschenrechner oder sonstigem Tool nachvollziehbar dar.**
- Runden Sie alle Ergebnisse auf **mindestens 3 Stellen** hinter dem Komma oder stellen Sie Ergebnisse als Bruch exakt dar.
- Lösen Sie, sofern möglich, die Aufgaben auf dem Angabenblatt.
- **Die Klausur enthält mehr Aufgaben, als Sie in der Bearbeitungszeit lösen können. Wählen Sie klug aus welche Aufgaben Sie lösen!** Sie müssen weder zum Bestehen noch für eine sehr gute Note alle Aufgaben korrekt bearbeiten. Zum Bestehen reichen ca. **35 Punkte**, eine sehr gute Note gibt es **ab ca. 70** Punkten.

Name: _____

Matrikelnummer: _____

Note: _____

Aufgabe	1	2	3	4	5	Σ
erreichbare Punkte	20	40	25	21	24	130
erreichte Punkte						

AUFGABE 1 WAHR ODER FALSCH?, 20 PUNKTE

Entscheiden Sie, ob die folgenden Aussagen wahr oder falsch sind. Begründen Sie Ihre Entscheidung (kurz).
Punktvergabe: w/f richtig: 1 Punkt; w/f richtig und Begründung sinnvoll: 2 Punkte

Aussage	wahr	falsch	kurze Begründung
(a) Wenn Mittelwerte, Varianzen und der (empirische) Korrelationskoeffizient zweier Merkmale vorliegen, ist eine Visualisierung der Daten unnötig.	☐	☐	
(b) Mit einem Histogramm kann man gut die empirische Verteilungsfunktion einer Stichprobe darstellen.	☐	☐	
(c) Der Interquartilabstand gibt die Länge des Intervalls an, in dem sich 25% aller Werte einer Stichprobe befinden.	☐	☐	
(d) Um das 0.5-Quantil einer Zufallsvariablen X mit Verteilungsfunktion $F(x)$ zu bestimmen, muss man den Wert der Umkehrfunktion von F an der Stelle 0.5 bestimmen.	☐	☐	
(e) Sie lesen, dass das durchschnittlich Einkommen eines Zürichers bei 7820 Franken liegt. Das heißt, die eine Hälfte der Bevölkerung verdient mehr, die andere weniger als diese Summe.	☐	☐	
(f) Für alle Ereignisräume Ω und alle Ereignisse $A, B \subseteq \Omega$ gilt, dass die bedingte Wahrscheinlichkeit eines Ereignisses immer kleiner ist, als die Wahrscheinlichkeit des Ereignisses selbst: $P(A\|B) < P(A)$.	☐	☐	
(g) Eine Bernoulli-Kette der Länge 100 besteht aus 100 unabhängig von einander unter identischen Bedingungen ausgeführten Bernoulli-Experimenten.	☐	☐	
(i) Alle Zufallsexperimente sind Laplace-Experimente.	☐	☐	
(j) Eine Bernoulli-verteilte Zufallsvariable ist auch Binomial-verteilt.	☐	☐	
(k) Kein Ereignis kann mit 110% Wahrscheinlichkeit eintreten.	☐	☐	

AUFGABE 2 STATISTIK

Alice arbeitet für den Webhoster Host123 und ist dafür zuständig, den automatischen Build-Prozess für Kunden-Webapplikationen zu optimieren. Hierfür analysiert sie in 7 zufällig ausgewählten Kundenrepositories die Anzahl der in den am häufigsten genutzten Programmiersprachen Java, Python, Ruby gespeicherten Dateien.

Ihre Ergebnisse fasst sie in der folgenden Tabelle zusammen:

Repository	R_1	R_2	R_3	R_4	R_5	R_6	R_7
Java	13	11	12	6	18	25	13
Python	13	11	10	3	15	19	7
Ruby	9	10	8	7	7	5	5

TEILAUFGABE 2.1 21 PUNKTE

Berechnen Sie **für die Anzahl der Java-Dateien** (Zeile 1 der Tabelle) die im Folgenden gesuchten Größen. Stellen Sie Ihren Rechenweg nachvollziehbar dar, bzw. geben Sie die Verwendung des Taschenrechners (TR), bzw. des genutzten Tools an!

a) (2 Punkte) Modalwert

b) (2 Punkte) Arithmetisches Mittel

c) (2 Punkte) Median

d) (2 Punkte) 90%-Quantil

e) (3 Punkte) empirische Standardabweichung

f) (3 Punkt) Interquartilsabstand

g) (2 Punkte) Spannweite (Range)

