Kontingentabellenanalyse

1. Elementares

Ausgangsbasis ist zunächst eine 2-dimensionale Kontingenztabelle mit den Häufigkeiten

	b_1	... b_j ...	b_l	Summe
a_1	n_{11}	... n_{1j} ...	n_{1l}	n_{1*}
... a_i ...	... n_{i1} ...	... ... n_{ij}... ...	... n_{il} ...	... n_{i*} ...
a_k	n_{k1}	... n_{kj} ...	n_{kl}	n_{k*}
Summe	n_{*1}	... n_{*j} ...	n_{*l}	n_{**}

zweier Merkmale a (mit *k* Ausprägungen) und b (mit *l* Ausprägungen). Unter der Hypothese, dass a und b voneinander unabhängig sind (also kein Zusammenhang besteht), ist die Wahrscheinlichkeit, dass die Kombination a_i und b_j auftritt:

$$p(a_i \text{ und } b_j) = p(a_i)\, p(b_j)$$

d.h. es ist zu erwarten, dass die Häufigkeiten sich wie folgt verhalten:

$n_{ij} \sim n_{i*} n_{*j} / n_{**}$ erwartete Unabhängigkeit

Daher heißen

$e_{ij} = n_{i*} n_{*j} / n_{**}$ Erwartungswerte

$r_{ij} = n_{ij} - e_{ij}$ Residuen

$\frac{(n_{ij} - e_{ij})}{\sqrt{e_{ij}}}$ standardisierte Rediduen

$\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}\left(1 - \frac{n_i}{n}\right)\left(1 - \frac{n_j}{n}\right)}}$ korrigierte standardisierte Residuen

Mittels des klassisichen χ^2-Tests (von Pearson) werden nun die beaobachteten Häufigkeiten n_{ij} mit den erwarteten e_{ij} verglichen und damit die Hypothese der Unabhängigkeit überprüft:

$$\chi^2 = \sum_{ij} \left(\frac{(n_{ij} - e_{ij})}{\sqrt{e_{ij}}} \right)^2$$

Diese Summe hat *(k-1)(l-1)* Freiheitsgrade. Sie ist allerdings nur dann χ^2-verteilt, wenn die einzelnen Quotienten innerhalb der Klammern, d.h. die standardisierten Residuen, annähernd normalverteilt sind.

Neben dem o.a. klassisichen χ^2-Test gibt es noch den Likelihood-Ratio χ^2-Test, LR-Test oder auch G-Test von Woolf genannt. Bei ihm werden nicht die Differenzen von beobachteten und

erwarteten Häufigkeiten verglichen, sondern deren Verhältnis:

$$G^2 = 2\sum_{ij} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

Diese Summe ist auch χ^2-verteilt mit *(k-1)(l-1)* Freiheitsgraden. Beide Testgrößen liefern annähernd identische Ergebnisse liefern.

2. Voraussetzungen und Alternativen

An dieser Stelle sind ein paar Anmerkungen zu den Voraussetzungen angebracht, d.h. unter welchen Bedingungen die standardisierten Residuen annähernd normalverteilt sind. Die dafür relevanten Größen sind die erwarteten Häufigkeiten e_{ij}. Die meisten Statistiker fordern heute:
- alle $e_{ij} >= 1$ und
- maximal 80% der $e_{ij} >= 5$.

Diese Regel stellte Cochran 1954 auf. Alan Agresti schwächte die zweite Regel in seinem Buch *Categorical Data Analysis* von 2002 so ab, dass der durchschnittliche Erwartungswert *n/(k*l) >= 5* sein muss, wobei *n* die Fallzahl ist. Der LR-Test wurde früher als robuster eingestuft, was aber nicht mehr gültig ist, wie Koehler and Larntz 1980 gezeigt haben. Für ihn gibt es u.a. die folgenden Verhaltensmuster (vgl. *A. Agresti*):
- Wenn viele $e_{ij} < 0,5$ sind, so reagiert der LR-Test konservativ (zu häufige Annahme von H_0),
- Wenn viele $0,5 < e_{ij} < 5$ sind, so reagiert der LR-Test liberal (zu häufige Annahme von H_1).

Diese Regeln gelten übrigens auch für die mehrdimensionalen χ^2-Tests (s.u.).

Welche Abhilfen gibt es nun, wenn die o.a. Voraussetzungen nicht erfüllt sind, einmal abgesehen von der trivialen Möglichkeit, Zeilen oder Spalten zusammenzufassen, um damit die Zeilen- oder Spaltensummen und damit die erwarteten Häufigkeiten zu vergrößern?

