(10 Punkte)

Aufgabe 1

Eine Vertriebsgesellschaft vermutet, dass es hinsichtlich der Nachfrage nach bestimmten Wochenzeitschriften regionale Unterschiede gibt. Um diese Vermutung statistisch zu überprüfen, zieht sie aus der Menge der Käufer von Wochenzeitschriften zwei voneinander unabhängige einfache Zufallsstichproben, und zwar eine aus der norddeutschen und eine aus der süddeutschen Region.
Die Käufer der beiden Stichproben werden gefragt, welche Wochenzeitschrift sie bevorzugen (keine Mehrfachnennungen!). Das Ergebnis dieser Befragung zeigt folgende Tabelle:

Zeitschrift	Nord (N)	Süd (S)	Summe
A	195	145	340
B	140	200	340
C	175	105	280
Summe	510	450	960

Mit einem χ^2-Test soll die Nullhypothese "Bevorzugte Zeitschrift ist unabhängig von der Region des Käufers" überprüft werden.

(a) Berechnen Sie die unter der Nullhypothese erwarteten Häufigkeiten für die sechs Felder der Tabelle:

Zeitschrift	Nord (N)	Süd (S)	Summe
A	E_1	E_2	340
B	E_3	E_4	340
C	E_5	E_6	280
Summe	510	450	960

(b) Als Teststatistik wird das Pearson'sche χ^2-Maß betrachtet. Die Teststatistik folgt unter der Nullhypothese asymptotisch einer χ^2-Verteilung mit k Freiheitsgraden. Wir groß ist k im betrachteten Beispiel (bitte begründen!)

(c) Als Teststatistik ergibt sich der Wert 31.8. Kann die Nullhypothese auf dem Niveau $\alpha = 0.01$ verworfen werden?

Hinweis: Die tabellierte Quantilsfunktion der χ^2_k-Verteilung für ausgewählte Anzahlen von Freiheitsgraden k finden Sie im Anhang.

(d) Benennen Sie ein alternatives Maß, das zur Untersuchung der betrachteten Nullhypothese verwendet werden kann.
Welcher (asymptotischen) Verteilung folgt das alternative Maß?

Aufgabe 2 **(5 Punkte)**

Betrachten Sie ein genetisches Merkmal mit der folgenden relativen Häufigkeitsverteilung der Genotype in einer Population: Der Genotyp ‚bb', der durch einen Test erkannt werden

aa	ab	bb
0.5	0.45	0.05

soll, sei der Verursacher einer Krankheit. Ein positives Testergebnis bedeutet hier also, dass der Test anzeigt 'Genotyp ist bb'. Der Test funktioniert jedoch nur fehlerhaft, wobei bei gegebenen Genotypen die folgenden Wahrscheinlichkeiten für ein positives Testergebnis bekannt seien:

Genotyp	aa	ab	bb
Wahrscheinlichkeit für positives Testergebnis	0.05	0.10	0.95

a) Bestimmen Sie die Wahrscheinlichkeit dafür, bei einer zufällig aus der Population ausgewählten Person ein positives Testergebnis zu erhalten.

b) Bestimmen Sie die Wahrscheinlichkeit, dass eine Person tatsächlich den Genotyp ‚bb' hat, wenn ein positives Testergebnis vorliegt. Wenn Sie bei der vorherigen Aufgabe kein Ergebnis erreicht haben, benutzen Sie für diese Rechnung für die Wahrscheinlichkeit bei einer zufällig ausgewählten Person ein positives Testergebnis zu erhalten den Wert 0.12.

c) Im Folgenden werden nur noch die Personen mit Genotyp ‚bb' betrachtet.
Nehmen Sie an, dass eine Person mit Genotyp ‚bb' mit Wahrscheinlichkeit 0.3 tatsächlich erkrankt und dass das Testergebnis für eine Person mit Genotyp ‚bb' unabhängig davon ist, ob die Person erkrankt ist oder nicht. Wie groß ist nun die Wahrscheinlichkeit, dass eine Person mit Genotyp ‚bb' erkrankt ist, wenn der Test ein positives Ergebnis ergibt.

