Dr. Katharina Best
Sommersemester 2011
28. April 2011

ulm university universität **uulm**

Reader – Teil 3: Kategoriale Daten

WiMa-Praktikum

1 Beschreibung kategorialer Daten

Von kategorialen Daten spricht man, wenn die Ausprägungen Kategorien sind. Sie können ungeordnet sein (wie männlich/weiblich) oder auch geordnet, aber nicht quantifiziert (wie Schulnoten). Im Allgemeinen haben wir eine vorher festgelegte Anzahl von Ausprägungen und betrachten Häufigkeiten von Beobachtungen.

Ein verbreiteter Fall ist das Vorhandensein zweier Merkmale in jeweils zwei Ausprägungen, die gezählt werden. Dies führt zur Darstellung in *Vier-Felder-Tafeln*. Dieser Fall kann zu beliebig vielen Ausprägungen der zwei Merkmale verallgemeinert werden, sagen wir I und J, deren Darstellung wir dann *Kontingenztafeln* nennen.

2 Fragestellungen

Die Analyse solcher Tafeln ist natürlich von der Fragestellung abhängig. Der Hintergrund ist die Frage nach dem Zusammenhang der Merkmale, wobei die Frage unterschiedlich gestellt werden kann. Statistisch bietet sich hier ein Test auf stochastische Unabhängigkeit, den wir hier als erstes vorstellen.

Wir sollten uns immer vergegenwärtigen, dass nicht Kausalbeziehungen, sondern nur mögliche Korrelationen untersucht werden.

2.1 Unabhängigkeitstest

Wir gehen von einer verbundenen i. i. d. Stichprobe $X_1, \ldots, X_n$ der Länge n aus, wobei $X_i = (X_{i1}, X_{i2})$, deren Verteilung der von $X = (\widetilde{X}_1, \widetilde{X}_2)$ folgt, also $X_1 \sim X$. Die erste Koordinate hält das erste Merkmal, die zweite das zweite fest. Haben wir I Ausprägungen $A_1, \ldots, A_I$ des ersten Merkmals und J Ausprägungen $B_1, \ldots, B_J$ des zweiten Merkmals zur Beschreibung, welche wir mit ihrer Nummer in einer Anordnung identifizieren, so können wir eine Häufigkeitstabelle aufstellen.

Dazu definieren

$$N_{ij}(X_1,\dots,X_n) = \left|\left\{k : 1 \le k \le n : X_{k_1} \in A_i, X_{k_2} \in B_j\right\}\right| \qquad \text{für } 1 \le i \le I, 1 \le j \le J$$

die Anzahl der Stichprobenzufallsvariablen $X_1,\dots,X_n$, die in der i-ten Zeile und der j-ten Spalten der Tafel gezählt werden, und n_{ij} die Anzahl der (konkreten) Stichprobenwerte $x_1,\dots,x_n$ entsprechend.

Weiter definieren wir Spalten-

$$n_{\bullet j} := \sum_{i=1}^{I} n_{ij} \qquad \text{und Zeilensummen} \qquad n_{i\bullet} := \sum_{j=1}^{J} n_{ij}$$

und die Wahrscheinlichkeiten p_{ij}, in dem (i,j)-Feld der Tafel zu landen, führen analog zu den Marginalwahrscheinlichkeiten

$$p_{\bullet j} := \sum_{i=1}^{I} p_{ij} \qquad \text{und} \qquad p_{i\bullet} := \sum_{j=1}^{J} p_{ij},$$

Die Hypothesen sind dann

$$H_0 : p_{ij} = \mathbb{P}\left((\widetilde{X}_1,\widetilde{X}_2) = (A_i,B_j)\right) = \mathbb{P}(\widetilde{X}_1 = A_i)\cdot\mathbb{P}(\widetilde{X}_2 = A_j) \qquad \text{für alle } i,j,$$
$$H_1 : p_{ij} = \mathbb{P}\left((\widetilde{X}_1,\widetilde{X}_2) = (A_i,B_j)\right) \neq \mathbb{P}(\widetilde{X}_1 = A_i)\cdot\mathbb{P}(\widetilde{X}_2 = A_j) \qquad \text{für mindestens ein Paar } (i,j).$$

Es besteht die Möglichkeit, einen exakten oder einen approximativen Test auszuführen.

2.1.1 Exakter Test nach Fisher

Da der exakte Test mit wachsenden I, J kompliziert wird, wollen wir ihn nur für den Vier-Felder-Fall vorstellen. Er basiert auf der hypergeometrischen Verteilung.

Unter der Nullhypothese ist bei festen Zeilen- und Spaltensummen die Kombinationswahl hypergeometrisch verteilt. Für die Teststatistik T gilt dann

$$\mathbb{P}(T=k) = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{22}}}{\binom{n}{n_{\bullet 1}}} \tag{1}$$
$$= \frac{n_{1\bullet}!n_{2\bullet}!n_{\bullet 1}!n_{\bullet 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!} \tag{2}$$

Der p-Wert berechnet sich hier durch

$$p := \sum_{\substack{k\in\{0,1,\dots,n\}:\\ \mathbb{P}(T=k)\le\mathbb{P}(T=n_{11})}} \mathbb{P}(T=k) \tag{3}$$

2.1.2 χ^2-Unabhängigkeitstest

Für größere I, J kann approximativ vorgegangen werden. Hier bietet sich die χ^2-Approximation an.