Die „klassische“ Methode ist die Yates- oder Kontinuitätskorrektur: Hierbei wird bei der Berechnung des χ^2-Wertes zu jedem Residuum $|n_{ij} - e_{ij}|$ eine Konstante, üblicherweise 0,5 oder 1 dazu addiert. In SPSS wurde diese Berechnung bei Kreuztabellen durchgeführt, in den neueren Versionen jedoch nicht mehr, und zwar wegen der nachfolgend beschriebenen Option. Die Yates-Korrektur wird jedoch in SPSS noch bei den loglinearen Modellen (*Loglinear -> Allgemein -> Modell*) angewandt. Dort ist dann ein Modell nur mit den beiden Haupteffekten, also kein saturiertes Modell, anzugeben. Der durchgeführte χ^2-Test beinhaltet dann die Korrektur. Vgl. dazu
- http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test
- http://en.wikipedia.org/wiki/Yates%27s_correction_for_continuity

Die andere Option ist der exakte Test, den SPSS zusätzlich zum asymptotischen anbietet. Bei ihm wird die Irrtumswahrscheinlichkeit (nahezu) exakt ausgerechnet. Es entfallen natürlich die Voraussetzungen, damit die standardisierten Residuen annähernd normalverteilt sind. Bei der Anwendung der Kreuztabellen in SPSS ist dazu der Button „Exakt“ zu betätigen. Dort gibt es dann zwei Alternativen: „Monte Carlo“ (nicht ganz so rechenintensiv) sowie „Exakt“ (sehr rechenintensiv), je nach Größe der Fallzahl *n*.

3. Analyse signifikanter χ^2-Werte

Liegt ein signifikanter χ^2-Wert vor, so interessiert, welche Zellen, d.h. welche Kombinationen von Werten a_i mit b_j, die „Verursacher“ sind. Hierzu können die o.a. standardisierten Residuen, besser noch die korrigierten standardisierten Residuen, herangezogen werden. Diese sind annähernd normalverteilt (mit Mittelwert 0 und Standardabweichung 1), so dass Werte außerhalb

des Bereichs -1,96 bis 1,96 als „auffällig“ eingestuft werden können. Bei positiven Residuen sind die entsprechenden Zellen auffällig stark besetzt, bei negativen Residuen auffällig schwach.

In SPSS werden 2-dimensionale Kontingenztabellen über *Deskriptive Statistiken -> Kreuztabellen* ausgegeben. Über die Option *Zellen* können alle o.a. Residuen ausgegeben werden. Falls über die Option *Statistiken* der χ^2-Test angefordert wird, so werden automatisch beide o.a. Tests durchgeführt.

Eine andere Möglichkeit der Analyse eines signifikanten Zusammenhangs besteht in der Aufteilung der Kontingenztabelle in Untertabellen. Dies empfiehlt sich, wenn spezifische apriori-Hypothesen vorliegen. Das entspricht in etwa den Kontrasten bei Varianzanalysen. Auch hier sind die Einzelvergleiche limitiert: Die Summe der Freiheitsgrade für die Einzelvergleiche darf nicht die Gesamtfreiheitsgrade *(k-1)(l-1)* übersteigen.

Im Zusammenhang mit nominalen Merkmalen wird ein Streuungs- bzw. Zusammenhangsmaß verwendet: die *Entropie*. Für ein Merkmal mit Wahrscheinlichkeiten p_i für die einzelnen Kategorien ist die Entropie H definiert:

$$H = -\sum p_i \log p_i$$

(*H=0* wenn ein $p_i=0$ bzw. *H=max* wenn alle p_i gleich groß sind.)
Mittels H lässt sich nun für ein Modell die Streuung errechnen und mittels Division durch die maximale Streuung (H_{max}) in ein Zusammenhangsmaß umrechnen, ähnlich wie bei der Varianzanalyse. Leider wird in SPSS H_{max} falsch errechnet.

Eine andere wichtige Statistik sind die Odds und Odds-Ratio. Setzt man die Häufigkeiten beider Ausprägungen b_1 und b_2 des dichotomen Merkmals *b* in Relation, z.B. n_{*1}/n_{*2}, so heißt dieser Quotient Odds und misst das Verhältnis des Auftretens von b_1 zum Auftreten von b_2 oder auch $\frac{p}{1-p}$. Vergleicht man nun diese Odds von b_1 für zwei Ausprägungen a_1 und a_2 eines anderen Merkmals *a*, so bildet man den Quotienten der beiden Odds, *Odds Ratio*, $\frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}$ von a_1 bzgl. a_2.

