Justus-Liebig-Universität Giessen 1. 8. 2024

Modulabschließende Prüfung zu

Grundlagen der Datenanalyse mit R

Module 07-M/BA-R1, 07-BDS-14, 07-BAI-13, 07-MDA-07 und 07-MDS-11

Sommersemester 2024

Bitte unbedingt beachten:

1. Diese Prüfung findet am 1. August 2024 von 9:00 Uhr bis 12:00 Uhr online oder in Präsenz im Hörsaal II der Physik im Heinrich-Buff-Ring 14 in Gießen statt. Es stehen Ihnen also **drei Zeitstunden** zur Bearbeitung der Prüfungsaufgaben zur Verfügung.

2. Zu Beginn der Prüfung finden Sie in Stud.IP im Ordner „Prüfungsaufgaben und -daten“ eine ZIP-komprimierte Datei mit einerseits einer PDF-Datei mit den Prüfungsaufgaben und andererseits sämtlichen der in den Aufgaben genannten Datensätzen. Ein Zugriff auf diesen Stud.IP-Ordner wird nur von 9:00 Uhr bis 9:10 Uhr möglich sein. Laden Sie daher die ZIP-komprimierte Datei sofort zu Beginn der Klausur in Ihr Arbeitsverzeichnis herunter und entpacken Sie sie dort!

3. Die Prüfungsaufgaben sind selbstständig zu lösen! Mit der Abgabe von Lösungsversuchen (hierzu siehe Punkt 7) erklären Sie automatisch und an Eides statt, dass Sie sie alleine und selbstständig erarbeitet haben.

4. Als Hilfsmittel für die Prüfung sind lediglich gestattet:
   a) Das Vorlesungsskript zu „Grundlagen der Datenanalyse mit R (R 1)“, die dazugehörigen Übungen und deren Lösungen samt Ihrer persönlicher Notizen (jeweils in elektronischer oder in Papierform) sowie
   b) ein privater Computer mit der installierten Software R in der Version 4.3.0 oder höher.

5. Zur Bearbeitung der Aufgaben ist die unter 4b genannte Software zu verwenden! (Sie dürfen auch eine geeignete IDE wie RStudio zum Einsatz bringen). Strukturieren Sie Ihren Programmcode übersichtlich mit kurzen Zeilen und versehen Sie ihn mit erläuternden Kommentaren!

6. Speichern Sie den Programmcode (einschließlich Ihrer Kommentare und eventueller Antworten) zu Ihrer Bearbeitung einer Aufgabe jeweils in einer **separaten** R-Datei (oder Rmd-Datei), d. h., behandeln Sie in einer solchen Datei nur eine Aufgabe.

   Geben Sie in jeder solchen Datei in der ersten Zeile Ihren eigenen Nach- und Vornamen sowie Ihre eigene Matrikelnummer (sofern vorhanden) an!

   Und benennen Sie diese Dateien **unbedingt** mit Ihrem Nach- und Vornamen sowie mit der Nummer der bearbeiteten Aufgabe analog zu folgendem Beispiel: **„MustermannErika1.R“**. (Bitte beachten: *Sie* heißen in der Regel *nicht* Erika Mustermann!)

   Tipp: **Speichern Sie** den Programmcode während der Bearbeitung sicherheitshalber **häufig und regelmäßig!**

7. Packen Sie **vor dem Ende der Bearbeitungszeit** alle Ihre R-Dateien, die den Programmcode Ihrer Bearbeitungen enthalten (nicht aber Ausgaberesultate wie Grafiken), **zusammen in einen einzelnen** ZIP-komprimierten Ordner und laden Sie diesen in Stud.IP in den dafür vorgesehenen Ordner „Prüfungsbearbeitungen“ dieser Veranstaltung hoch. (Dieser Ordner wird nach dem Ende der Bearbeitungszeit umgehend geschlossen.)

8. Bei technischen Problemen oder anderen schwerwiegenden Schwierigkeiten während der Prüfung melden Sie sich bitte umgehend bei Dr. Gerrit Eichner unter gerrit.eichner@math.uni-giessen.de oder Tel. +49 (0)641 99 32104.