Verwenden Sie erwartungstreue, konsistente Schätzfunktionen, um basierend auf der obigen Stichprobe, Schätzwerte für die Anzahl der Java-Dateien pro Repository zu berechnen für

h) (2 Punkte) den wahren Mittelwert μ

i) (3 Punkte) die wahre Varianz σ^2

Zur Erinnerung:

Repository	R_1	R_2	R_3	R_4	R_5	R_6	R_7
Java	13	11	12	6	18	25	13
Python	13	11	10	3	15	19	7
Ruby	9	10	8	7	7	5	5

TEILAUFGABE 2.2 13 PUNKTE

a) (4 Punkte) Finden Sie jeweils einen Unterschied, bzw. eine Gemeinsamkeit zwischen der Anzahl der Python- und der Python-Dateien und der Java und der Ruby- Dateien in den Repositories, indem Sie Ihre Berechnungen aus Aufgabe 2.1 mit den folgende Kennwerten für Java und Ruby vergleichen:

Sprache	Mittelwert	Median	Std	IQR	Range
Python	11.1	11.0	5.242	5.5	16.0
Ruby	7.3	7.0	1.890	2.5	5.0

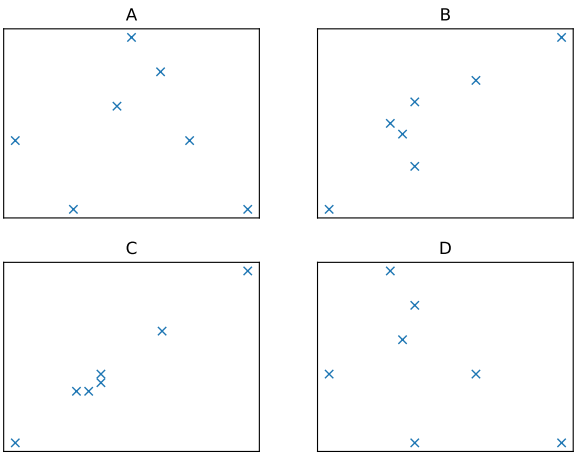

Abbildung 1: Welcher Scatterplot entspricht welcher Sprachkombination?

Zur Erinnerung:

Repository	R_1	R_2	R_3	R_4	R_5	R_6	R_7
Java	13	11	12	6	18	25	13
Python	13	11	10	3	15	19	7
Ruby	9	10	8	7	7	5	5

b) (4.5 Punkte) Die Scatter-Plots in Abbildung 1 illustrieren den Zusammenhang zwischen der Anzahl der Dateien in

(1) Java und Python	(2) Java und Ruby	(3) Python und Ruby

Geben Sie an, welches der Bilder (A,B,C,D) welchen Zusammenhang darstellt.

c) (4.5 Punkte) **Schätzen** Sie die empirischen Korrelationskoeffizienten $r_{J,P}$ zwischen der Anzahl der Dateien in Java und Python, Java und Ruby, sowie Python und Ruby.
Rechnen Sie diese nicht aus, sondern wählen Sie unter den folgenden Werten. Interpretieren Sie den Wert (was sagt dieser über den Zusammenhang der beiden Metriken aus?) Begründen Sie Ihre Meinung!

$r_{J,P}$	a) -0.485	b) 0.998	c) 0.911	d) -0.089
$r_{J,R}$	a) -0.485	b) 0.998	c) 0.911	d) -0.089
$r_{P,R}$	a) -0.485	b) 0.998	c) 0.911	d) -0.089

TEILAUFGABE 2.3 6 PUNKTE

Alice möchte für den wahren Mittelwerte der Anzahl der Python-Dateien pro Repository mit Hilfe ihrer Stichprobe der Größe $n = 7$, Mittelwert $\bar{x} = 11.143$ und empirischer Standardabweichung $s = 5.242$ (siehe Teilaufgabe 2.2) ein 90%-Konfidenzintervall angeben.

a) (2 Punkte) Geben Sie an, welches der richtige Ansatz ist, um das gesuchte Konfidenzintervall zu berechnen. Begründen Sie Ihre Meinung.

