Prof. Dr. Klaus Eckhardt

Kontingenzanalyse

veröffentlicht im Internet unter aufgabomat.de

Inhalte: Kontingenz- bzw. Kreuztabelle, Randsummen, Chi-Quadrat-Unabhängigkeitstest, Chi-Quadrat-Verteilung, Kontingenzkoeffizient, korrigierter Kontingenzkoeffizient

Gliederung:

1 Einleitung 1
2 Kontingenz- bzw. Kreuztabelle 2
3 Chi-Quadrat-Unabhängigkeitstest 3
4 Kontingenzkoeffizient 6

1 Einleitung

Die Kontingenzanalyse dient der Untersuchung, ob zwischen zwei nominal- oder ordinalskalierten Merkmalen bzw. Variablen ein Zusammenhang besteht. Es könnte beispielsweise um die folgenden Themen gehen:

- Wirksamkeit unterschiedlicher Medikamente
- Erkrankungsrisiken in Abhängigkeit von der Lebensführung
- Erwerbslosigkeit in Abhängigkeit vom Bildungsabschluss
- Konsumverhalten in unterschiedlichen Altersgruppen
- Parteienpräferenzen von Männern und Frauen.

Teil der Kontingenzanalyse, so wie sie hier dargestellt wird, ist der Chi-Quadrat-Test, der unter anderem auch als Verteilungstest[1] verwendet wird. Deutet sich in Verbindung mit der Kontingenzanalyse kein Zusammenhang zwischen den untersuchten Variablen an, so ergibt sich daraus zwangsläufig die Folgerung, dass sie als voneinander unabhängig anzusehen sind. Man spricht daher auch vom Chi-Quadrat-Unabhängigkeitstest. Die stochastische Unabhängigkeit von Zufallsvariablen ist nun aber eine notwendige Voraussetzung vieler weiterer statistischer Verfahren wie den t-Test oder den F-Test. Mithilfe des Chi-Quadrat-Tests lässt sich prüfen, ob diese Voraussetzung erfüllt ist.

Der Chi-Quadrat-Test liefert allerdings nur dann korrekte Ergebnisse, wenn eine ausreichend große Anzahl N von Messwerten vorliegt (Faustregel: $N > 50$). Ist dies nicht der Fall, so muss auf einen anderen Test zurückgegriffen werden. Weiterführende Informationen dazu sind der Fachliteratur zu entnehmen.

[1] Siehe beispielsweise Skript „Chi-Quadrat-Verteilungstest“ unter aufgabomat.de.

2 Kontingenz- bzw. Kreuztabelle

Gegeben seien zwei Zufallsvariablen X und Y. Die Variable X trete in I unterschiedlichen Ausprägungen x_i auf, die Variable Y in J unterschiedlichen Ausprägungen y_j. In einer Datenerhebung wird erfasst, mit welcher absoluten Häufigkeit Kombinationen dieser Merkmalsausprägungen x_i und y_j vorkommen.

Beispiel: Zur Behandlung einer Krankheit wurde ein neues Medikament entwickelt. Um zu untersuchen, ob es den Behandlungserfolg erhöht, werden zwei Probandengruppen gebildet. Gruppe A wird mit dem Medikament behandelt, Gruppe B mit einem Placebopräparat, d. h. einem Pseudomedikament ohne Wirkstoff. Ein solches Experiment wird als **Blindversuch** bezeichnet. Nach einem vorgegebenen Zeitraum wird der Gesundheitszustand der Probanden beurteilt: Hat sich der Zustand weiter verschlechtert, ist er unverändert geblieben, ist eine Besserung eingetreten oder ist die jeweilige Person sogar geheilt? In diesem Beispiel ist folglich

- Merkmal X: Behandlungsmethode in den I = 2 Ausprägungen x_1 = A und x_2 = B
- Merkmal Y: Gesundheitszustand in den J = 4 Ausprägungen y_1 = verschlechtert, y_2 = unverändert, y_3 = verbessert und y_4 = geheilt.

