9: Kreuztabellen

Statistische Verfahren in der Geographie

Till Straube <straube@geo.uni-frankfurt.de>

Institut für Humangeographie
Goethe-Universität Frankfurt

1 Lernziele dieser Sitzung

Sie können…

- eine Kreuztabelle erstellen und interpretieren.
- den Kontingenzkoeffizienten χ^2 errechnen.
- die Maßzahlen ϕ bzw. CI errechnen und interpretieren.

2 Bivariate Verteilungen mit nominalen Variablen

In der bivariaten Statistik (Sitzungen 7 und 8) ging es bisher um Zusammenhänge zwischen zwei metrischen Variablen. In dieser Sitzung geht es um statistische Verfahren der bivariaten Statistik, bei denen für beide Variablen nur das Nominalskalenniveau vorausgesetzt ist. (Für Skalenniveaus s. Sitzung 1.)

Mit den Werten von nominalen Variablen lassen sich die in Sitzung 7 und 8 besprochenen Parameter (z.B. Kovarianz) nicht errechnen, weil wir mit ihnen nicht die notwendigen Rechenoperationen (Addition, Subtraktion) durchführen können. Stattdessen sind die beobachteten Häufigkeiten Ausgangslage für die im Folgenden besprochenen Verfahren.

2.1 Beispiel

Wir fragen uns, ob es einen Zusammenhang zwischen dem Studienfach von Studierenden an einer Universität und ihrem präferierten Transportmittel für den Pendelweg zum Campus gibt. Insbesondere interessiert uns, ob ein Zusammenhang zwischen dem Studium der Geistes- und Sozialwissenschaften und der Fahrradnutzung besteht.

Beide Variablen sind nominalskaliert: die erhobenen Werte können in Kategorien eingeordnet werden (Studienfach: Geographie, Politikwissenschaft, BWL, …; Transportmittel: Bus, Fahrrad, zu Fuß, …).

Um die Variablen im Sinne unserer Fragestellung zu vereinfachen, wandeln wir beide Variablen in *dichotome* Variablen um (die dann nur zwei Werte annehmen können). Wir beschränken uns auf die Erhebung von „Fahrrad“ oder „anderes Transportmittel“ einerseits und „Geistes-/Sozialwissenschaft“ oder „anderes Studienfach“ andererseits. Die (verkürzte) Tabelle der Rohdaten einer Zufallsstichprobe der Größe $n = 90$ könnte dann so aussehen wie Tabelle 1.

Tabelle 1: Ungeordnete Rohdaten der Erhebung

	Studienfach	Transportmittel
1	anderes Studienfach	Fahrrad
2	anderes Studienfach	Fahrrad
3	Geistes-/Sozialwissenschaft	anderes Transportmittel
4	Geistes-/Sozialwissenschaft	anderes Transportmittel
5	Geistes-/Sozialwissenschaft	anderes Transportmittel
6	Geistes-/Sozialwissenschaft	Fahrrad
...	...	...
85	anderes Studienfach	anderes Transportmittel
86	anderes Studienfach	anderes Transportmittel
87	Geistes-/Sozialwissenschaft	anderes Transportmittel
88	anderes Studienfach	anderes Transportmittel
89	anderes Studienfach	anderes Transportmittel
90	Geistes-/Sozialwissenschaft	anderes Transportmittel

Tabelle 2: Kreuztabelle der Beispieldaten

	Fahrrad	anderes Transportmittel	
Geistes-/Sozialwissenschaft	11	28	39
anderes Studienfach	9	42	51
	20	70	90

3 Kreuztabelle

Die Kreuztabelle (auch Kontingenztabelle, Kontingenztafel, engl. *contingency table*) ist eine übersichtliche Zusammenfassung der Rohdaten. Sie spannt die beiden Variablen in Spalten- und Zeilenrichtung auf, so dass in jeder Zelle die Häufigkeit einer bestimmten Wertekombination steht.

Bei zwei dichotomen Variablen ergeben sich zwei Spalten und zwei Zeilen, also vier Tabellenfelder. Wir sprechen in diesem Fall auch von einer 2×2-Tabelle.

3.1 Beispiel

Die Kreuztabelle für unser Beispiel ist in Tabelle 2 dargestellt. Die Spaltenüberschriften sind die beiden Werte der dichotomen Variable „Transportmittel“, und die Zeilennamen sind die beiden Werte für „Studienfach“. In den Zellen stehen die Häufigkeiten. Es lässt sich also z.B. ablesen, dass die Kombination „Fahrrad“ und „anderes Studienfach“ neun mal vorkommt.

Am rechten Rand der Tabelle stehen die Summen für die Zeilen, am unteren Rand die Summen der Spalten. Ganz unten rechts steht die Gesamtsumme (Größe der Stichprobe).

