Stochastik-Praktikum **WS 2023-2024**

Prof. Dr. Matthias Birkner
Jan Lukas Igelbrink

Blatt 10

Aufgabe 1 (Simpson-Paradoxon) (3 + 3 = 6 Punkte)
Wir wollen ein Beispiel für das Simpson-Paradoxon betrachten. Laden Sie den Datensatz *gehalt.txt* von der Praktikumsseite herunter und lesen Sie ihn mit `read.table` ein. Der Datensatz enthält die Gehälter, in Tsd., von Personen mit den Beruf *Hundedompteur* und *Katzenmasseur* sowie deren Berufserfahrung in Jahren.

a) Schlägt sich eine größere Berufserfahrung in einem höheren Gehalt nieder? Bestimmen Sie die Regressionsgerade und den mittleren quadratischen Abstand für Gehalt gegen Berufserfahrung. Tun Sie dies einmal nur für die *Hundedompteure*, einmal für die *Katzenmasseure* und einmal für beide Berufe zusammen.

b) Erstellen Sie einen Punktplot der Daten, die x-Achse sei die Berufserfahrung und die y-Achse das Gehalt. Markieren Sie die Datenpunkte der *Katzenmasseure* grün und die der *Hundedompteure* blau. Ergänzen Sie den Plot um die drei in b) berechneten Regressionsgeraden. Was fällt auf?

Aufgabe 2 (1+1+1 = 3 Punkte)
Lesen Sie die Datei *800m.txt* ein. Dieser Datensatz enthält die Bestzeiten im 800m-Lauf der Herren für die olympischen Spiele 1896-2020. (Letztere fanden 2021 statt.)

a) Finden und plotten Sie die Regressionsgerade für Bestzeit gegen Jahr.

b) Interpretieren Sie alle Outputs von `lm`.

c) Schreiben Sie eine Funktion, welche die n letzten Datenpunkte entfernt und dann basierend auf dem verbliebenen Datensatz mit Hilfe einer Regressionsgeraden die n fehlenden Bestzeiten schätzt. Führen Sie diese Funktion aus für $n = 1, ..., 15$ und plotten Sie das Ergebnis in Abhängigkeit von n im Vergleich zu den echten Daten.

Aufgabe 3 (Moore'sches Gesetz) (1 + 3 = 4 Punkte)
In dieser Aufgabe wollen wir uns mit dem sogenannten Mooreschen Gesetz beschäftigen. Der spätere Intel-Mitbegründer Gordon Moore behauptete 1965, dass sich die Zahl der Transistioren pro Chip jedes Jahr verdoppeln würde, später im Jahr 1975 war Moore nicht mehr so optimistisch und sprach nun von 2 Jahren. Dies erschien seinem Kollegen David House wiederum zu pessimistisch, weshalb er den Verdopplungszeitraum auf 18 Monate bezifferte.

a) Laden Sie vom Reader den Datensatz *prozessoren.txt* herunter. Dieser Datensatz enthält das Erscheinungsjahr, die Transistoren, die Taktung und die Energiedichte einer Auswahl von Prozessoren, welche von Intel über den Zeitraum von 1971 bis 2014 veröffentlicht wurden. Erstellen Sie einen Plot mit dem Erscheinungsjahr als x-Achse und den Logarithmus der Transistorenzahl als y-Achse.

b) Berechnen Sie die Regressionsgerade. Entscheiden Sie, welche der drei oben genannten Schätzungen am besten die Entwicklung der Transistorenzahl beschreibt und berechnen Sie die Verdopplungsrate (in Jahren) mithilfe der Steigung Ihrer Regressionsgeraden.

Falls Sie interessiert sind, finden Sie unter `https://www.computerhistory.org/siliconengine/moores-law-predicts-the-future-of-integrated-circuits/` einige historische Quellen.

Aufgabe 4 (2 + 3 = 5 Punkte)
Die Datei `Lottozahlen.txt` enthält für jede der Zahlen 1 bis 49 die absolute Häufigkeit, mit der diese Zahl in 4767 Ziehungen vorkam (ohne Zusatzzahl, siehe z.B. `https://www.dielottozahlende.net/lotto-6-aus-49/statistiken/haeufigkeit-der-lottozahlen`).

a) Bestimmen Sie für jede der Zahlen ein approximatives Konfidenzintervall zum Irrtumsniveau $\alpha = 0.05$ für die Wahrscheinlichkeit, diese Zahl in einer Ziehung zu sehen. Wie viele dieser Konfidenzintervalle überdecken den theoretischen Wert?

b) Ist das Ergebnis überraschend? Führen Sie dazu eine Simulationstudie mit R durch. Simulieren Sie 10^3 mal die 4767 Ziehungen und führen Sie **a)** für jede diese Simulation durch. Zählen Sie jedes Mal, wieviele Konfidenzintervalle den wahren Wert überdecken, und erstellen Sie hiervon ein Histogramm.