Kontingenztafeln

Ein ähnliches Prinzip wie bei der Prüfung auf Gleich- bzw. Normalverteilung wird auch beim Vergleich zweier qualitativer, mehrkategorieller („polytomer") Variablen angewendet. Hierfür werden die Messwerte in einer $c \times r$-Tabelle („Kontingenztafel") aufgelistet, wobei die c Spalten den verschiedenen Ausprägungen der ersten und die r Reihen den Kategorien der zweiten Variablen entprechen.

VORAUSSETZUNGEN

1.) Die Eintragungen sind absolute Häufigkeiten.
2.) Jeder Fall (jede Vp) geht nur einmal in die Daten ein.
3.) Alle Eintragungen sind wechselseitig unabhängig.
4.) Die erwarteten Werte sollten nicht zu klein sein ($e_{ij} \geq 5$), da sonst die χ^2-Verteilung nicht gut genug angenähert wird.

VORGEHENSWEISE

Hypothesen

$H_0 \ldots$ Die Verteilungen der beiden Stichproben sind gleich.

$H_1 \ldots$ Die beiden Verteilungen sind unterschiedlich.

absolute Häufigkeiten

Ermittlung der Summen für jede Kategorie:

$n \ldots$ Gesamtzahl der Einträge

$n_{i.} \ldots$ Zeilensummen

$n_{.j} \ldots$ Spaltensummen

$r \ldots$ Anzahl der Reihen („rows")

$c \ldots$ Anzahl der Spalten („columns")

$$\begin{array}{cccc|c} o_{11} & o_{12} & \dots & o_{1c} & n_{1.} \\ o_{21} & o_{22} & \dots & o_{2c} & n_{2.} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ o_{r1} & o_{r2} & \dots & o_{rc} & n_{r.} \\ \hline n_{.1} & n_{.2} & \dots & n_{.c} & n \end{array}$$

Die erwarteten Werte

Die e_{ij} werden anhand der Randsummen bestimmt:

$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Herleitung:

Die relativen Häufigkeiten sind $\frac{n_{.j}}{n}$ für die Zeilen und $\frac{n_{i.}}{n}$ für die Spalten. Um den erwarteten Anteil für eine bestimmte Kombination „Zeile $\times$ Spalte“ zu berechnen, werden diese einfach multipliziert ($\rightarrow$ $\frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n^2}$). Für den erwarteten absoluten Wert (e_{ij}) muss diese relative Häufigkeit mit der Gesamtzahl der Werte (n) multipliziert werden, was - nach Kürzen - obigen Term $\frac{n_{i.}n_{.j}}{n}$ ergibt.

Anmerkung:

Die Werte der letzten Zeile bzw. Spalte können auch berechnet werden, indem man von der jeweiligen Randsumme alle anderen Werte (der entsprechenden Zeile/Spalte) subtrahiert.

Testgröße

$$\chi^2 = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\text{mit } df = (r-1) \cdot (c-1)$$

Signifikanzprüfung

Verglichen wird diese Prüfgröße mit dem kritischen Wert (für α, $df = (r-1) \cdot (c-1)$, einseitige Fragestellung) aus der Tabelle. Die H_0 wird verworfen, falls $\chi^2 > \chi^2_{krit.}$.

KONTINGENZKOEFFIZIENT

Man kann obige Hypothesen auch folgendermaßen interpretieren:

$H_0 \ldots$ Es besteht kein Zusammenhang zwischen den untersuchten Variablen.

$H_1 \ldots$ Die beiden Variablen hängen zusammen.

Es wäre wünschenswert – insbesondere, falls der Test signifikant ist – ein Maß für die Stärke des Zusammenhangs zur Verfügung zu haben. Bei quantitativen Variablen ist dies die Korrelation, für qualitative Merkmale lernen wir nun den *Kontingenzkoeffizient* („coefficient of contingency“, CC) kennen:

Berechnung von CC

$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Korrektur: $CC \rightarrow CC^*$

Für $m = min(r, c)$ ist der maximale Wert, den CC annehmen kann:

$$CC_{max} = \sqrt{\frac{m-1}{m}}$$

Damit verschiedene Kontingenzkoeffizienten auch miteinander vergleichbar bleiben, setzt man die CC-Werte in Relation zu den maximal erreichbaren Werten CC_{max}:

$$CC^* = \frac{CC}{CC_{max}}$$

Eigenschaften, Unterschiede zur Korrelation

a) Der Kontingenzkoeffizient ist (im Gegensatz zur Korrelation) immer positiv. Man kann also am CC^*-Wert nicht erkennen, in welche Richtung der Zusammenhang geht, sondern nur, wie stark dieser ist.

b) Der Kontingenzkoeffizient ist immer *kleiner* als 1. Im Gegensatz dazu kann die Korrelation auch den Wert 1 annehmen.

c) Das Quadrat der (Produktmoment-)Korrelation kann als Bestimmtheitsmaß („erklärter Varianzanteil“) interpretiert werden. Für das Quadrat des Kontingenzkoeffizienten gibt es keine vergleichbare Interpretationsmöglichkeit.