Softwarehinweis

In R kann eine einfache Kreuztabelle mit dem Befehl `table()` ausgegeben werden.

Tabelle 3: Allgemeine Bezeichnungen in der Kreuztabelle

	Spalte 1	Spalte 2	...	Spalte ℓ	
Zeile 1	n_{11}	n_{12}	...	$n_{1\ell}$	$n_{1\cdot}$
Zeile 2	n_{21}	n_{22}	...	$n_{2\ell}$	$n_{2\cdot}$
...	...	...	...	...	...
Zeile k	n_{k1}	n_{k2}	...	$n_{k\ell}$	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot\ell}$	n

3.2 Verallgemeinerung

In Tabelle 3 ist das allgemeingültige Format für Kreuztabellen festgehalten. Dabei sind folgende Besonderheiten zu beachten:

- Das Symbol k steht für die Anzahl der Zeilen, ℓ für die Anzahl der Spalten.
- Die Häufigkeiten für Merkmalskombinationen in den Tabellenfeldern werden durch n_{ij} symbolisiert, wobei i für die laufende Nummer der Zeile steht, und j für die laufende Nummer der Spalte.
- Die Teilsummen an den Rändern werden mit Punktnotation bezeichnet. Dabei steht die Zeilensumme $n_{i\cdot}$ für die Summe *aller* Felder in Zeile i (Zeilensumme) und $n_{\cdot j}$ für die Summe *aller* Felder in Spalte j (Spaltensumme).
- Die Gesamtsumme unten rechts wird hier mit n gekennzeichnet und steht wie gewohnt für die Gesamtgröße der Stichprobe.

4 Berechnung der erwarteten Werte

Bestünde *kein* Zusammenhang zwischen den Variablen, dann wäre zu erwarten, dass sich die Kombinationen gleichmäßig auf die Tabellenfelder aufteilen, und zwar ausgehend von den Teilsummen für die Zeilen und Spalten.

Der Erwartungswert für ein Tabellenfeld (also der „durchschnittliche“ Wert, wenn es keinen Zusammenhang zwischen den beiden Variablen gibt) berechnet sich durch die Formel:

$$m_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \tag{1}$$

Es wird also das Produkt der Zeilen- und der Spaltensumme geteilt durch die Gesamtsumme.

4.1 Beispiel

Die beobachtete Häufigkeit für die Kombination „Geistes-/Sozialwissenschaft“ (Zeile 1) und „anderes Transportmittel“ (Spalte 2) ist 28. Aber was wäre der Erwartungswert bei den gegebenen Summen? Wir

Tabelle 4: Kreuztabelle der Beispieldaten

	Fahrrad	anderes Transportmittel	
Geistes-/Sozialwissenschaft	11	28	39
	(8,67)	(30,33)	
anderes Studienfach	9	42	51
	(11,33)	(39,67)	
	20	70	90

setzen einfach die entsprechenden Werte in die Gleichung 1 ein:

$$\begin{aligned} m_{12} &= \frac{n_{1\cdot} \cdot n_{\cdot 2}}{n} \\ &= \frac{39 \cdot 70}{90} \\ &\approx 30{,}33 \end{aligned}$$

Diese Rechnung lässt sich für alle Tabellenfelder durchführen. Die Kreuztabelle kann dann um diese erwarteten Werte in Klammern ergänzt werden (s. Tabelle 4).

5 Berechnung des Kontingenzkoeffizenten χ^2

Sind für alle Tabellenfelder die Beobachtungs- und Erwartungswerte gegeben, lässt sich für jedes Tabellenfeld ein Wert berechnen, der diese Werte in Relation setzt. Die Summe dieser Werte über die gesamte Tabelle hinweg wird Kontingenzkoeffizient genannt und mit χ^2 („Chi-Quadrat“) abgekürzt.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2}$$

Bei der Formel steht k wieder für die Anzahl der Zeilen (und i für ihre laufende Nummer) und ℓ für die Anzahl der Spalten (und j für ihre laufende Nummer).

Das doppelte Summenzeichen mag etwas verwirrend sein, bedeutet aber nur, dass die Zeilen spaltenweise summiert werden, und dann die Summe dieser Zeilensumme genommen wird – d.h. dass einfach alle Tabellenfelder aufsummiert werden.

Der χ^2-Wert kann (ähnlich wie der F-Wert aus Sitzung 6) nur positive Werte annehmen. Er bildet die Grundlage für die im Folgenden besprochenen Kennwerte ϕ und CI sowie für den in Sitzung 10 zu besprechenden χ^2-Test.