Bei I Ausprägungen des Merkmals X und J Ausprägungen des Merkmals Y gibt es insgesamt I · J verschiedene Kombinationen der Merkmalsausprägungen x_i und y_j. In einer **Kontingenz- bzw. Kreuztabelle** wird aufgeführt, mit welcher absoluten Häufigkeit diese Merkmalskombinationen im Datensatz vorkommen (Tabelle 1). Man spricht auch von einer I×J-Tabelle („I Kreuz J-Tabelle“). Die Tabelle wird ergänzt um eine zusätzliche Zeile und eine zusätzliche Spalte, in denen die jeweiligen **Randsummen** aufgeführt werden. Dabei handelt es sich um die Summen aller Werte in jeweils einer Zeile (**Zeilensummen**) und die Summen aller Werte in jeweils einer Spalte der Tabelle (**Spaltensummen**). Die Randsummen stellen die absolute Häufigkeit dar, mit der die Merkmalsausprägungen x_i und y_j jeweils für sich betrachtet aufgetreten sind.

		Merkmal Y				
		y_1	y_2	…	y_J	Randsumme
Merkmal X	x_1	n_{11}	n_{12}	…	n_{1J}	N_1
	x_2	n_{21}	n_{22}	…	n_{2J}	N_2
	…	…	…	…	…	…
	x_I	n_{I1}	n_{I2}	…	n_{IJ}	N_I
	Randsumme	M_1	M_2	…	M_J	

Tabelle 1: I×J-Kontingenztabelle für zwei Merkmale X und Y.

- x_i: Ausprägungen des Merkmals X, i = 1, …, I
- y_j: Ausprägungen des Merkmals Y, j = 1, …, J
- n_{ij}: Anzahl der Beobachtungen mit der Merkmalskombination X = x_i und Y = y_j
- Randsumme N_i: absolute Häufigkeit des Vorkommens der Merkmalsausprägung x_i
- Randsumme M_j: absolute Häufigkeit des Vorkommens der Merkmalsausprägung y_j

Beispiel:

	verschlechtert	unverändert	verbessert	geheilt	Summe
A	19	20	26	25	90
B	42	51	16	28	137
Summe	61	71	42	53	

Tabelle 2: Beispiel für eine Kontingenztabelle.

Es lässt sich in der Regel eher beurteilen, ob eine Abhängigkeit zwischen den Merkmalen bzw. Variablen vorliegen könnte, wenn man die absoluten Häufigkeiten in relative Häufigkeiten

umrechnet. Im vorliegenden Beispiel ist die absolute Zahl der geheilten Probanden in der Gruppe B zwar höher als in der Gruppe A, Gruppe B umfasst aber zugleich eine insgesamt größere Anzahl von Personen, sodass dieser Befund wenig aussagekräftig ist. Teilt man die Werte in jeder Zeile durch die zugehörige Zeilensumme ($N_1 = 90$, $N_2 = 137$), so wird ersichtlich, dass die Zahl der Probanden, deren Zustand sich gebessert hat oder die geheilt wurden, in der Gruppe B vergleichsweise geringer ist als in der Gruppe A (Tabelle 3).

	verschlechtert	unverändert	verbessert	geheilt
A	0,21	0,22	0,29	0,28
B	0,31	0,37	0,12	0,20

Tabelle 3: Relative Häufigkeiten des Gesundheitszustands in den beiden Behandlungsgruppen.