1) $[\bar{x} - z_{0.95} \cdot \frac{s}{\sqrt{n}}; \bar{x} + z_{0.95} \cdot \frac{s}{\sqrt{n}}]$
2) $[\bar{x} - z_{0.9} \cdot \frac{s}{\sqrt{n}}; \bar{x} + z_{0.9} \cdot \frac{s}{\sqrt{n}}]$
3) $[\bar{x} - t_{7;0.95} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{7;0.95} \cdot \frac{s}{\sqrt{n}}]$
4) $[\bar{x} - t_{7;0.9} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{6;0.9} \cdot \frac{s}{\sqrt{n}}]$
5) $[\bar{x} - t_{6;0.9} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{6;0.9} \cdot \frac{s}{\sqrt{n}}]$
6) $[\bar{x} - t_{6;0.95} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{6;0.95} \cdot \frac{s}{\sqrt{n}}]$

b) (2 Punkte) Mit Hilfe des richtigen Ansatzes berechnet sich das Konfidenzintervall für den wahren Erwartungswert der Anzahl der Python-Dateien pro Repository als $[7.293; 14.993]$. Was sagt dieses Intervall aus?

c) (2 Punkte) Sind Sie mit diesem Intervall zufrieden? Wie können Sie es ggf. vergrößern oder verkleinern?

Aufgabe 3 Wahrscheinlichkeitsrechnung

Teilaufgabe 3.1 10 Punkte

Auch Bob arbeitet beim Webhoster Host123 und ist für das Kunden-Management zuständig. Die gängige Praxis ist die, dass jede:r Entwickler:in einer Kundenfirma einen **Nutzernamen**, welcher **den ersten 12 Zeichen** des SHA1-Hashs ihres/seines vollen Namens entspricht, erhält. Zusätzlich zum Nutzernamen ist jede:r Nutzer:in eine **Commit-ID** für das Repository zugeordnet, welche aus den **ersten 6** Zeichen des Nutzernamens besteht.

Bob beobachtet, dass dieses Verfahren beim automatisierten Buildprozess manchmal für Probleme sorgt, weil es passieren kann, dass die Commit-ID nur aus Zahlen besteht, oder dass zwei Entwickler:innen die selbe Commit-ID zugeordnet bekommen.

Bob nimmt vereinfacht an, dass eine SHA1-Hash eine Hexadezimal-Zahl ist, also aus den Ziffern 0-9 sowie den Buchstaben a-f besteht, wobei die einzelnen Zeichen mehrmals vorkommen können. Dann überlegt er sich Folgendes:

a) (1 Punkt) Wie viele mögliche Nutzernamen N_N gibt es?

b) (1 Punkt) Wie viele mögliche Commit-IDs N_C gibt es?

c) (2 Punkte) Mit welcher Wahrscheinlichkeit $P(Z)$ besteht eine Commit-ID nur aus Ziffern?

d) (3 Punkte) Mit welcher Wahrscheinlichkeit $P(G)$ haben zwei Nutzer:innen die selbe Commit-ID?

e) (3 Punkte) Eine Umstellung des Verfahrens der Nutzernamen und Commit-ID Zuordnung ist zeitaufwändig und teuer, aber einige Kunden beschweren sich über Probleme beim Buildprozess.
Würden Sie an Bobs Stelle die gängige Praxis weiterlaufen lassen oder würden Sie eine andere Strategie wählen? Begründen Sie Ihre Meinung Hilfe Ihrer obigen Überlegungen.

TEILAUFGABE 3.2 15 PUNKTE

Charlie arbeitet ebenso für Host123 und ist zusätzlich Lehrbeauftragter für die Vorlesung Software-Engineering. Als Praxisbeispiel für seine Vorlesung zum Thema "Unit-Tests" befragt er einige Entwickler:innen, ob sie Unit-Tests verwenden oder nicht. Er findet heraus, dass der Anteil der Repositories mit Unit-Tests von der verwendeten Programmiersprache abhängt. Daher konzentriert er sich auf Repositories, die alle jeweils nur Code in Java, Python oder (XOR) Ruby enthalten. Für diese stellt er fest, dass

- 30% der Repositories mit Unit-Tests in Java, 50% in Python geschrieben sind,
- 80% der Repositories ohne Unit-Tests in Java, 10% in Python geschrieben sind,
- insgesamt 70% der Repositories in Java geschrieben sind.

