Praktikum zur Vorlesung «Grundlagen der Statistik»

WS 2022/23
Prof. Dr. M. Schott
(Teilweise übernommen von Prof. Dr. D. Heider)

Allgemeine Hinweise

Das Praktikum findet in vom **15.02.2023** bis zum **24.02.2023** statt. Persönlich werden wir im Raum 03A14 (HS III A3) in der Hans-Meerwein-Straße 6 zur Verfügung stehen. An den ersten Tagen sind Aufgaben zu lösen, deren Lösungen in einem Praktikumsbericht im Umfang von etwa 10-15 Seiten dokumentiert werden sollen. Natürlich müssen Sie die Aufgaben alleine lösen, jedoch stehen wir während dieser Zeit für Rückfragen bereit. Insbesondere sollten hier alle erstellten Graphen abgebildet und erläutert werden. Im Anhang des Berichts müssen alle R-Befehle zu den Übungsaufgaben dokumentiert sein. Die Berichte müssen einzeln angefertigt werden. Bei Plagiaten werden die betroffenen Berichte als nicht bestanden gewertet.

An den letzten beiden Tagen müssen Sie in einer 5-minütigen Präsentation ihre Lösung der Aufgabe 8 vorstellen.

Vorrausetzungen zum Bestehen:

- Abgabe des Praktikumberichts
- Anwesenheit (online bzw. in Person) an allen drei Tagen für mindestens eine Stunde. Sollten Sie gute Gründe haben, warum Sie nicht persönlich anwesend sein können, sprechen Sie dies bitte vorher mit den Übnungsleitern bzw. Prof. Schott ab.
- Präsentation der Aufgabe 8 (Zeitplan wird während des Praktikums besprochen)

Tag 1: Daten und Plots

Aufgabe 1: Datensatz OECD

Den folgenden Aufgaben liegt der Datensatz OECD zu Grunde, er enthält Variablen (Stand 2009), die das Wohlergehen von Kindern in den Mitgliedstaaten der OECD messen sollen. Abgefragt wurde:

- Einkommen: das durchschnittliche Einkommen der Eltern [in tausend US Dollar pro Kind]
- Armut: der Anteil [immer in Prozent] an Kindern in einem armen Elternhaus
- Bildung: der Anteil an Kindern, die ohne eine Grundausstattung (Bücher, Schreibtisch, Computer, Internet) für Bildung auskommen
- WenigRaum: der Anteil an Kindern, die auf zu wenig Raum wohnen
- Umwelt: der Anteil an Kindern, die unter schlechten Umweltbedingungen leben
- Lesen: mittlerer PISA-Score zur Lesefähigkeit
- Geburtsgewicht: der Anteil an Kindern, die bei der Geburt weniger als 2.5 kg wiegen
- Säuglsterblichkeit: Säuglingssterblichkeit (< 1 Jahr) [x in Tausend]
- Sterblichkeit: Sterblichkeit (< 20 Jahre) [x in 100 000]
- Selbstmord: Selbstmord von Jugendlichen im Alter von 15 bis 19 [x in 100 000]
- Bewegung: der Anteil an 11, 13 und 15 jährigen Jugendlichen, die sich regelmäßig bewegen
- Rauchen: der Anteil an 15 jährigen Jugendlichen, die mindestens einmal die Woche rauchen
- Alkohol: der Anteil an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren
- Bullying: der Anteil an Kindern, die angeben in der Schule bedroht zu werden
- Schule: der Anteil an Kindern, die angeben die Schule zu mögen
- Geo: Faktor, der die geographische Lage eines Landes beschreibt, dabei steht E für Europa und R für den Rest der Welt.

Lösen Sie folgende Aufgaben:

a) Laden Sie sich die Datei oecdM.csv herunter.

b) Berechnen Sie die Mittelwerte und Varianzen der einzelnen Variablen mit dem geeigneten apply Befehl.

c) Überprüfen Sie, ob die Niederlande in der Länderliste des Datensatze auftaucht. Gibt es auch einen Eintrag für China? (Benutzen Sie die R-Hilfe, um herauszufinden, wie man auf die Ländernamen zugreifen kann.)

d) In welchem Land waren die meisten Jugendlichen mindestens zweimal betrunken? Wie hoch ist der maximale Prozentsatz?

e) In welchem Land ist die Säuglingssterblichkeit am geringsten? Wie hoch ist sie in diesem Land?

f) In welchen Ländern ist der Prozentsatz an Jugendlichen, die sich regelmäßig bewegen, kleiner als der Durchschnitt?