Erweckt die erste Analyse der Daten den Eindruck, dass ein Zusammenhang vorliegen könnte, schließen sich zwei Fragen an:

- Lässt sich der Verdacht eines Zusammenhangs erhärten? Kann davon ausgegangen werden, dass der Zusammenhang, der sich in den Stichprobenwerten zu zeigen scheint, kein bloßer Zufall ist? Diese Frage wird mithilfe des **Chi-Quadrat-Unabhängigkeitstests** beantwortet (Abschnitt 3).
- Falls der Test auf einen Zusammenhang schließen lässt: Wie ist dessen Stärke zu beurteilen? Zu diesem Zweck wird der so genannte **Kontingenzkoeffizient** berechnet (Abschnitt 4).

3 Chi-Quadrat-Unabhängigkeitstest

1. Null- und Alternativhypothese formulieren
   Irrtumswahrscheinlichkeit festlegen

   Im Fall des Chi-Quadrat-Unabhängigkeitstests lauten Null- und Alternativhypothese
   H_0: X und Y sind voneinander unabhängig.
   H_1: X und Y sind voneinander abhängig.

   Die Wahl der Irrtumswahrscheinlichkeit ist abhängig vom Zweck des Tests. Angenommen, der Test wird durchgeführt, um die Voraussetzung der stochastischen Unabhängigkeit bei weitergehenden Analysen zu prüfen. In diesem Fall ist es besonders kritisch, wenn die Nullhypothese irrtümlich beibehalten wird, denn in diesem Fall würden die nachfolgenden Auswertungen unter einer falschen Annahme durchgeführt und unwissentlich zu unbrauchbaren Ergebnissen führen. Die Wahrscheinlichkeit für den Fehler zweiter Art sollte daher klein sein. Die einzige Möglichkeit, dies zu erreichen, ist, die Wahrscheinlichkeit für den Fehler erster Art, die Irrtumswahrscheinlichkeit α, relativ groß zu wählen. In diesem Fall würde man darum eher mit $\alpha = 0{,}10$ als mit dem meistgebrauchten $\alpha = 0{,}05$ arbeiten. Im vorliegenden Beispiel stellt sich dagegen die Frage nach den Kosten des Medikaments und eventuellen Nebenwirkungen. Daher sollte hier vorrangig die Wahrscheinlichkeit reduziert werden, die Nullhypothese irrtümlich zu verwerfen und damit irrtümlich von einer positiven Wirkung des Medikaments auszugehen. Im Beispiel wird deshalb $\alpha = 0{,}01$ gesetzt.

2. Ausgehend von der Nullhypothese erwartete absolute Häufigkeiten n_{ij}^* berechnen

   Die Nullhypothese H_0 lautet, dass X und Y voneinander unabhängig sind. Nach dem Multiplikationssatz der Wahrscheinlichkeitsrechnung für stochastisch unabhängige Ereignisse gilt für die Wahrscheinlichkeit des Auftretens der Merkmalskombination $(x_i ; y_j)$

$$P[(X = x_i) \cap (Y = y_j)] = P(X = x_i)\, P(Y = y_j), \tag{1}$$

wobei P als Abkürzung für das lateinische Wort „probabilitas“ (Wahrscheinlichkeit) steht. Im Folgenden sei N die Gesamtzahl der in der Datenerhebung erfassten Werte:

$$\begin{aligned} N &= \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \\ &= n_{11} + n_{12} + \ldots + n_{1J} + n_{21} + n_{22} + \ldots + n_{2J} + \ldots n_{I1} + n_{I2} + \ldots + n_{IJ}. \end{aligned}$$

Die relativen Häufigkeiten, mit denen die Merkmalsausprägungen x_i und y_j im Datensatz vorkommen, sind Schätzwerte für $P(X = x_i)$ und $P(Y = y_j)$. Unter Zuhilfenahme der Randsummen N_i und M_j:

$$P(X = x_i) \approx \frac{N_i}{N} \tag{2}$$

$$P(Y = y_j) \approx \frac{M_j}{N}\,. \tag{3}$$

Die bei Unabhängigkeit der Merkmale bzw. Variablen zu erwartende absolute Häufigkeit für die Merkmalskombination $(x_i\,;\, y_j)$ ergibt sich, indem man die Wahrscheinlichkeit $P[(X = x_i) \cap (Y = y_j)]$ für die Merkmalskombination mit der Gesamtzahl N der erfassten Werte multipliziert:

$$\begin{aligned} n_{ij}^* &= P[(X = x_i) \cap (Y = y_j)]\, N \\ &\approx \frac{N_i}{N}\,\frac{M_j}{N}\, N \\ &\approx \frac{N_i\, M_j}{N} \end{aligned} \tag{4}$$