Hinweis: Verwenden Sie zur Modellierung des Szenarios die folgenden Ereignisse:

- J, P, R : ein Repository enthält nur Code in Java, Python bzw. Ruby
- T : ein Repository enthält Unit-Tests

a) (5 Punkte) Modellieren Sie das beschriebene Szenario z.B. durch einen Wahrscheinlichkeitsbaum oder mit Hilfe einer Stichpunktliste. Geben Sie alle aus dem Text ableitbaren Wahrscheinlichkeiten der beschriebenen Ereignisse und ihrer jeweiligen Gegenereignisse an. Verwenden Sie ggf. die Variable x für nicht gegebene aber benötigte Wahrscheinlichkeiten.

b) (4 Punkte) Wie groß ist der Anteil der Repositories, die Unit-Tests enthalten?

c) (3 Punkte) Mit welcher Wahrscheinlichkeit ist ein zufällig ausgewähltes Repository ein Python-Repository und enthält Unit-Tests?

d) (3 Punkte) Mit welcher Wahrscheinlichkeit enthält ein Repository, von dem man weiß, dass es Unit-Tests enthält, nur Python-Dateien?

AUFGABE 4 ZUFALLSVARIABLEN I

TEILAUFGABE 4.1 21 PUNKTE

Dave und Eve möchten einige Funktionalitäten des Rollenspiels “Das schwarze Auge ” (DSA) als Handyapp implementieren. Im Verlauf eines DSA-Abenteuers müssen die Spieler (genannt **Helden**) oftmals sogenannte **Proben** bestehen, bei dem der Held seinen Würfelwurf mit einem bestimmten Wert vergleichen muss. Ein Gelingen der Probe sorgt dafür, dass das geschieht, was der Held beabsichtigt, ein Misslingen hingegen lässt ihn bei der Tätigkeit scheitern.

In einem echten DSA-Spiel werden 3 W20 verwendet, das ist Dave und Eve aber zu kompliziert. Dave verwendet daher für den Würfelwurf eines Heldens die **diskreten** Zufallsvariablen D_1, D_2, D_3, Eve verwendet die stetige Zufallsvariable E.

Konkret:

- Dave nimmt an, dass ein Held in jedem Versuch mit zwei fairen W4 (Würfel 1 und Würfel 2) würfelt. Die Ergebnisse von Würfel 1 und Würfel 2 werden addiert:
  - D_1, D_2 sind jeweils **diskrete ZV**, welche die Augenzahl von Würfel 1 bzw. 2 beschreiben und mit gleicher Wahrscheinlichkeit eine Zahl von 1-4 annehmen.
  - $D_3 = D_1 + D_2$ ist eine **diskrete ZV** und beschreibt die Zahl, die durch die Addition der Augenzahlen der beiden Würfeln entsteht.
- Eve erzeugt den Würfelwert eines Heldens mittels der **stetigen ZV** E die reelle Zahlen im Intervall $[2;8]$ annimmt. Die Verteilungsdichte- und Verteilsungsfunktion sind durch die Funktionen f_E und F_E gegeben.

Verteilungsdichtefunktion von E:

$$f_E(x) = \begin{cases} 0 & \text{für } x < 2 \\ \frac{1}{90}(4x+1) & \text{für } 2 \le x \le 5 \\ \frac{1}{90}(41-4x) & \text{für } 5 \le x \le 8 \\ 0 & \text{für } x > 8 \end{cases}$$

Verteilungsfunktion von E:

$$F_E(x) = \begin{cases} 0 & \text{für } x < 2 \\ \frac{1}{90}(2x^2+x-10) & \text{für } 2 \le x \le 5 \\ \frac{1}{2} + \frac{1}{90}(41x-2x^2-155) & \text{für } 5 \le x \le 8 \\ 1 & \text{für } x > 8 \end{cases}$$

a) (5 Punkte) Geben Sie für D_1 und D_3 jeweils die Wahrscheinlichkeitsverteilung an.

b) (2 Punkte) Dave findet die von Eve ermittelte Verteilungsfunktion unglaubwürdig. Wie kann Eve ihm beweisen, dass die Verteilungsfunktion korrekt ist?

c) (4 Punkte) Stellen Sie in einem gemeinsamen Diagramm die Wahrscheinlichkeitsverteilungsfunktion von D_3 und E dar. Achten Sie auf korrekte Beschriftungen. Ihr Diagramm muss **nicht exakt** sein, soll aber den Unterschied zwischen den Verteilungsfunktionen klar machen.

d) (4 Punkte) Berechnen Sie für die Zufallsvariablen D_3 und E jeweils die Wahrscheinlichkeit, dass ein Held eine Würfelprobe zum Wert 5 besteht, dass der Wert von D_3 und E also höchstens 5 ist.

e) (6 Punkte) Berechnen Sie den Erwartungswert der Zufallsvariablen D_1, D_3 und E. **Vermeiden Sie lange Rechnungen,** indem Sie den Erwartungswert von D_3 und E abschätzen und Ihre Schätzung begründen, bzw. Ihr Wissen über Zufallsvariablen nutzen!

