

Vierfeldertafeln und Chi-Quadrat-Unabhängigkeitstest

Abstract

Werden in einer statistischen Erhebung zwei Merkmale A und B untersucht, die bei jeder Beobachtung entweder vorhanden sind oder nicht, können die Resultate in einer Vierfeldertafel (engl.: four-fold table, 2x2 frequency table, contingency table) zusammengefasst werden. Mit einem Chi-Quadrat-Unabhängigkeitstest (χ^2 -Unabhängigkeitstest) kann die Unabhängigkeit zwischen diesen Merkmalen A und B geprüft werden. Als Erweiterung ist dies auch möglich, wenn das Merkmal A in k_A und das Merkmal B in k_B verschiedenen Ausprägungen vorkommt.

Die für eine Behandlung im Unterricht notwendige Theorie wird für die Vierfeldertafeln an Beispielen entwickelt und zusammengestellt. Die Eigenschaften der Normalverteilung sowie der χ^2 -Verteilung sollten bereits bekannt oder wenigstens – beispielsweise mit geeigneten Simulationen – plausibel gemacht worden sein.

1. Zusammenhang zweier Merkmale

Bei vielen einfachen, in der Praxis häufig vorkommenden statistischen Erhebungen werden zwei Merkmale A und B untersucht, die bei jeder Beobachtung entweder vorhanden sind oder nicht. Es stellt sich die Frage, ob auf Grund einer solchen Stichprobe die Nullhypothese H_0 : "Die Merkmale A und B sind unabhängig voneinander" weiter beibehalten werden muss, oder ob sie zugunsten der Alternativhypothese H_1 : "Die Merkmale A und B sind nicht unabhängig voneinander" verworfen werden darf, was bedeutet, dass – mit dem unvermeidbaren Fehler 1. Art – auf einen Zusammenhang zwischen diesen beiden Merkmalen in der Grundgesamtheit geschlossen werden darf.

Die Unabhängigkeit der beiden Merkmale kann mit einem Chi-Quadrat-Unabhängigkeitstest (χ^2 -Unabhängigkeitstest) geprüft werden. Als Faustregel sollte dabei aber jede der erwarteten Häufigkeiten mindestens 5 und keine dieser Häufigkeiten Null sein, wodurch 20 die minimale untere Grenze der Beobachtungsanzahl in der Stichprobe darstellt. Für kleinere Werte muss die Kontinuitätskorrektur nach Yates (Frank Yates, 1902 – 1994) oder Fisher's exakter Test (R. A. Fisher, 1890 – 1962) verwendet werden, worauf hier aber nicht eingegangen werden soll.

Das Vorgehen und die notwendigen Rechenschritte werden an Hand von Beispielen demonstriert.

2. Ein Beispiel: Rauchen Männer häufiger als Frauen?

In einer Stichprobe seien 100 Frauen und Männer danach gefragt worden, ob sie rauchen.

Wir nehmen folgende Zahlen als Resultat dieser fiktiven Umfrage an:

| Merkmale: | A (Rauchend): | \bar{A} (Nicht rauchend): | Total Zeilen: |
|-------------------|---------------|-----------------------------|-------------------|
| B (Mann): | a = 20 | b = 40 | a + b = 60 |
| \bar{B} (Frau): | c = 8 | d = 32 | c + d = 40 |
| Total Spalten: | a + c = 28 | b + d = 72 | n = a+b+c+d = 100 |

Tabelle 1: Vierfeldertabelle für die Merkmale A: "Rauchend" und B: "Mann"

Aus der Tabelle wird klar, dass bei gegebenen, festen Zeilen- und Spalten-Summen (s. Tabelle 1: "Total Zeilen", resp. "Total Spalten") nur eine einzige der vier Zahlen a, b, c und d gegeben sein muss, woraus sich die anderen drei dann zwangsläufig ergeben: Dies gibt bereits einen ersten Hinweis auf die so genannte "Anzahl der Freiheitsgrade".

Nach der Nullhypothese H_0 sind die beiden zu A resp. B gehörigen Ereignisse voneinander unabhängig, woraus nach dem Multiplikationssatz für bedingte Wahrscheinlichkeiten folgt:

$$H_0: P(A|B) = P(A) \quad (\text{Gl. 1})$$

Diese Wahrscheinlichkeiten müssen aus den relativen Häufigkeiten geschätzt werden. Es müsste also – bei gültiger Nullhypothese H_0 – angenähert gelten:

$$\frac{a}{a+b} \approx \frac{a+c}{n} \quad (\text{Gl. 2})$$

Analoge Beziehungen ergeben sich für b, c und d.