9. Viel Erfolg!

Name und Matrikelnummer: .....

Aufgabe	1	2	3	4	5	6	7	Σ	%	Note:
Erreichb. Pkt.	7	5	4	6	7	17	14	60	120	
Err. Pkt.										

P. _____

. .2024, _____

1. Manche Hausratversicherungspolicen verlangen einen Schätzwert für die Wiederbeschaffungskosten des zu versichernden Hausrates. In einem Versuch, diesen Schätzwert für einen Buchbestand von insgesamt 1554 Büchern zu ermitteln, indem man lediglich eine Zufallsstichprobe von 100 Exemplaren (statt aller) genauer analysierte, wurde neben den laut Verlagskatalogen und anderen Quellen gültigen Wiederbeschaffungspreisen (in Britischen Pence) auch die Dicke eines jeden Buches in Millimeter bestimmt. Die erhobenen Daten finden Sie in der Datei `Buecherwerte.csv`.

   a) Lesen Sie mit `read.csv` die Daten in R ein und kontrollieren Sie Struktur und Inhalt der eingelesenen Daten! *(2 Punkte)*

   b) Rechnen Sie allein aus den Buchpreisen der Stichprobe die Wiederbeschaffungskosten für den Gesamtbestand hoch! *(1 Punkt)*

   c) Bestimmen Sie nun den durchschnittlichen Wiederbeschaffungspreis pro Millimeter für die Stichprobe und rechnen Sie damit die Wiederbeschaffungskosten für den Gesamtbestand, dessen Gesamtdicke 25182 Millimeter betrug, hoch! *(1 Punkt)*

   d) Fertigen Sie ein Streudiagramm der Buchpreise gegen die Buchdicken an! *(1 Punkt)*

   e) Die Erhebung der Buchdicke war in der Hoffnung geschehen, die Schätzung der Buchpreise mithilfe dieser „Ersatzvariablen“ vereinfachen zu können. In Anbetracht des eben erzeugten Streudiagrammes: Ließe sich aus der Dicke (in mm) eines Buches – wenigstens tendenziell – sein Wiederbeschaffungspreis schätzen? Wenn ja, hielten Sie diesen Schätzwert für zuverlässig, und wenn nein, warum nicht? *(2 Punkte)*

2. In die PatientInnendatenbank einer medizinischen Praxis haben sich im Laufe der Jahre (z. B. wegen fehlerhafter Namensschreibung) sogenannte Dubletten (= Mehrfacheintragungen) eingeschlichen. D. h., es befinden sich möglicherweise zahlreiche doppelte, dreifache und potenziell noch erheblich höhere Vielfachheiten an Eintragungen darin. In der Datei `Geburtsdaten.txt` finden Sie die Geburtsdaten aller Einträge der Datenbank. Es soll nun für diese Geburtsdaten herausgefunden werden, wie oft eine jede Vielfachheit (≥ 2) unter ihnen auftritt. Dabei ist weder gefragt, wie sie aussehen, noch wo sie sind.

   a) Lesen Sie mit `scan` die Geburtsdaten (als `character`) in R ein und kontrollieren Sie Struktur der eingelesenen Daten! *(2 Punkte)*

   b) Tabellieren Sie (ohne die Verwendung irgendeiner Schleife), wie häufig jede Vielfachheit (≥ 2) auftritt! *(3 Punkte)*

3. Ein Produktionsverfahren für ein Werkstück liefere angeblich höchstens 3 % Ausschuss. Es werden 100 unabhängig voneinander damit hergestellte Werkstücke der Qualitätskontrolle unterzogen. Fünf der 100 Stücke ergeben sich zu Ausschuss.

a) Wieviel Stücke Ausschuss wären zu erwarten? *(1 Punkt)*

b) Spricht die Beobachtung von fünf Stücken Ausschuss gegen die angebliche Schwelle von 3 %? Begründen Sie Ihre Antwort und ziehen Sie dazu die Wahrscheinlichkeit heran, mit der ein mindestens so schlechtes Ergebnis in dem Produktionsverfahren zu beobachten wäre, wenn die Ausschussquote tatsächlich 3 % betrüge! *(3 Punkte)*

