Institut für Angewandte Mathematik
Sommersemester 2024
Andreas Eberle, Stefan Oberdörster

12. Übungsblatt „Einführung in die Statistik"

Abgabe bis Dienstag 9.7.

---

1. (Chiquadrat-Test auf Unabhängigkeit) 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. (Bedingte Austauschbarkeit, Kontingenztafeln und Chiquadrat-Statistik) Betrachten Sie die Situation aus der Vorlesung, d.h. $(X_1, Y_1), \ldots, (X_N, Y_N)$ sind unabhängig und identisch verteilt mit Werten in $\{a_1, \ldots, a_K\} \times \{b_1, \ldots, b_L\}$. Sei Π eine von $X = (X_1, \ldots, X_N)$ und $Y = (Y_1, \ldots Y_N)$ unabhängige Zufallspermutation aus $\mathcal{S}_N$, und $T(X, Y)$ die Chiquadrat-Statistik. Zeigen Sie:

a) Sind X und Y unabhängig, dann hat $(X, \Pi Y)$ dieselbe Verteilung wie (X, Y).

b) Die folgenden Aussagen gelten für alle Beobachtungswerte $x = (x_1, \ldots x_N)$ und $y = (y_1, \ldots y_N)$, und für alle $k \in \{1, \ldots, K\}$ und $l \in \{1, \ldots, L\}$:

$$\begin{aligned} \mathbb{E}\left[H_{kl}(x, \Pi y)\right] &= \bar{H}_{kl}(x, y), && (1) \\ \mathbb{E}\left[H_{kl}(x, \Pi y)^2\right] &= \bar{H}_{kl}(x, y)\left(1 + (H_k(x) - 1)(H_l(y) - 1)/(N - 1)\right), && (2) \\ \mathbb{E}\left[T(x, \Pi y)\right] &= (K - 1)(L - 1)N/(N - 1). && (3) \end{aligned}$$

c) Folgern Sie: Ist Y bedingt austauschbar gegeben X im Sinne von a), dann gilt

$$\begin{aligned} \mathbb{E}\left[H_{kl}(X, Y)\right] &= \mathbb{E}\left[\bar{H}_{kl}(X, Y)\right] \quad \text{für alle } k, l, \text{ und} && (4) \\ \mathbb{E}\left[T(X, Y)\right] &= (K - 1)(L - 1)N/(N - 1). && (5) \end{aligned}$$

3. (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		

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

b) Ändert sich Ihr Befund, wenn Sie zusätzlich erfahren, dass sich die oberen beiden Zeilen der Tabelle 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?

4. (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, $Y_0, \ldots, Y_n$ die Beobachtungen, und $\xi_1, \ldots, \xi_n$ unabhängige zufällige Störungen mit $\mathbb{E}(\xi_k) = 0$ und $\mathrm{Var}(\xi_k) = 1$.

a) Zeigen Sie: Im Fall von standardnormalverteilten ξ_k 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?

5. (Datensätze mit R analysieren) Auf der webpage "ggobi.org/book" finden Sie unter "Data" diverse Datensätze im csv-Format, deren Zusammensetzung in der pdf-Datei "Data Descriptions" beschrieben ist. Einen solchen Datensatz können Sie in R mit dem Befehl x< −read.csv("http://.... .csv") einlesen und unter x abspeichern. Mit x$y können Sie dann auf den Teildatensatz zum Merkmal y zugreifen. Dies ist äquivalent zu x[[k]], wenn das Merkmal y in der k-ten Spalte steht.

a) Lesen Sie den Datensatz "Tips" ein, und lassen Sie sich diesen mit View(x) ausgeben. Erstellen Sie einen Scatterplot der Höhe des Trinkgelds (x$tip) in Abhängigkeit von der Höhe der Rechnung (x$totbill). Erstellen Sie auch Boxplots und Histogramme dieser Merkmale.

b) Es liegt nahe, dass die Höhe des Trinkgelds in vielen Fällen proportional zur Höhe der Rechnung ist. Daher betrachten wir das Verhältnis r< −xtip/xtotbill. Erstellen Sie auch hier einen Scatterplot und einen Boxplot.

c) Berechnen Sie Mittelwerte und getrimmte Mittelwerte für verschiedene Werte $\tau \in [0, 0.5]$. Wie kommt die Abhängigkeit der getrimmten Mittelwerte von τ zustande?

d) Erstellen Sie mithilfe von qqnorm(r) und qqline(r) einen Normal-Q-Q Plot. Was ist in einem solchen Plot dargestellt, und was können Sie daraus ablesen?

e) Berechnen Sie Konfidenzintervalle zum Niveau 95% für den Erwartungswert und den Median von r.

f) Untersuchen Sie die Abhängigkeit der Höhe der Gesamtrechnung von den anderen (kategoriellen) Merkmalen grafisch. Dies geht zum Beispiel mit dem Befehl boxplot(totbill∼sex+smoker+time,data=x).