Prof. D. Metzler SS 2026

ÜBUNGEN ZUR WAHRSCHEINLICHKEITSRECHNUNG UND STATISTIK FÜR BIOLOGEN

Blatt 10

Aufgabe 1 Die Datei `Lottozahlen.txt` enthält für jede der Zahlen 1 bis 49 die absolute Häufigkeit, mit der diese Zahl in 4644 Ziehungen von 1955 bis 2008 des Lottos 6 aus 49 vorkam (ohne Zusatzahl, siehe z.B. `http://www.dielottozahlen.de`). Bestimmen Sie für jede der Zahlen ein Konfidenzintervall zum Irrtumsniveau $\alpha = 0{,}05$ für die Wahrscheinlichkeit, diese Zahl in einer Ziehung zu sehen. Wieviele dieser Konfidenzintervalle überdecken den theoretischen Wert? Ist das Ergebnis überraschend? – führen Sie beispielsweise mit `R` eine kleine Simulationsstudie durch.

Aufgabe 2 Mukoviszidose ist eine menschliche Erbkrankheit, die von einem Gendefekt auf Chromosom 7 hervorgerufen wird. Das Allel für Mukoviszidose ist rezessiv, d.h. nur homozygote Individuen erkranken tatsächlich. Nehmen wir an, eine Population befindet sich (bezüglich dieses Gens) im Hardy-Weinberg-Gleichgewicht, und unter je 3000 Geburten findet sich ein an Mukoviszidose erkranktes Kind.
a) Welcher Anteil der Population trägt genau eine Kopie des defekten Allels?
b) Nehmen wir an, ein gesundes Paar habe ein an Mukoviszidose erkranktes Kind. Wie wahrscheinlich ist es dann, dass ein weiteres Kind ebenfalls krank sein wird?
c) Nehmen wir an, ein gesundes Paar habe bereits ein gesundes Kind. Wie wahrscheinlich ist es dann, dass das zweite Kind krank sein wird?

Aufgabe 3 Die Datei `800m.csv` enthält die Bestzeiten im 800m-Lauf der Herren für die Jahre 1970 bis 2015 in Sekunden[1]. Berechnen und zeichnen Sie die (kleinste-Quadrate-)Regressionsgerade für Jahr gegen Bestzeit. Wie hätten Sie Ende 2015 die Bestzeit für 2016 prognostiziert? Wie sicher wären Sie sich damals bei Ihrer Schätzung gewesen?

Aufgabe 4 Erzeugen Sie mehrere zufällige Datensätze mit der Normalverteilung und anderen Verteilungen und vergleichen Sie die simulierten Daten mittels Normal-QQ-Plot mit der Normalverteilung (z.B. in R mit `qqnorm(rnorm(15))`). Betrachten Sie dann die Normal-QQ-Plots in der Datei `R2QQPlotsRaten.pdf` und raten Sie, welche der neun Datensätze unabhängige Stichproben aus einer Normalverteilung darstellen.

Aufgabe 5 Die Datei `2QQPlotsDichtepolygone.pdf` enthält Normal-QQ-Plots für Daten aus 6 verschiedenen Verteilungen. Skizzieren Sie Dichtepolygone, die zeigen, wie die jeweilige Verteilung von einer Normalverteilung (mit entsprechendem Erwartungswert und entsprechender Varianz) abweicht.

Aufgabe 6 Folgende Tabelle zeigt Dauer des Studiums (in Semestern) und Einstiegsgehalt (in Tausend €) der Absolventen eines Jahres am Fachbereich Mathematik und Informatik der Yule-Simpson-Universität:

Semester	12	14	16	12	15	14	13	14	11	13	10	12	14	13	14	15
Gehalt	39.4	38.2	37.4	39.5	32.8	35.3	39.1	35.2	37.9	35.7	41	40.9	34.2	38.4	36.2	38.4
Semester	9	11	9	9	12	13	11	10	10	10	9	10	12	10		
Gehalt	33.7	35.9	36.1	34.2	29.9	31.9	33.3	36.2	33.8	32.9	33.3	35.1	34.2	35.3		

[1] Wikipedia, `https://en.wikipedia.org/wiki/800_metres`, 14.06.2016

(a) Schlägt sich (für diese Absolventen) ein längeres Studium in einem höheren Anfangsgehalt nieder? Bestimmen Sie die Regressiongerade für Studiendauer gegen Anfangsgehalt.

(b) Ändert sich Ihr Befund, wenn Sie zusätzlich erfahren, dass die oberen beiden Zeilen der Tabelle sich auf die Absolventen des Fachs Informatik, die unteren beiden sich auf die Absolventen des Fachs Mathematik beziehen, und Sie dieselbe Regression jeweils innerhalb dieser beiden Gruppen durchführen?

(c) Führen Sie auch eine Regressionsanalyse durch, in der Sie sowohl dis Studiendauer als auch das Studienfach als erklärende Variablen für das Einstiegsgehalt berücksichtigen. Überprüfen Sie durch eine Varianzanalyse und weitere Kriterien, ob Sie zusätzlich einen Interaktionsterm zwischen Studienfach und Studiendauer im Modell haben sollten.

(d) Visualisieren Sie in geeigneter Weise die Daten und die Regressionsmodelle ohne und mit Berücksichtigung des Studienfachs.