AUFGABE 5 ZUFALLSVARIABLEN II

Auch Frank, Grace und Henry arbeiten daran einige Funktionalitäten von DSA als Handyapp zu implementieren. Sie haben sich entschieden, an der Simulation der Zauberin Evanora mitzuarbeiten.

TEILAUFGABE 5.1 6 PUNKTE

Frank konzentriert sich auf die zeitlichen Abläufe. Vereinfacht gesagt arbeitet Evanora so, dass sie zu zufälligen Zeitpunkten Besuch von Helden empfängt. Evanora berät diese und heilt das Problem dann mit einer zufälligen Menge Zaubertrank. Um dies abzubilden, verwendet Frank folgende Zufallsvariablen:

- Z: Zeit, die zwischen dem Eintreffen zweier Helden bei Evanora liegt, ist exponentialverteilt mit Erwartungswert $\mathrm{E}(Z) = \frac{1}{5}$ [Stunden].
- A: die Anzahl der Helden, die Evanora pro Stunde besuchen ist entsprechend eine Poisson-verteilte Zufallsvariable mit Rate $\lambda_A = 5$ [pro Stunde].

a) (2 Punkte) Mit welcher Wahrscheinlichkeit kommen zwei Besuchern im Abstand von genau $\frac{1}{5}$ Stunden zu Evanora?

b) (2 Punkte) Mit welcher Wahrscheinlichkeit kommen zwei Besucher im Abstand von weniger als $\frac{1}{5}$ Stunden zu Evanora?

c) (2 Punkte) Mit welcher Wahrscheinlichkeit ist die Anzahl der Besucher, die Evanora pro Stunde besuchen genau gleich 5?

TEILAUFGABE 5.2 11 PUNKTE

Grace konzentriert sich auf die Menge des Zaubertranks die ein Besucher von Evanora bekommt. Hierzu verwendet sie die Zufallsvariable

- M: Menge des Zaubertrank, die ein Besucher bekommt ist normalverteilt mit Erwartungswert 110 und Standardabweichung 30 [ml].

a) (2 Punkte) Mit welcher Wahrscheinlichkeit bekommt ein Besucher mindestens 100 ml Zaubertrank?

b) (2 Punkte) Wie groß ist der Anteil der Besucher, die zwischen 80 und 140 ml Zaubertrank bekommen ungefähr?

c) (3 Punkte) Welche Mengen an Zaubertrank werden nur in 10 % aller Ausgabefälle überschritten?

d) (4 Punkte) In welchem, symmetrisch um den Erwartungswert gelegenen Intervall, liegt die Menge des ausgegeben Zaubertranks in 90 % der Fälle?

TEILAUFGABE 5.3 7 PUNKTE

Henry modelliert einen anderen Aspekt der Arbeitsweise von Evanora: Leider bezahlt ein Besucher nur mit einer Wahrscheinlichkeit von $p_1 = 0.75$ für seinen Zaubertrank, umgekehrt ist ein Zaubertrank aber auch nur mit einer Wahrscheinlichkeit von $p_2 = 0.85$ wirksam. Diese Wahrscheinlichkeiten sind für alle Besucher bzw. Tränke jeweils gleich und voneinander unabhängig. Die 12 ist Evanoras Glückszahl, daher macht sie nach einer Serie von 12 Besuchern immer eine kleine Pause.

Henry verwendet zur Beschreibung dieses Szenarios die folgenden Zufallsvariablen

- G: Anzahl der zahlenden Besucher in einer Serie von 12 Besuchern.
- W: Anzahl der wirksamen Zaubertränke in einer Serie von 12 ausgegebenen Zaubertränken.

a) (2 Punkte) Geben Sie an, wie und mit welchen Parametern die Zufallsvariablen W, G verteilt sind.

b) (2 Punkte) Wie groß ist die Wahrscheinlichkeit, dass unter 12 von Evanora ausgegebenen Zaubertränken weniger als 3 sind, die wirken?

c) (3 Punkte) Mit welcher Wahrscheinlichkeit liegt die Anzahl der zahlenden Besucher in einer Serie von 12 Besuchern zwischen 7 und 10 einschließlich?