Beispiel: Merkmal X: Behandlungsmethode, Merkmal Y: Gesundheitszustand, N = 227 (Tabelle 2)
Zeilensummen: $N_1 = 90$, $N_2 = 137$
Spaltensummen: $M_1 = 61$, $M_2 = 71$, $M_3 = 42$, $M_4 = 53$

$$n_{11}^* = \frac{N_1 M_1}{N} = \frac{90 \cdot 61}{227} = 24 \qquad n_{12}^* = \frac{N_1 M_2}{N} = \frac{90 \cdot 71}{227} = 28 \qquad n_{13}^* = \frac{N_1 M_3}{N} = \frac{90 \cdot 42}{227} = 17 \qquad n_{14}^* = \frac{N_1 M_4}{N} = \frac{90 \cdot 53}{227} = 21$$

$$n_{21}^* = \frac{N_2 M_1}{N} = \frac{137 \cdot 61}{227} = 37 \qquad n_{22}^* = \frac{N_2 M_2}{N} = \frac{137 \cdot 71}{227} = 43 \qquad n_{23}^* = \frac{N_2 M_3}{N} = \frac{137 \cdot 42}{227} = 25 \qquad n_{24}^* = \frac{N_2 M_4}{N} = \frac{137 \cdot 53}{227} = 32$$

	verschlechtert	unverändert	verbessert	geheilt
A	24	28	17	21
B	37	43	25	32

Tabelle 4: Bei Gültigkeit von H_0 zu erwartende absolute Häufigkeiten.

3. Klassen so zusammenfassen, dass n_{ij}^* überall mindestens 5 beträgt

   Die dadurch eventuell reduzierte Anzahl der Merkmalsausprägungen wird im Folgenden mit I* und J* bezeichnet.

4. Teststatistik berechnen
   Gültigkeit der Nullhypothese beurteilen

   Nun wird der Wert der so genannten **Teststatistik** berechnet, einer Zufallsvariable, die dann, wenn die Nullhypothese des Tests gilt, bekannte Eigenschaften aufweist. Der Formulierung der Teststatistik liegen die folgenden Gedanken zugrunde:
   - Je ähnlicher n_{ij} und n_{ij}^* sind, d. h. je kleiner die Differenzen $n_{ij} - n_{ij}^*$, desto wahrscheinlicher gilt die Nullhypothese.
   - Um zu verhindern, dass sich negative und positive Differenzen gegenseitig aufheben, werden die Differenzen quadriert.
   - Eine Differenz fällt umso weniger ins Gewicht, je mehr Werte ohnehin in der jeweiligen Klasse bzw. Merkmalskombination zu erwarten sind. Daher werden die quadrierten Differenzen noch durch n_{ij}^* geteilt.

   Die Teststatistik des Chi-Quadrat-Unabhängigkeitstests ist

$$\chi^2 = \sum_{i=1}^{I^*} \sum_{j=1}^{J^*} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} \tag{5}$$

   Das Symbol χ ist der griechische Großbuchstabe Chi.

   Falls die Nullhypothese H_0 zutrifft und Anzahl der Messwerte genügend groß ist (Faustregel: N > 50), ist χ^2 eine Chi-Quadrat-verteilte Zufallsvariable. Aus Gleichung 5 ist ersichtlich, dass Chi-Quadrat-verteilte Variablen nur Werte ≥ 0 annehmen. Die Wahrscheinlichkeitsdichtefunktion der **Chi-Quadrat-Verteilung** ist daher nach links durch die Null begrenzt, nach rechts aber unbegrenzt und damit asymmetrisch. Entsprechendes gilt für die zugehörige Verteilungsfunktion (Abbildung 1).