Aufgabe 2: Häufgkeiten und Stripcharts

a) Wieviele Länder im Datensatz oecdM gehören zu Europa, wieviele zum Rest der Welt? Stellen Sie das Ergebnis in einem Kuchendiagramm dar und verwenden Sie dazu die Farben grün (green) und blau (blue).

b) Visualisieren Sie die Variable Lesen, getrennt nach dem Faktor Geo, in einem vertikalen Stripchart. Welche Aussage können Sie mit diesem Stripchart treffen?

Aufgabe 3: Quantile und Plots

a) Erstellen Sie einen Boxplot für die Variable “Bildung”. Was fällt Ihnen auf?

b) Untermauern Sie die Beobachtung aus Aufgabe (a) durch Berechnung einiger Quantile mit Hilfe der Funktion quantile().

c) Stellen Sie zudem die aufsteigend geordneten Werte der Variable Bildung mit Hilfe der Funktion plot() als Kurve dar.

d) Begründen Sie anhand Ihrer Beobachtungen, dass das 75% Quantil der Daten einen guten Trennpunkt zwischen Ländern mit “guter” und “schlechter” Grundausstattung für Bildung darstellt.

Tag 2: Statistische Test, Regression, ANOVA

Aufgabe 4: Annahmen des t-Tests

a) Ziehen Sie 100 exponential-verteilte Zufallszahlen mit dem Parameter $\lambda = 0.1$ und speichern Sie diese in dem Objekt X1. Erstellen Sie analog ein Objekt X2 von 100 Zufallszahlen, für die Sie erneut 100 exponential-verteilte Zufallszahlen HX2 (H wie Hilfsvektor) mit dem Parameter $\lambda = 0.1$ ziehen und anschließend alle Elemente von 20 subtrahieren. Ein Element i des Vektors X2 berechnet sich also als X2[i] = 20−X2[i].

b) Führen Sie den t−Test durch, um zu untersuchen, ob die beiden Objekte unterschied- liche Mittelwerte besitzen.

c) Führen Sie den Wilcoxon-Rangsummen-Test durch, um zu untersuchen, ob die beiden Objekte unterschiedliche Mediane besitzen.

Aufgabe 5: Testen an PISA-Daten

Der PISA-Test ist eine standardisierte Bewertung von (15-jährigen) Schülern unter den teil- nehmenden Staaten. Ziel der Regierungen ist, eine Datenbasis zur länderübergreifenden For- schung zu ermöglichen. Im Datensatz PISA.csv finden Sie die Ergebnisse einiger ausgewählter OECD-Staaten, getrennt nach dem Geschlecht (Variable sex: 1 Female, 2 Male, Perc Sex gibt den Anteil an). Folgende Variablen sind von Interesse:

- R00-R06: Mittlerer Score zur Lesekompetenz im Jahr 2000 bzw. 2006
- M00-M06: Mittlerer Score zur Kompetenz in der Mathematik im Jahr 2000 bzw. 2006
- S00-S06:Mittlerer Score in den Naturwissenschaften (science) im Jahr 2000 bzw. 2006

a) Laden Sie den Datensatz PISA.csv von der Homepage herunter und lesen sie ihn ein.

b) Untersuchen Sie deskriptiv (Boxplots aller 6 Variablen), ob sich die drei PISA-Scores des Jahres 2006 im Vergleich zum Jahr 2000 verändert haben. (Gehen Sie hierbei und im weiteren nicht näher auf irgendwelche Geschlechtsunterschiede ein).

c) Untersuchen Sie mit einem geeigneten Test, ob sich die drei PISA-Scores signifikant verändert haben.

Aufgabe 6: Produktionskontrolle bei Hustensaft

Laut Produktbeschreibung enthält der Hustensaft Mikasolvan 40g/Liter des Wirkstoffes Halsruhe. Die Produktion des Saftes wird angehalten, wenn die Konzentration des Wirkstoffes deutlich zu niedrig ist.

Um die Produktion zu überprüfen, wurden 9 Flaschen zufällig entnommen und die Konzentration von Halsruhe gemessen. Die Messwert finden Sie in der Datei Hustensaft.csv. Begründen diese Daten einen Produktionsstopp?

Aufgabe 7: Lineare Regression

Wir betrachten den Datensatz Suess.csv. Er beschreibt die Auswirkung von Feuchtigkeit und Süße auf den Geschmack einer Süßigkeit.

