

12. Übungsblatt „Einführung in die Statistik“

Abgabe bis Mittwoch 6.7., 16 Uhr.

1. (Chiquadrat-Test) Bei einem χ^2 -Test auf Unabhängigkeit in einer 2×3 Kontingenztafel sind Daten verloren gegangen. Es ist nur noch zu erkennen, dass die erwartete Häufigkeit in einer bestimmten Klasse 10 ist und die beobachtete Häufigkeit 20 beträgt. Lässt sich mit dieser Information etwas anfangen wenn das Niveau des Tests bei 1.5% bzw. 5% bzw. 10% liegt?

2. (Simpson Paradox) Die folgende Tabelle zeigt Dauer des Studiums (in Semestern) und Einstiegsgehalt (in 1000 Euro) der Absolventen eines Jahres am Fachbereich Mathematik und Informatik der (hypothetischen) Yule-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		

- Schlägt sich (für diese Absolventen) ein längeres Studium in einem höheren Anfangsgehalt nieder? Bestimmen Sie die Regressionsgerade für Studiendauer gegen Anfangsgehalt.
- 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 diesselbe Regression jeweils innerhalb dieser beiden Gruppen durchführen?

3. (Autoregressives Modell) Zur Beschreibung zeitlicher Entwicklungen mit deterministischer Wachstumstendenz und zufälligen Störungen verwendet man oft das folgende autoregressive Modell:

$$Y_k = \gamma Y_{k-1} + \sqrt{v} \xi_k, \quad 1 \leq k \leq n.$$

Dabei sind $\gamma \in \mathbb{R}$ und $v > 0$ unbekannte Parameter, $Y_0, \dots, Y_n$ die Beobachtungen, und $\xi_1, \dots, \xi_n$ unabhängige zufällige Störungen mit $\mathbb{E}(\xi_k) = 0$ und $\text{Var}(\xi_k) = 1$.

- Zeigen Sie: Im Fall von standardnormalverteilten ξ_k 's und Startvariable $Y_0 = 0$ lautet die Likelihood-Funktion für $\theta = (\gamma, v)$

$$L(\gamma, v; y) = (2\pi v)^{-n/2} \exp \left[-\frac{1}{2v} \sum_{k=1}^n (y_k - \gamma y_{k-1})^2 \right].$$

- b) Machen Sie einen Ansatz für den quadratischen Fehler, und bestimmen Sie den Kleinste-Quadrate-Schätzer für γ . Ist dieser Schätzer erwartungstreu?

4. (Lineares Modell) Für zwei Objekte mit unbekannten Gewichten w_1 und w_2 werden die Gewichte der Einzelobjekte, die Summe und Differenz der Gewichte gemessen. Jede der vier Messungen sei mit einem unabhängigen Fehler derselben Varianz behaftet. Es wurden die Messwerte 3.4, 2.9, 6.3, 0.7 beobachtet.

- a) Stellen Sie ein lineares Modell auf.
- b) Bestimmen Sie die Kleinste-Quadrate-Schätzer für w_1 , w_2 und $w_1 - w_2$, sowie deren geschätzte Streuung.
- c) Testen Sie die Hypothese $H_0 : w_1 = w_2$ gegen $H_1 : w_1 \neq w_2$ (unter der Annahme normalverteilter Messfehler).

5. (Lineare Regression mit R) Auf der Vorlesungshomepage finden Sie eine Datei mit dem in der Vorlesung betrachteten Datensatz *carmileage*.

- a) Bestimmen und plotten Sie den Benzinverbrauch und die entsprechenden Regressionsgeraden in Abhängigkeit von den übrigen Kenngrößen. Beachten Sie, dass der Benzinverbrauch (in Gallon pro Meile) das Inverse der Kenngröße MPG (miles per gallon) ist.
- b) Bestimmen und plotten Sie auch die logarithmierten Werte, und die entsprechenden Regressionsgeraden. Vergleichen Sie die Ergebnisse aus a) und b).
- c) Schätzen Sie die Koeffizienten im multiplen Regressionsmodell für den Benzinverbrauch in Abhängigkeit von den anderen Kenngrößen, und interpretieren Sie die Ergebnisse.
- d) Führen Sie die Befehle `summary(...)` und `plot(..)` für die oben betrachteten Regressionsmodelle aus. Welche Statistiken und Grafiken werden dargestellt?

Literatur: Petzoldt, Skript zur Datenanalyse mit R, Abschnitt 7.3.