4. Mehrdimensionale Kontingenztabellen

Die Ausgabe mehrdimensionaler Tabellen empfiehlt es sich, ebenfalls die o.a. *Kreuztabellen* zu benutzen, da nur dort die relativen Häufigkeiten und Residuen ausgegeben werden. Wird der χ^2-Test angefordert, so wird dieser nur für den Zusammenhang der Spaltenvariablen mit der Zeilenvariablen durchgeführt. Werden z.B. mehrere Zeilenvariablen spezifiziert, so wird jede dieser mit der Spaltenvariablenvariablen in Beziehung gesetzt. Werden dagegen Schichtvariablen spezifiziert, so wird ein χ^2-Test für den Zusammenhang von Zeilen- und Spaltenvariablen durchgeführt, und zwar für jede Ausprägung bzw. Kombination von Ausprägungen der Schichtvariablen sowie ein globaler χ^2-Test, bei dem alle Schichtvariablen ignoriert werden.

5. Analyse mittels loglinearer Modelle

Vorab ein Wort zum Begriff des loglinearen Modells: Wird das o.a. Modell der Unabhängigkeit logarithmiert, so erhält man aus

$$n_{ij} = n_{i*} n_{*j} / n_{**}$$

$$\ln n_{ij} = \ln n_{i*} + \ln n_{*j} - \ln n_{**}$$

oder anders geschrieben erkennt man, wie durch das Logarithmieren sich das Modell als ein lineares Modell formulieren lässt:

$$\ln n_{ij} = \lambda_{A(i)} + \lambda_{B(j)} + \lambda_0$$

wobei wie bei der Varianzanalyse sowohl die $\lambda_{A(i)}$ als auch die $\lambda_{B(j)}$ sich zu 0 aufaddieren. Die λ geben die Abweichungen von Zeilen bzw. Spalten vom Mittel an und sind daher ein Maß für den Effekt von A bzw. B. Sie errechnen sich (in diesem einfachen Fall):

$$\lambda_0 = \sum_i \sum_j \ln n_{ij}$$

$$\lambda_i = \left(\sum_j \ln n_{ij} \right) / k - \lambda_0$$

SPSS nimmt allerdings eine andere Parametrisierung vor:

$$\ln n_{ij} = \lambda_{A(i)} I_{A(i)} + \lambda_{B(j)} I_{B(j)} + \lambda_0$$

wobei die $I_{A(i)}$ und $I_{B(j)}$ wie bei der Varianzanalyse Design- oder Indikatorvariablen mit den Werten 0 und 1 sind. Hiermit lassen sich entsprechend einem Zusammenhangsmodell die erwarteten Häufigkeiten schätzen. SPSS berechnet nur die nicht-redundanten Parameter, da die anderen in eine Berechnung der erwarteten Häufigkeiten nicht einfließen.

Mittels loglinearer Modelle werden ähnliche Fragestellungen gelöst wie mit der Varianzanalyse, nämlich der Zusammenhang mehrerer Merkmale, insbesondere die Analyse des Einflusses mehrerer nominaler Merkmale auf eine abhängige Variable, die bei der Varianzanalyse metrisch skaliert ist, hier jedoch nominal. Die Analyse gestaltet sich jedoch recht unterschiedlich: Bei der Varianzanalyse erhält man automatisch für jeden Einflussfaktor einen Signifikanztest sowie für alle möglichen Interaktionen. Daraus lässt sich das Zusammenhangsmodell ablesen als eine Summe von Effekten. Bei den loglinearen Modellen muss ein Zusammenhangsmodell vorgegeben werden, das dahingehend überprüft wird, ob es mit den Daten vereinbar ist. Ist das nicht der Fall, muss das Modell modifiziert und das erneut überprüft werden. Es gibt ein Modell, das immer passt: das *saturierte Modell*, das Zusammenhangsmodell, das alle Einflüsse und Interaktionen beinhaltet. Nur: aus der Tatsache, dass dieses Modell passt, kann nicht geschlossen werden, dass die darin enthaltenen Terme signifikante Bestandteile sind. Es muss vielmehr nach einem minimalen Modell mit möglichst wenigen Termen gesucht werden, das keine signifikante Abweichung mit den Daten zeigt.

Zur Überprüfung des Modells - inwieweit dieses mit den Daten vereinbar ist - wird üblicherweise, z.B. wegen der vorteilhafteren Voraussetzungen, der o.a. Likelihood-Ratio-Test verwendet, alternativ ist auch Pearson‘s χ^2-Test möglich. Wie auch im 2-dim. Fall werden die auf Basis des Modells geschätzten Häufigkeiten mit den beobachteten verglichen. Ein signifikantes Ergebnis indiziert ein unpassendes Modell.