4. Betrachten Sie die folgende (4×4)-Matrix:

$$H := \begin{pmatrix} \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix}$$

a) Ermitteln Sie – ohne R! – die (sehr einfache) Formel, mit der die Elemente $H_{i,j}$ obiger Matrix in Abhängigkeit ihrer Zeilen- und Spaltenindices i bzw. j berechnet werden! *(2 Punkte)*

b) Verwenden Sie lediglich die Funktionen `matrix`, `seq` (bzw. „`:`“), `rep`, `col` und/oder `row` sowie elementare Arithmetik, um die obige (4×4)-Matrix mit R zu reproduzieren (ohne dass Sie diese Matrix „einfach eintippen“)! *(3 Punkte)*

c) Verallgemeinern Sie nun Ihren Programmcode zu einer Funktion mit einem Argument `n`, die zu diesem n die entsprechende $(n \times n)$-Matrix erzeugt! *(1 Punkt)*

5. Es werde angenommen, dass der Erwartungswert a der aufzuwendenden Arbeitszeit in Personenstunden für die Produktion von $n \in \mathbb{N}$ Werkstücken einer gewissen Sorte durch die Beziehung $a(n) = 10 + 2 \log n$ beschrieben wird. Die tatsächlich zu beobachtenden Arbeitszeiten mögen von dieser Beziehung durch normalverteilte Störungen abweichen, sodass in einem Experiment mit $k = 30$ Versuchsläufen bei verschiedenen Werkstückezahlen $n_1, \ldots, n_k \in [1, 100]$ faktisch die Arbeitszeiten $A_i = 10 + 2 \log n_i + \varepsilon_i$ für $i = 1, \ldots, k$ mit unabhängigen und identisch normalverteilten ε_i mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1/2$ beobachtet werden.

a) Wählen Sie „vernünftige“ Werte für die n_i und erzeugen Sie reproduzierbar die dazugehörigen A_i! *(4 Punkte)*

b) Fertigen Sie eine Grafik des Graphen der oben beschriebenen Beziehung zwischen a und n an (aber für reellwertige $n \geq 1$) und überlagern Sie ihr das Streudiagramm der in Teil a erzeugten Wertepaare (n_i, A_i) so, dass sie alle sichtbar sind! *(3 Punkte)*

6. In einer Spinnerei werden maschinell Wollknäuel mit einem Sollgewicht von 50 g gewickelt. Es wurden 20 zufällig ausgewählte Knäuel gewogen. Die ermittelten Gewichte in g finden Sie in der Datei `Wollknaeuelgewichte2024.txt`.

a) Lesen Sie mit `scan` Gewichtsdaten in R ein und kontrollieren Sie Struktur der eingelesenen Daten! *(2 Punkte)*

b) Der Hersteller der Wickelmaschine hat angegeben, dass die Wickelgewichte normalverteilt seien. Überprüfen Sie diese Aussage mit einem geeigneten explorativen Werkzeug! Was halten Sie von der Aussage? (3 Punkte)

c) Es besteht der Verdacht, dass das mediane Wickelgewicht des Produktionsprozesses vom gewünschten Sollgewicht abweicht. Es ist Ihre Aufgabe, datenanalytisch zu untersuchen, ob dieser Verdacht gerechtfertigt ist. Wählen und nennen Sie ein dazu geeignetes statistisches Verfahren und seine Voraussetzungen! Sind die Voraussetzungen für die vorliegenden Daten erfüllt bzw. überhaupt überprüfbar? (5 Punkte)

d) Formalisieren Sie die beiden Aussagen, zwischen denen in dem von Ihnen in Teil c gewählten Verfahren zu entscheiden ist, und benennen Sie sie unter Verwendung der üblichen Fachausdrücke! Legen Sie sodann den Wert einer entscheidenden Größe fest und benennen Sie auch sie mit dem üblichen Fachausdruck! Wie lautet die zugehörige Entscheidungsregel? (5 Punkte)

e) Führen Sie schließlich die Datenanalyse durch und teilen Sie Ihre Entscheidung mit! (2 Punkte)

7. Die menschliche Lunge besteht aus fünf sogenannten Lappen: dem linken unteren, linken oberen, rechten unteren, rechten mittleren und dem rechten oberen Lungenlappen. Zur Beurteilung des Status einer Lunge ist u. a. die lokale Gewebedichte von Interesse, die in sogenannten "Hounsfield Units", kurz HU, gemessen wird.