Aufgabe 3 **(11 Punkte)**

Gegeben sei eine Stichprobe $X = (X_1, \ldots, X_n)$ von unabhängig identisch verteilten Zufallsvariablen, wobei für die Dichte von X_i, $i = 1, \ldots n$, gilt:

$$f(x_i; \theta) = (\theta - 1)\, x_i^{-\theta}\, I_{[1,\infty)}(x_i), \quad \theta > 1.$$

a) Bestimmen Sie den Maximum-Likelihood-Schätzer für θ.

b) Bestimmen Sie den Standardfehler des Maximum-Likelihood-Schätzers.

c) Bestimmen Sie ein 95%-Wald-Konfidenzintervall für die Stichprobe
$x = (1.8, 2, 2.1, 2.1, 2.3, 2.5, 2.5, 2.7, 2.9, 3.1)$.
Hinweis: $\sum_{i=1}^{n} \log(x_i) = 8.62$

d) Abbildung 1 zeigt die normierte Log-Likelihoodfunktion für die gegebenen Daten. Begründen Sie kurz (Stichpunkte) warum und inwiefern ein Likelihood-Intervall für $\hat{\theta}_{ML}$ in dieser konkreten Datensituation bessere Ergebnisse als ein Wald-Intervall liefern würde.

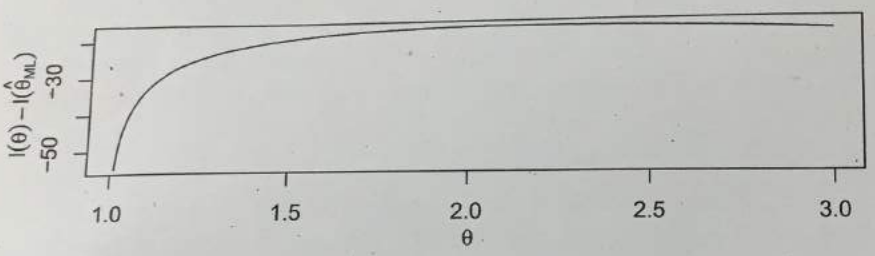


Abbildung 1: Normierte Log-likelihood Funktion.

Aufgabe 4 **(7 Punkte)**

Gegeben sei eine Zufallsvariable X mit stetiger Dichte

$$f(x) = \begin{cases} \lambda + c \cdot (1 - \lambda) \cdot x & \text{für } x \in [0, 1], \\ 0 & \text{sonst.} \end{cases}$$

und $\lambda \in \mathbb{R}$.

(a) Zeigen Sie, dass $c = 2$ gilt.

(b) Zeigen Sie, dass der Erwartungswert $E(X) = \frac{2}{3} - \frac{\lambda}{6}$.

(c) Überprüfen Sie, ob der Schätzer $\hat{\lambda} = 4 - \frac{6}{n} \sum_{i=1}^{n} X_i$ erwartungstreu für λ ist.

(d) Bestimmen Sie $P(X = 0.3)$.

Aufgabe 5

(7 Punkte)