3. Zu erwartende Werte

Für eine normalverteilte Grösse x mit Mittelwert 0 und Streuung 1 hat x^2 eine χ^2 -Verteilung mit dem Freiheitsgrad 1. Die kumulierten Werte dieser Verteilung sind tabelliert, und die zugehörigen Verteilungsfunktionen finden sich in vielen Büchern über Statistik; die benötigten Werte könnten bei Bedarf aber auch mit einer geeigneten Simulation gefunden werden. Weitergehende Kenntnisse des Funktionsverlaufs sind hier nicht nötig. Gebraucht wird hingegen der folgende Satz:

Sind x_1, x_2, \dots, x_n voneinander unabhängige, normalverteilte Grössen mit Mittelwert 0 und Streuung 1, so hat die Summe der Quadrate $\chi^2 = x_1^2 + x_2^2 + \dots + x_n^2$ eine χ^2 -Verteilung mit n Freiheitsgraden. Ist die Streuung nicht 1, sondern σ , so ist klar, dass die Grösse

$$\chi^2 = \sum_{k=1}^n \frac{x_k^2}{\sigma^2} \quad (\text{Gl. 3})$$

verwendet werden muss.

Der unter der Hypothese H_0 zu erwartende Wert a_e von a kann aus den Zeilen- und Spalten-Summen in der ursprünglichen Vierfeldertafel (Tabelle 1) gemäss der oben angegebenen Gleichung Gl. 2 berechnet werden. Zusammen mit den entsprechenden Gleichungen für b_e , c_e und d_e ergibt sich:

$$a_e = \frac{(a+b)(a+c)}{n}, \quad b_e = \frac{(a+b)(b+d)}{n}, \quad c_e = \frac{(a+c)(c+d)}{n} \quad \text{und} \quad d_e = \frac{(b+d)(c+d)}{n} \quad (\text{Gl. 4})$$

Die im gegebenen Beispiel zu erwartende Werte, die im Allgemeinen nicht ganzzahlig sein werden, sind in der folgenden Tabelle wiedergegeben:

| Merkmale: | A (Rauchend): | \bar{A} (Nicht rauchend): | Total: |
|-------------------|---------------|-----------------------------|---------------------|
| B (Mann): | $a_e = 16.8$ | $b_e = 43.2$ | $a + b = 60$ |
| \bar{B} (Frau): | $c_e = 11.2$ | $d_e = 28.8$ | $c + d = 40$ |
| Total: | $a + c = 28$ | $b + d = 72$ | $n = a+b+c+d = 100$ |

Tabelle 2: Vierfeldertafel mit den unter H_0 zu erwartenden Werten

4. Bestimmung der Testgrösse

Die Testgrösse χ^2 ergibt sich aus folgenden Überlegungen: Ist p die Wahrscheinlichkeit, dass in einer Beobachtung das Merkmal A auftritt, so ist die Anzahl a – unter der Annahme, dass die Nullhypothese H_0 gültig ist – binomialverteilt mit dem Mittelwert $(a + b) p$ und der Varianz $\sigma_a^2 = (a + b) p (1 - p)$. Das Entsprechende gilt für c , mit Mittelwert $(c + d) p$ und Varianz $\sigma_c^2 = (c + d) p (1 - p)$. Die unbekannte Wahrscheinlichkeit p muss aus den Häufigkeiten abgeschätzt werden:

$$p \approx \frac{a + c}{n}, \quad (\text{Gl. 5})$$

wobei n gleich der Gesamtzahl aller Beobachtungen ist: $n = a + b + c + d$.

Bei Gültigkeit der Hypothese H_0 sind die Quotienten $\frac{a}{a + b}$ und $\frac{c}{c + d}$ zwei voneinander unabhängige, normalverteilte Zufallsgrössen mit gleichem Mittelwert. Ihre Differenz

$$x := \frac{a}{a + b} - \frac{c}{c + d} \quad (\text{Gl. 6})$$

ist daher bekanntlich wiederum normalverteilt, und zwar mit dem Mittelwert Null und der Summe der Einzelvarianzen als Varianz:

$$\sigma^2 = \frac{\sigma_a^2}{(a + b)^2} + \frac{\sigma_c^2}{(c + d)^2} = p(1 - p) \left(\frac{1}{a + b} + \frac{1}{c + d} \right). \quad (\text{Gl. 7})$$

Wie oben bereits erwähnt (vgl. Gl. 3; Spezialfall), ist das Quadrat einer standardnormalverteilten Grösse χ^2 -verteilt mit Freiheitsgrad 1. Genau dies ist für den Quotienten $\left(\frac{x^2}{\sigma^2} \right)$ der Fall. Mit den konkreten, im Experiment gefundenen Werten von a , b , c und d kann darum gerade dieser Quotient $\left(\frac{x^2}{\sigma^2} \right)$ als Vergleichs- respektive Testgrösse χ^2 verwendet werden.