   Die Datei `Lungenlappendichte.csv` enthält die Absoluthäufigkeitsverteilungen der Dichtewerte in den Lungenlappen eines lungengesunden Probanden, die in einem sehr fein gerasterten CT-Bild seiner Lunge registriert wurden. (Die HU-Skala erstreckt sich hier in Einserschritten von -1024 bis 3050.) Sie enthält in ihren Spalten `LU`, `LO`, `RU`, `RM` und `RO` die Absoluthäufigkeiten, mit denen die in der Spalte `HUDichte` stehenden HU-Werte im Bild des **l**inken **u**nteren, **l**inken **o**beren, **r**echten **u**nteren, **r**echten **m**ittleren bzw. **r**echten **o**beren Lungenlappens jeweils beobachtet wurden.

   a) Lesen Sie die Daten in `Lungenlappendichte.csv` mithilfe von `read.csv2` in einen Data Frame ein und kontrollieren Sie dessen Struktur und Inhalt! *(2 Punkte)*

   b) Leiten Sie eine neue Spalte `Gesamt` ab, die für jeden HU-Wert seine über die fünf Lungenlappen hinweg kumulierte Gesamthäufigkeit enthält! *(2 Punkte)*

   c) Die Auflösung der HU-Skala soll vergröbert werden, damit die Häufigkeitsverteilung der in Spalte `Gesamt` stehenden Daten überhaupt einigermaßen übersichtlich wird. Leiten Sie zu diesem Zweck mithilfe der Funktion `cut` zunächst einen Faktorvektor aus der Spalte `HUDichte` ab, wobei Sie ihren Wertebereich von -1100 bis 3100 in Intervalle der Breite 100 einteilen! Sorgen Sie dafür, dass der Faktorvektor den adäquaten Modus einer ordinal skalierten Variablen hat! *(2 Punkte)*

   Nutzen Sie dann jenen Faktorvektor, um die Elemente der Spalte `Gesamt` entsprechend gruppiert zu summieren! (Tipp: `tapply` in Verbindung mit `sum`.) *(1 Punkt)*

   d) Fertigen Sie schließlich für die „gruppierte" Häufigkeitsverteilung von eben ein Säulendiagramm mit *horizontalen* Säulen an (also ein Balkendiagramm)! Sorgen Sie mit den Argumenten `las` und `mar` der Funktion `par` dafür, dass die Beschriftung beider Achsen

horizontal ist und dass insbesondere für die Beschriftung der vertikalen Achse genügend Platz ist! Charakterisieren Sie kurz die dargestellte Verteilung! *(3 Punkte)*

e) Rekonstruieren Sie aus den vorliegenden Häufigkeiten in `Gesamt` und den dazugehörigen Werten in `HUDichte` die ursprünglichen Rohdaten! (Es sollte ein Vektor mit mehr als 3,6 Mio. Elementen sein.) *(1 Punkt)*

f) Einen Normal-QQ-Plot für eine derart große Stichprobe wie die in Teil e rekonstruierte anzufertigen, ist aufgrund der Rechenzeit sinnlos. Alternativ kann man eine (oder mehrere) hinreichend große, zufällige Teilstichprobe(n) aus der Stichprobe ziehen und für diese die Zulässigkeit der Normalverteilungsannahme prüfen. Ziehen Sie mithilfe der Funktion `sample` (siehe ihre Hilfeseite!) reproduzierbar (!) etwa 0,5 % der Stichprobe von Teil e als zufällige Teilstichprobe und beurteilen Sie dafür, ob die Normalverteilungsannahme haltbar ist! *(3 Punkte)*