In einer Serie von Spielen mit zwei Spielern hat am Anfang Spieler 1 ein Guthaben von 1 Euro und Spieler 2 ein Guthaben von 2 Euro. In jeder Runde gewinnt einer der beiden Spieler mit Wahrscheinlichkeit 1/3 (und erhält dann 1 Euro vom Gegner), oder das Spiel endet unentschieden mit Wahrscheinlichkeit 1/3. Es wird solange gespielt, bis mindestens ein Spieler ruiniert ist. Dies ist der Fall, wenn einer der beiden Spieler einen Euro an den Gegner zahlen müsste, aber nur noch ein Guthaben von 0 Euro hat. In diesem Fall bleibt das Guthaben bei 0 Euro und das Spiel ist beendet. Sei X_i das Kapital von Spieler 1 nach der i-ten Runde, wobei X_0 das Startkapital bezeichnet. Die Kapitalentwicklung von Spieler 1 wird als zeithomogener Markovprozess $\boldsymbol{X} = (X_0, X_1, X_2, \ldots)$ auf dem Zustandsraum $S = \{0, 1, 2, 3\}$ modelliert. Das heißt, $X_2 = 1$ bedeutet, dass nach zwei Durchgängen des Spiels Spieler 1 über ein Kapital von 1 verfügt (und Spieler 2 demnach über ein Kapital von 2).

(a) Bestimmen Sie die Übergangsmatrix $\boldsymbol{P} \in \mathbb{R}^{4\times 4}$.

(b) Bestimmen Sie die Verteilung $\boldsymbol{\mu}^{(n)}$ von X_n für $n = 0, 1, 2$.

(c) Bei welcher der folgenden Verteilungen handelt es sich um die stationäre Verteilung der Markovkette?

$$\boldsymbol{\pi}_1 = \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right) \quad \text{oder} \quad \boldsymbol{\pi}_2 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

Aufgabe 6

(13 Punkte)

Zahlreiche deutsche Städte erstellen sogenannte Mietspiegel, um Mietern und Vermietern eine "objektive" Entscheidungshilfe in Mietfragen zur Verfügung zu stellen. Bei der Erstellung von Mietspiegeln wird aus der Gesamtheit aller in Frage kommenden Wohnungen eine repräsentative Zufallsstichprobe gezogen und die Daten werden von Interviewern anhand von Fragebögen ermittelt. Auf der Basis einer solchen Erhebung mit insgesamt 2053 Wohnungen wird ein lineares Regressionsmodell für die Nettomiete pro Quadratmeter mit Einflussgrößen "Wohnfläche" (in m^2, `wfl`), "Anzahl Zimmer" (`rooms`), "Beste Lage" (`best`, mit Ausprägungen 1 falls Wohnung in bester Lage und 0 sonst), "Baujahr" (`bj`) geschätzt:

$$Y_i = \beta_0 + \beta_1 x_{\texttt{wfl}} + \beta_2 x_{\texttt{rooms}} + \beta_3 x_{\texttt{best}} + \beta_3 x_{\texttt{bj}} + \epsilon_i$$

Unten ist die Tabelle der Koeffizientenschätzer gegeben.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-38.31	4.09	-9.38	0.00
wfl	0.01	0.00	1.44	0.15
wohnbest	16.80	3.50		0.00
bj	0.02	0.00	11.87	0.00
rooms	-0.71	0.09	-7.53	0.00

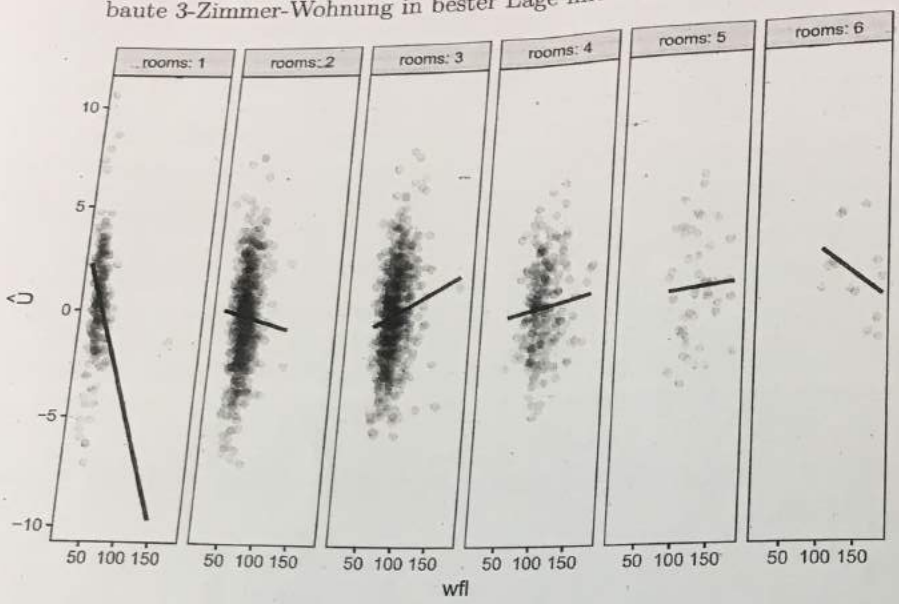
Tabelle 1: Koeffizientenschätzer des Linearen Modells.