Überlegen Sie sich:
(a) Wie groß sollten I, J werden, damit es funktioniert?
(b) Wieso greift hier die χ^2-Approximation?

Die Wahrscheinlichkeiten für die einzelnen Einträge der Kontingenztafel lassen sich mit den ML-Schätzern

$$\hat{p}_{ij} \overset{\text{unter } H_0}{=} \hat{p}_{i\bullet} \cdot \hat{p}_{\bullet j} \overset{ML}{=} \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} \tag{4}$$

schätzen. Unter der Nullhypothese gilt dann für die geschätzte Anzahl in einem Eintrag der Tafel

$$\hat{m}_{ij} = n\hat{p}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}. \tag{5}$$

In diesem Fall ist die Teststatistik $T : \mathcal{X} \to [0, \infty)$ mit

$$T(x_1, \ldots, x_n) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}},$$

Dann ist $\lim_{n\to\infty} T(X_1, \ldots, X_n) \sim \chi^2_{(I-1)(J-1)}$, wobei sich die Anzahl der Freiheitsgrade ergibt aus $(I \cdot J - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1)$.

Mit dem gleichen Testverfahren können wir auch eine andere Fragestellung untersuchen, die nach der Homogenität der Verteilungen eines Merkmals in verschiedenen Stichproben.

2.2 Homogenitätstest

Beim Homogenitätstest soll geprüft werden, ob J verschiedene Stichproben der selben Wahrscheinlichkeitsverteilung folgen. Das heißt, dass die Zugehörigkeit zu einer Population nicht als Realisierung einer Zufallsvariablen zu sehen ist. Die Spalten selbst sind voneinander unabhängig.

Eine typische Fragestellung ist z. B. ob das Wahlverhalten in Rheinland-Pfalz und Baden-Württemberg unterschiedlich ist (bzw. am 27. März 2011 war).

Die Hypothesen sind dann

$$\begin{aligned} H_0 &: p_i := p_{i1} = p_{i2} = \ldots = p_{iJ} && \text{für alle } 1 \leq i \leq I, \\ H_1 &: p_{ij} \neq p_i && \text{für mindestens ein } i \text{ und ein } j. \end{aligned}$$

Auch hier kann entweder ein exakter oder ein approximativer Test durchgeführt werden.

2.2.1 Exakter Test nach Fisher auf Homogenität

Überlegen Sie sich:
(a) Wie Sie einen exakten Test auf Homogenität ausführen würden.
(b) Wie dabei der p-Wert definiert wäre.

2.2.2 χ^2-Homogenitätstest

Die Wahrscheinlichkeiten $p_1, \ldots, p_I$ lassen sich mit den ML-Schätzern

$$\hat{p}_i = \frac{n_{i\bullet}}{n} \qquad \text{für } 1 \leq i \leq I \tag{6}$$

schätzen. Die Teststatistik

$$T(x_1, \ldots, x_n) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - n_{\bullet j}\hat{p}_i\right)^2}{n_{\bullet j}\hat{p}_i}, \tag{7}$$

ist dann approximativ χ^2-verteilt mit $(I-1)(J-1)$ Freiheitsgraden.

2.3 Anpassungstest

Der Anpassungstest ist von der Fragestellung dem Homogenitätstest ähnlich. Wir haben allerdings nur eine Stichprobe, welche gegen eine vorgegebene Verteilung $p_1, \ldots, p_I$ mit $0 < p_i < 1$ und $\sum_{i=1}^{I} p_i = 1$ getestet werden soll.

Wir haben demnach ein Merkmal mit den verschiedenen Ausprägungen $A_1, \ldots, A_I$. Die Hypothesen sind

$$\begin{aligned} H_0 &: \mathbb{P}(X = A_i) = p_i \qquad \text{für alle } 1 \leq i \leq I \\ H_1 &: \mathbb{P}(X = A_i) \neq p_i \qquad \text{für ein } 1 \leq i \leq I \end{aligned}$$

Wir untersuchen die Teststatistik $T : \mathcal{X} \to [0, \infty)$ mit

$$T(x_1, \ldots, x_n) = \sum_{j=1}^{I} \frac{\left(n_j - n \cdot p_j\right)^2}{n \cdot p_j},$$

die approximativ χ^2_{I-1}-verteilt ist.

Der Anpassungstest lässt sich verallgemeinern, indem wir zu einer Verteilung mit der Verteilungsfunktion F, gegen welche wir testen wollen, Intervalle $A_1 = (a_1, b_1], \ldots, A_r = (a_I, b_I]$ vorgeben und die Hypothesen zu

$$\begin{aligned} H_0 &: \mathbb{P}(X \in A_i) = p_i \qquad \text{für alle } 1 \leq i \leq I \\ H_1 &: \mathbb{P}(X \in A_i) \neq p_i \qquad \text{für ein } 1 \leq i \leq I \end{aligned}$$

anpassen. Wegen der Konvergenz der empirischen Verteilungsfunktion gegen die Verteilungsfunktion (Satz von Glivenko-Cantelli) spiegelt unsere erstellte Häufigkeitstabelle die Verteilung wieder.