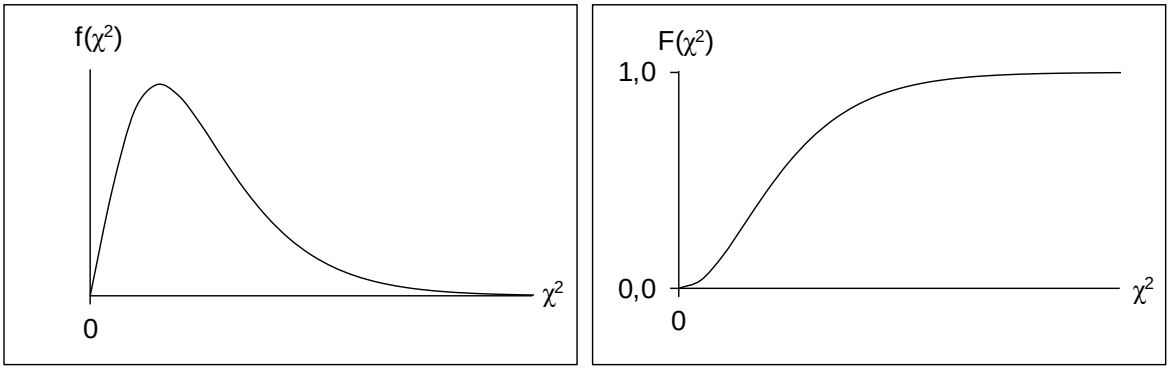

Abbildung 1: Wahrscheinlichkeitsdichte- und Verteilungsfunktion der Chi-Quadrat-Verteilung.

   Die Chi-Quadrat-Verteilung hat nur einen Parameter, den so genannten **Freiheitsgrad**, für den hier symbolisch f geschrieben wird. Im Fall des Unabhängigkeitstests ist

$$f = (I^* - 1)\,(J^* - 1). \tag{6}$$

Beispiel:
$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*}$$
$$= \frac{(n_{11}-n_{11}^*)^2}{n_{11}^*} + \frac{(n_{12}-n_{12}^*)^2}{n_{12}^*} + \frac{(n_{13}-n_{13}^*)^2}{n_{13}^*} + \frac{(n_{14}-n_{14}^*)^2}{n_{14}^*}$$
$$+ \frac{(n_{21}-n_{21}^*)^2}{n_{21}^*} + \frac{(n_{22}-n_{22}^*)^2}{n_{22}^*} + \frac{(n_{23}-n_{23}^*)^2}{n_{23}^*} + \frac{(n_{24}-n_{24}^*)^2}{n_{24}^*}$$
$$= \frac{(19-24)^2}{24} + \frac{(20-28)^2}{28} + \frac{(26-17)^2}{17} + \frac{(25-21)^2}{21}$$
$$+ \frac{(42-37)^2}{37} + \frac{(51-43)^2}{43} + \frac{(16-25)^2}{25} + \frac{(28-32)^2}{32}$$
$$= 14{,}8$$

Je kleiner der Wert der Teststatistik ist, desto wahrscheinlicher ist es, dass die Nullhypothese zutrifft. $\chi^2 = 0$ bedeutet, dass keinerlei Unterschied zwischen empirischer und erwarteter Häufigkeitsverteilung besteht. Der Chi-Quadrat-Unabhängigkeitstest wird daher rechtsseitig durchgeführt, d. h. das Annahmeintervall wird nur nach rechts durch ein Quantil der Chi-Quadrat-Verteilung begrenzt. Entsprechende Quantilwerte lassen sich beispielsweise einer Quantiltabelle entnehmen (Tabelle 5).