Setzen wir für p den Schätzwert $p = \frac{a + c}{n}$ gemäss Gl. 5 ein, ergibt sich dafür

$$\chi^2 = \left(\frac{x^2}{\sigma^2} \right) = \frac{\left(\frac{a}{a + b} - \frac{c}{c + d} \right)^2}{\frac{a + c}{n} \left(1 - \frac{a + c}{n} \right) \left(\frac{1}{a + b} + \frac{1}{c + d} \right)} \quad (\text{Gl. 8})$$

was etwas unhandlich erscheint, aber leicht vereinfacht werden kann:

$$\chi^2 = \frac{n \cdot (ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)} \quad (\text{Gl. 9})$$

Das so berechnete χ^2 ist auch gleich der Summe der Quotienten aus den Abweichungsquadraten der vier Zahlen von ihren Erwartungswerten und dem jeweiligen Erwartungswert selber, was sich durch Einsetzen sofort verifizieren lässt:

$$\chi^2 = \frac{(a - a_e)^2}{a_e} + \frac{(b - b_e)^2}{b_e} + \frac{(c - c_e)^2}{c_e} + \frac{(d - d_e)^2}{d_e} \quad (\text{Gl. 10})$$

5. Entscheidung

Wie oben schon angedeutet, ist die Zahl der Freiheitsgrade gleich 1, da aus den vier beobachteten Zahlen a, b, c und d die Wahrscheinlichkeit p_1 für das Merkmal A und die Wahrscheinlichkeit p_2 für das Merkmal B bestimmt wurden und ausserdem der Zusammenhang $a + b + c + d = n$ besteht.

In unserem Beispiel wird

$$\chi^2 = \frac{100 \cdot (20 \cdot 32 - 40 \cdot 8)^2}{(20 + 40)(20 + 8)(40 + 32)(8 + 32)} = 2.1164 \quad (\text{Gl. 11})$$

Dieses beobachtete χ^2 liegt weit unter dem Wert von 6.63, was einer kumulierten Wahrscheinlichkeit der χ^2 -Verteilung von 99% entspricht, aber auch unter der Grenze von 3.84, was einer kumulierten Wahrscheinlichkeit von 95% entspricht: Diese Zahlen sprechen – auf einem 5% – Niveau – nicht gegen die Nullhypothese H_0 : Männer und Frauen scheinen gleich häufig zu rauchen.

Anders sieht es aus, wenn – in einer zweiten Stichprobe – für $a = 12$, $b = 48$, $c = 16$ und $d = 24$ gefunden wird, was zu $\chi^2 = 4.7619 (> 3.84)$ führt. In diesem Fall ist die Nullhypothese auf einem 5%-Niveau zu verwerfen, wobei eine Irrtumswahrscheinlichkeit $\alpha < 5\%$ in Kauf genommen werden muss. Anschaulich argumentiert: Ein so grosser Wert für χ^2 lässt sich durch das "Wirken des Zufalls" nicht mehr erklären: Es besteht ein Zusammenhang zwischen den Merkmalen A und B.

Bei diesem Test kann nur zweiseitig getestet werden. Wenn aber zugunsten der Alternativhypothese H_1 entschieden werden muss, ergibt sich die Art des Zusammenhangs unmittelbar aus dem Zahlenmaterial selber: Frauen rauchen – in diesem zweiten fiktiven Beispiel – häufiger als Männer.

6. Erweiterung

Kommt das Merkmal A in k_A und das Merkmal B in k_B Ausprägungen vor, kann ein dem Vierfeldertest analoger Test für $k_A \cdot k_B$ – Felder durchgeführt werden. Die Prüfgrösse ergibt sich zu

$$\chi^2 = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (\text{Gl. 12})$$

wobei n_{ij} die Anzahl der Stichprobenelemente mit der Ausprägungskombination A_i und B_j bezeichnet, und e_{ij} deren erwartete Häufigkeit unter H_0 bezeichnet. Die Anzahl der Freiheitsgrade ist dabei $(k_A - 1) \cdot (k_B - 1)$.

7. Zusammenfassung

Vierfeldertafeln und der Chi-Quadrat-Unabhängigkeitstest liefern mit bescheidenem Aufwand Grundlagen, um eine Vielzahl praktisch wichtiger Fragen korrekt entscheiden zu können, wie z.B.:

- Überleben mit Medikament A mehr an der gleichen Krankheit Erkrankte das erste Jahr nach der Diagnose als mit Medikament B?
- Liefert Produzent A eines Artikels mehr Ausschuss als Produzent B?
- Unterscheiden sich die Aufnahmezahlen von Studentinnen und Studenten an einer Hochschule?
- etc.

Wegen des guten Verhältnisses zwischen didaktischem Aufwand und erzielbarem Ertrag sowohl für die Theorie als auch für die Praxis sollte dieses Thema in keinem Statistikkurs fehlen.

Literatur:

Theorie:

Van der Waerden, B.L., Mathematische Statistik, Springer-Verlag 1957,

Weiss, C., Basiswissen Medizinische Statistik, Springer-Verlag 2001,

Bortz, J., Lehrbuch der Statistik, Springer-Verlag 1979.

Leicht verständliche Einführung in die praktische Anwendung statistischer Methoden, mit vielen Beispielen, in denen reale, publizierte Zahlen aus dem medizinischen Bereich verwendet werden:

Altman Douglas G., Practical Statistics for Medical Research, Chapman & Hall, London, 1991.

Lehrbuch zur schliessenden Statistik auf einführendem Niveau für Studenten der Wirtschaftswissenschaften, mit vielen durchgerechneten praktischen Beispielen aus verschiedenen Bereichen:

Polasek Wolfgang, Schliessende Statistik, Springer-Verlag 1996.