- *Geschmack*: Geschmackspunktzahl (Integer)
- *Feuchtigkeit*: Feuchtigkeitspunktzahl (Integer)
- *Suesse*: Süßegrad (Integer)

a) Laden Sie den Datensatz Sues.csv und speichern Sie ihn in einem Dataframe sues ab.

b) Benutzen Sie die Funktion coplot() um einen Plot von Geschmack abhängig von der Feuchtigkeit der Süßigkeit, bedingt auf den Süßegrad zu erstellen. Benutzen Sie für eine bessere Darstellung die Optionen pch = c(5,18), rows = 1 und columns = 3. Gibt es bei gegebener Feuchtigkeit einen Einfluss der Süße auf den Geschmack?

c) Fitten Sie das Modell
$Geschmack_i = \beta_0 + \beta_1 \cdot Feuchtigkeit_i + \beta_2 \cdot Suesse_i + \varepsilon_i.$

d) Plotten Sie die Residuen gegen *Feuchtigkeit* · *Suesse*.

e) Bestimmen Sie nun die Parameter des Modells mit Interaktionsterm

$$Geschmack_i = \beta_0 + \beta_1 \cdot Feuchtigkeit_i + \beta_2 \cdot Suesse_i + \beta_3 \cdot (Feuchtigkeit_i \cdot Suesse_i) + \varepsilon_i.$$

f) Verbessert sich die Anpassung des Modells an die Daten? Ist der Koeffizient β_3 der In- teraktion signifikant von 0 verschieden? Plotten Sie erneut die Residuen gegen Feuchtigkeit· Suesse und vergleichen Sie die Ergebnisse mit dem Plot aus Aufgabe (d).

Aufgabe 8: Korrelationen

In dieser Aufgabe müssen Sie eine eigene Untersuchung durchführen überlegen. Verwenden Sie dazu die Datenbank der EU https://ec.europa.eu/eurostat/ und laden sich dort mindestens zwei Datensätze A und B herunter (z.B Geburtenrate und Anzahl der Verkauften PKWs), welche eine Größe in Abhängigkeit der Jahreszahl angeben. Stellen Sie sicher, dass zumindest einer der Datensätze nicht nur eine Aufschlüsselung nach Jahreszahl sondern auch über die verschiedenen Länder der EU beinhaltet.

a) Visualisieren Sie beide Größen A und B in Abhängigkeit der Jahreszahl
b) Visualisieren Sie die Abhängigkeit ihrer Größen A und B von einander und berechnen Sie die Korrelation und die Signifikanz ihrer Korrelation.
c) Bestimmen Sie die Länder welche die größte und die kleinste Korrelation mit Deutschland der Größe A bzw. B aufweisen. Visualisieren Sie dies.

Tag 3: Multivariate Verfahren

Dieses Übungsblatt ist dem Datensatz SPECTF aus der Veröffentlichung Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. von Kurgan et al. im Jahr 2001 gewidmet.

- In der ersten Spalte befindet sich die binäre Zielvariable in zwei Klassen: normal und abnormal
- Die anderen p = 44 Spalten sind Parameter die Kardiale Single Proton Emission Computed Tomography (SPECT) Bilder mit n = 267 Beobachtungen beschreiben
- Alle 44 Parameter sind stetig und liegen in dem Intervall zwischen 0 und 100
- Die Daten sind schon präprozessiert und zu finden als RData-File auf der Kurs-Seite.

Aufgabe 9: Deskriptives

a) Lesen Sie den SEPCTF Datensatz mit Hilfe des load()-Befehls ein und überprüfen Sie welche Information das RData-File enthält und die Dimension der Daten SPECTF.

b) Berechnen Sie für jeden Parameter die Spannweite (Maximum-Minimum). Betrachten Sie die Summary und erstellen Sie ein Histogramm der Spannweiten.

c) Berechnen Sie für jeden Parameter die Varianz. Greifen Sie die 6 Parameter mit der höchsten Varianz heraus, erstellen Sie für diese je einen Boxplot und fassen Sie die Einzelboxplots in einer Graphik zusammen.

Aufgabe 10: Logistische Regression, Kreuzvalidierung und ROC-Kurven

In dieser Aufgabe sollen Sie anhand der SPECTF-Daten ein Modell erstellen um die binäre Zielvariable vorherzusagen. Dieses Modell soll mit Hilfe von ROC-Kurven evaluiert werden.

a) Informieren Sie sich selbstständig über Kreuzvalidierungsverfahren.

b) Implementieren Sie eine 10-fache Kreuzvalidierung mittels einer for-Schleife für den SPECTF Datensatz. Innerhalb dieser Schleife soll eine logistische Regression anhand des der Trainingsdaten (Zielvariable $\sim$ Parameter), sowie eine Vorhersage anhand der Testdaten gemacht werden.

c) Berechnen Sie mit Hilfe der R-package pROC und ROCR die ROC-Kurve, die Kon- fidnezintervalle und den AUC der Vorhersage aus Aufgabe (b).

d) Plotten Sie die ROC-Kurve und machen Sie im gleichen Plot den AUC-Wert und das Konfidenzinterval sichtbar. Speichern Sie den Plot in einem PDF-File ROC SPECTF.pdf.