f	p		
	0,90	0,95	0,99
1	2,7	3,8	6,6
2	4,6	6,0	9,2
3	6,3	7,8	11,3
4	7,8	9,5	13,3
5	9,2	11,1	15,1
6	10,6	12,6	16,8
7	12,0	14,1	18,5
8	13,4	15,5	20,1
9	14,7	16,9	21,7
10	16,0	18,3	23,2

f	p		
	0,90	0,95	0,99
11	17,3	19,7	24,7
12	18,5	21,0	26,2
13	19,8	22,4	27,7
14	21,1	23,7	29,1
15	22,3	25,0	30,6
16	23,5	26,3	32,0
17	24,8	27,6	33,4
18	26,0	28,9	34,8
19	27,2	30,1	36,2
20	28,4	31,4	37,6

Tabelle 5: Quantile χ^2_p der Chi-Quadrat-Verteilung.

Beispiel: 0,99-Quantil der Chi-Quadrat-Verteilung mit f = (2 – 1) (4 – 1) = 3: $\chi^2_{0,99} = 11{,}3$
$\chi^2 > \chi^2_{0,99} \Rightarrow H_0$ wird verworfen.
Das neue Medikament scheint zu wirken.

4 Kontingenzkoeffizient

Zur Charakterisierung der Stärke eines Zusammenhangs zwischen nominal- oder ordinalskalierten Merkmalen X und Y dient der **Kontingenzkoeffizient**

$$K = \sqrt{\frac{\chi^2}{\chi^2 + N}}\,. \qquad (7)$$

Darin ist χ^2 der Wert der Teststatistik des Chi-Quadrat-Unabhängigkeitstests (Abschnitt 3, Gleichung 5) und N der Stichprobenumfang. Der Wertebereich des Kontingenzkoeffizienten ist $[0; K_{max}]$ mit

$$K_{max} = \sqrt{\frac{\min\{I;J\}-1}{\min\{I;J\}}}\,. \tag{8}$$

I ist die Anzahl der Ausprägungen, die das Merkmal X annimmt, J die Anzahl der Ausprägungen, die das Merkmal Y annimmt (Abschnitt 2).

Der Wertebereich des Kontingenzkoeffizienten ist damit nach oben hin unbegrenzt, was es im Allgemeinen unmöglich macht, Kontingenzkoeffizienten unterschiedlicher Paare von Merkmalen miteinander sinnvoll zu vergleichen. Der Kontingenzkoeffizient wird daher normiert:

$$K_{korr} = \frac{K}{K_{max}}\,. \tag{9}$$

K_{korr} ist der so genannte **korrigierte Kontingenzkoeffizient**. Durch die Normierung hat er den Wertebereich [0; 1].

Beispiel: Merkmal X: Behandlungsmethode, Merkmal Y: Gesundheitszustand, N = 227 (Tabelle 2)

$$K = \sqrt{\frac{14{,}8}{14{,}8+227}}$$
$$= 0{,}25$$

$$K_{max} = \sqrt{\frac{\min\{2;4\}-1}{\min\{2;4\}}}$$
$$= 0{,}71$$

$$K_{korr} = \frac{0{,}25}{0{,}71}$$
$$= 0{,}35$$

Zu beachten ist:

- Ein hoher Kontingenzkoeffizient ist kein Beweis für einen kausalen Zusammenhang, sondern lediglich ein Hinweis darauf, dass ein solcher Zusammenhang bestehen könnte.
- Ein Einzelwert des Kontingenzkoeffizienten ist in der Regel kaum zu interpretieren. Sinnvoll ist die Berechnung solcher Werte in erster Linie für Vergleiche.

Die Eingangsdaten des in diesem Skript behandelten Beispiels und weitere Übungsaufgaben finden sich im Internet unter der Adresse aufgabomat.de in der Rubrik Statistik.