a) Betrachten Sie die obige Tabelle und rechnen Sie den fehlenden t-Wert für `wohnbest` aus. Beschreiben Sie das Vorgehen zur Berechnung des p-Werts für den Regressionskoeffizienten der Variable `wfl` (keine Berechnung gefragt).

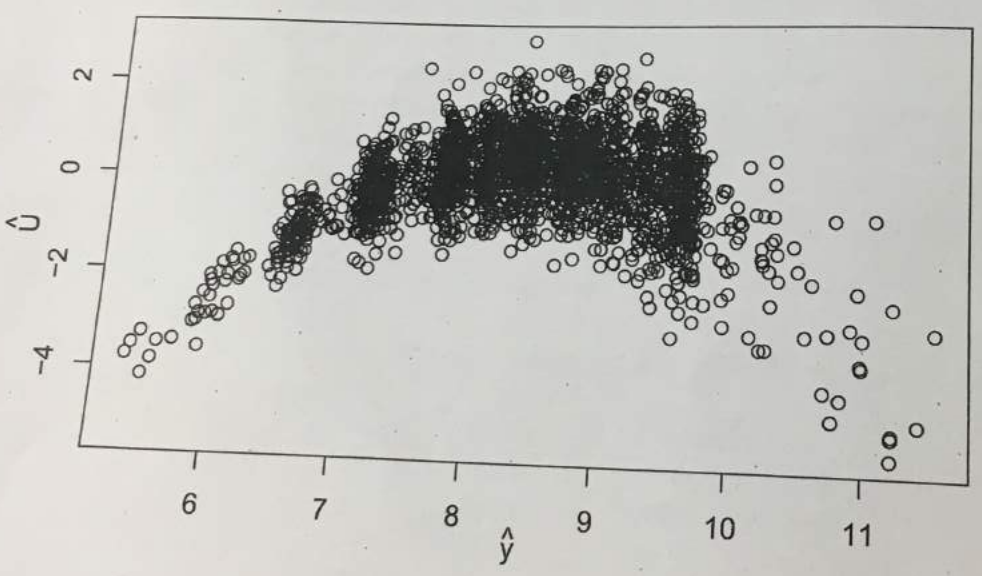
b) Geben Sie eine Interpretation des Intercepts an. Ist es hier inhaltlich sinnvoll diesen zu interpretieren? Begründen Sie Ihre Antwort.

c) Interpretieren Sie die Koeffizientenschätzer der Variablen x_{wfl} und x_{rooms}.

d) Berechnen Sie die nach dem obigen Modell prognostizierte Miete für eine 1990 erbaute 3-Zimmer-Wohnung in bester Lage mit einer Wohnfläche von $75m^2$.



e) Betrachten Sie die obige Grafik. Erläutern Sie warum das Modell möglicherweise um einen Interaktionseffekt zwischen `wfl` und `rooms` erweitert werden sollte. Nehmen Sie in Ihrer Antwort Bezug auf das beobachtete Muster in den Residuenplots.



f) Betrachten Sie die obige Grafik. Welche Annahmen des linearen Regressionsmodells werden hier verletzt?