

3.5 Chi-Quadrat Tests für diskrete Daten

Gegeben seien Zählraten wie $X_i =$ Anzahl Fälle in der Klasse i . Wie testen wir, ob diese Daten mit einem Modell konsistent sein?

3.5.1 Beispiele

(1) Ist die Anzahl der Tore in einem Fußballspiel eine poissonverteilte ZV mit $\lambda = 3$? Daten aus 18 Spielen aus der 1. und 2. Bundesliga stehen zur Verfügung:

7, 9, 2, 1, 7, 3, 5, 3, 6 und 2, 3, 5, 3, 5, 2, 3, 0, 5

(2) Gibt es eine geschlechtsspezifische Abneigung gegen das Studium an der Mathematisch-Naturwissenschaftlichen Fakultät der Uni Augsburg? Im Sommersemester 1999 waren die Zahlen

	Frauen	Männer
MNF	433	693
Rest	5505	4641

Methode

Wir berechnen die unter dem Modell zu erwartenden Werten und vergleichen sie mit den beobachteten Werten.

Beispiel (1) Anzahl Tore

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Sei e_i die erwartete Anzahl Spiele mit i Toren unter den 18 Spielen und o_i die beobachtete Zahl:

Tore	0	1	2	3	4	5	6	≥ 7
o_i	1	1	3	5	0	4	1	3
e_i	0.89	2.69	4.03	4.03	3.02	1.81	0.91	0.60

Heuristische Idee. Die (quadrierten) Abweichungen

$$(o_i - e_i)^2$$

messen den Unterschied zum Modell für eine Klasse, aber sie sind nicht gut untereinander vergleichbar. Wir wollen nicht die (o_i, e_i) Paare

$$(100, 110) \quad \text{und} \quad (10, 20)$$

gleich bewerten. Deshalb arbeiten wir mit den relativen Abweichungen

$$\frac{(o_i - e_i)^2}{e_i}$$

Wir summieren sie über alle Klassen, und verwenden die Teststatistik

$$T = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Für H_0 stellt sich heraus, dass unter gewissen Annahmen

$$T \sim \chi_7^2$$

wo 7 die Anzahl der Klassen minus 1 ("Anzahl Freiheitsgrade") ist. Wir werden H_0 verwerfen, wenn T groß ist.

$$P(T > 14.1) = 0.05 \quad P(T > 18.5) = 0.01$$

Für die Daten finden wir $T = 16.79$, so dass wir H_0 zu einem Testniveau von 5% verwerfen können, aber nicht zu einem Niveau von 1%.

Der Durchschnitt der Tore pro Spiel war $71/18 = 3.944$. Wenn wir einen Test mit diesem λ ausführen, finden wir $T = 7.03$, aber wir müssen die Benutzung des Schätzers berücksichtigen. Da ein Parameter geschätzt worden ist, verlieren wir einen Freiheitsgrad und nehmen $T \sim \chi_6^2$ an.

$$P(T > 12.6) = 0.05$$

H_0 wird in diesem Fall akzeptiert.

Beispiel (2) Fakultäten

H_0 : kein Geschlechtseinfluß — der Anteil in MNF entspricht dem Gesamtanteil der Frauen an Studierenden.

Insgesamt gibt es 5938 Frauen und 5334 Männer mit einem Frauenanteil von $\frac{5938}{11272}$. Unter H_0 erwarten wir

$$\frac{5938}{11272} * 1126$$

Frauen in MNF. Die erwarteten Werte für die Tabelle sind

	Frauen	Männer
MNF	593.17	532.83
Rest	5344.83	4801.17

Aus demselben heuristischen Argument wie vorher nehmen wir als Teststatistik

$$X^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Approximativ gilt

$$X^2 \sim \chi_{(r-1)(c-1)}^2$$

wo r die Anzahl von Zeilen ("rows") sei und c die Anzahl von Spalten ("columns"). $r = 1, c = 1 \Rightarrow X^2 \sim \chi_1^2$

$$X^2 = 101.5 \Rightarrow H_0 \text{ verwerfen.}$$

3.5.2 (Zweidimensionale) Kontingenztafeln im allgemeinen

Gegeben seien zwei Klassifikationen der Daten, $i = 1, \dots, r$ und $j = 1, \dots, c$. Diese Information kann in einer $r \times c$ Tabelle mit Einträgen

$$n_{ij} = \text{Anzahl F\u00e4lle in } i \text{ und } j$$

zusammengefasst werden. Wir interessieren uns meistens f\u00fcr die Hypothese

H_0 : keine Assoziation zwischen den Klassifikationen.

Sei

$$p_{ij} = P(\text{Zeilenklasse} = i, \text{Spaltenklasse} = j)$$

$$p_{i.} = P(\text{Zeilenklasse} = i)$$

$$p_{.j} = P(\text{Spaltenklasse} = j)$$

$$\text{dann } H_0 \Rightarrow p_{ij} = p_{i.} p_{.j}$$

Wir schätzen $p_{i.}, p_{.j}$ mit

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \hat{p}_{.j} = \frac{n_{.j}}{n}$$

$$\Rightarrow e_{ij} = n \left(\frac{n_{i.}}{n} \right) \left(\frac{n_{.j}}{n} \right)$$

$$= \frac{(\text{Zeilensumme } i)(\text{ Spaltensumme } j)}{n}$$

Als Teststatistik haben wir

$$X^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Für $n \rightarrow \infty$ gilt $X^2 \sim \chi^2_{(r-1)(c-1)}$.

Als Faustregel (Cochran 1952!) hat man $e_{ij} \geq 5 \quad \forall ij$,
aber das ist sicherlich zu streng.

Freiheitsgrade

es sind rc Zellen (Kombinationen) rc

n wird festgelegt -1

$p_{i.}$ werden geschätzt $-(r - 1)$

$p_{.j}$ werden geschätzt $-(c - 1)$

Es bleiben $(r - 1)(c - 1)$

$$Y = X^2 \sim \chi_k^2$$

$$f(y) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} y^{k/2-1} e^{-y/2} \quad x \geq 0$$

$$E[Y] = k$$

$$V[Y] = 2k$$

Für $k = 1, 2$ hat die Dichte ein Maximum bei $y = 0$.

Für $k \geq 3$ ist $f(0) = 0$ und das Maximum liegt bei $y = k - 2$.