Wie wird nun ein Modell aufgestellt? Angenommen, es liegt eine 3-dimensionale Kontingenztabelle der Variablen A, B und C vor. Ein einfacher Effekt, z.B. „A“ trägt den unterschiedlichen Häufigkeiten der einzelnen Ausprägungen von A Rechnung. Er sollte daher immer für jeden Faktor im Modell vorhanden sein. Ein Interaktionseffekt, z.B. A*B, steht für einen Zusammenhang zwischen A und B und ist symmetrisch. Da üblicherweise nur hierarchische Modelle betrachtet werden, impliziert A*B die einfachen Effekte A und B. Eine 3-er Interak-

tion, z.B. A*B*C impliziert analog alle darin enthaltenen 2-er Interaktionen A*B, A*C und B*C, sowie natürlich die Effekte A, B und C. Vorsicht: Während bei der Varianzanalyse der Effekt eines Faktors auf die abhängige Variable z.B. mit A bezeichnet wird, geschieht dies bei loglinearen Modellen immer über die Angabe der Interaktion mit der abhängigen Variablen, z.B. A*C. Analoges gilt für die Effektparameter $\lambda_{A(i)}$ (Varianzanalyse) bzw. $\lambda_{A(i)C}$ (loglineares Modell).

Bei der Modellsuche kann in SPSS *Loglinear -> Modellauswahl* helfen, dort die Option *Rückwärtselimination verwenden*. Ausgangsbasis ist das saturierte Modell (Voreinstellung). Zum einen wird schrittweise jeweils ein Term eliminiert, solange das restliche Modell passend bleibt. Zum anderen werden alle Terme ab eines bestimmten Grades (z.B. 2-er oder 3-er Interaktionen) dahingehend auf Signifikanz geprüft, ob durch eine Elimination des Terms das Modell unpassend wird (Optionen: *Assoziationstabelle*).

Während in SPSS *Loglinear -> Allgemein* beliebige loglineare Modelle untersucht werden, sind Logit-Modelle - ein Spezialfall der loglineare Modelle - in SPSS unter *Loglinear -> Logit* auf solche beschränkt, bei denen ähnlich der Varianzanalyse eine Variable *y* die abhängige darstellt und die übrigen Variablen x_i die unabhängigen Einflussfaktoren. Im Gegensatz zu den allgemeinen Modellen, in d enen die Zellenhäufigkeiten als Poisson-verteilt angenommen werden, werden diese bei Logit-Modellen als multinomial verteilt angenommen. Das Logit-Modell ist eine Verallgemeinerung des logistischen Modells von 2 auf *k* Ausprägungen der abhängigen Variablen *Y*. Es wird üblicherweise wie folgt dargestellt:

$$\log\left(\frac{P(y=j)}{P(y=k)}\right) = \alpha_j + \beta_{j1}x_1 + \ldots + \beta_{ji}x_i \quad mit\ j=1,\ldots,\ k\text{-}1$$

wobei *j* eine Ausprägung von *y* ist und α_j sowie $\beta_{j1},\ldots,\beta_{ji}$ Modellkoeffizienten sind.

Die Wahrscheinlichkeit $P(y=j)$, dass also *y* die Ausprägung *j* annimmt, errechnet sich daraus als:

$$P(y=j) = \frac{\exp(\alpha_j + \beta_{j1}x_1 + \ldots + \beta_{ji}x_i)}{1 + \sum_{\ldots}^{k-1} \exp(\alpha_m + \beta_{m1}x_1 + \ldots + \beta_{mi}x_i)}$$

Bei Logit-loglinearen Modellen können auch metrische Prädiktoren als *Kovariate* berücksichtigt werden.

Die Parameter λ, mit deren Hilfe ein signifikanter Effekt erklärt werden kann, können in SPSS mit Signifikanztest ausgegeben werden, allerdings nur die nicht-redundanten. Da sich die λ zu 0 aufsummieren, werden für die Haupteffekte *(k-1)* Parameter und für die Interaktionen *(k-1)(l-1)* ausgegeben. Und gerade letztere sind die besonders interessanten.

6. Abhängige Stichproben - Messwiederholungen

Statistische Modelle für den allgemeinen Fall (Messwiederholungen bei nominalen Merkmalen) gibt es zwar inzwischen (u.a. *event history models* von Nancy Tuma, verfügbar im Programm RATE), aber nicht in einer leicht anzuwendenden Form oder in Standardsoftware verfügbar. Für den Fall einer varianzanalytischen Fragestellung mit einer dichotomen abhängigen Variablen gibt es eine Reihe von Methoden, von Cochran‘s Q-Test bis zur Anwendung der klassischen parametrischen Varianzanalyse. Näheres dazu im Skript *Nichtparametrische Varianzanalysen - praktische Anwendung.* Somit besteht ein Notbehelf in der Dichotomisierung der abhängigen Variablen.

Haiko Lüpsen
Regionales Rechenzentrum der Universität zu Köln

29.5.2013