5.1 Beispiel

Ein möglicher Zwischenschritt ist es, diese Teilwerte von χ^2 für die einzelnen Tabellenfelder auszurechnen und in der Kreuztabelle zu notieren. Die Teilwerte werden dann für jedes Tabellenfeld mit der

Tabelle 5: Kreuztabelle der Beispieldaten mit Teilwerten für χ^2

	Fahrrad	anderes Transportmittel	
Geistes-/Sozialwissenschaft	11	28	39
	(8,67)	(30,33)	
	0,626	0,179	
anderes Studienfach	9	42	51
	(11,33)	(39,67)	
	0,479	0,137	
	20	70	90

Formel

$$\frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{3}$$

berechnet und sind in Tabelle 5 in blau dargestellt.

Zum Beispiel ergibt sich der Teilwert für χ^2 für die Kombination „anderes Studienfach“ – „Fahrrad“ durch Einsetzen in Gleichung 3:

$$\begin{aligned}\frac{(n_{21} - m_{21})^2}{m_{21}} &\approx \frac{(9 - 11{,}33)^2}{11{,}33} \\ &= \frac{-2{,}33^2}{11{,}33} \\ &\approx \frac{5{,}43}{11{,}33} \\ &\approx 0{,}479\end{aligned}$$

Der χ^2-Wert lässt sich nun bestimmen, indem diese Teilwerte aufsummiert werden:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij} - m_{ij})^2}{m_{ij}} \\ &\approx 0{,}626 + 0{,}179 + 0{,}479 + 0{,}137 \\ &= 1{,}421\end{aligned}$$

Mit diesem Wert $\chi^2 \approx 1{,}421$ können wir noch nicht so viel anfangen – wir wissen aber, dass er ein Maß dafür ist, wie sehr unsere beobachtete Verteilung von einer zu erwarteten Verteilung (vorausgesetzt, es gibt keinen Zusammenhang) abweicht.

6 Berechnung des ϕ-Koeffizienten

Der ϕ-Koeffizient ist der Korrelationskoeffizient für zwei dichotome Variablen (wobei er in der hier besprochenen Version nur positive Werte annehmen kann). Er ist jedoch *nicht* ohne weiteres mit dem

Korrelationskoeffizienten r (aus Sitzung 7) vergleichbar.

Der Wert für ϕ kann aus χ^2 berechnet werden mit:

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{4}$$

6.1 Beispiel

In unserem Beispiel ergibt sich also für ϕ durch Einsetzung in Gleichung 4:

$$\begin{aligned}\phi &= \sqrt{\frac{\chi^2}{n}} \\ &\approx \sqrt{\frac{1{,}421}{90}} \\ &\approx 0{,}126\end{aligned}$$

Es wird ersichtlich, dass es eine leichte Korrelation der Variablen gibt. Aber in welche Richtung? Dafür müssen wir auf die Kreuztabelle blicken: Der beobachtete Wert für die Wertekombination „Fahrrad“ und „Geistes-/Sozialwissenschaft“ beträgt $n_{11} = 11$ und liegt über dem Erwartungswert $m_{11} = 8{,}67$. Damit ist klar: Das Studium von Geistes- und Sozialwissenschaften korreliert *positiv* mit der Fahrradnutzung für den Pendelweg.

Ob diese Korrelation auch statistisch relevant ist, kann mit dem χ^2-Test (Sitzung 10) überprüft werden.

7 Berechnung des Cramér-Index

Bisher wurden in dieser Sitzung nur Verteilungen von zwei dichotomen Variablen besprochen. Nun gibt es aber auch nominalskalierte bivariate Verteilungen, in denen die Merkmale mehr als zwei Werte annehmen können (also nicht dichotom sind). In diesem Fall ist der Cramér-Index (auch Cramérs v, engl. *Cramér index*) ein geeigneter Kennwert für die Abhängigkeit der Variablen.

Die Formel für den Cramér-Index lautet

$$CI = \sqrt{\frac{\chi^2}{n \cdot (\min(k, \ell) - 1)}} \tag{5}$$

wobei der Ausdruck $\min(k,\ell)$ für den *kleineren* Wert aus Zeilenanzahl k und Spaltenanzahl ℓ steht.

In einer 2×2-Tabelle ist dieser Wert identisch mit dem ϕ-Koeffizienten.

7.1 Beispiel

Hätten wir im Beispiel die Erhebung nicht auf dichotome Variablen reduziert, sondern die Wissenschaftsdisziplinen und Verkehrsmittel direkt erhoben, so würde sich die Kreuztabelle vielleicht wie in **??** darstellen.

Dabei werden die Erwartungswerte wie gehabt mit Gleichung 1 und die Teilwerte für χ^2 mit der Gleichung 3 errechnet.

	Transportmittel			
Studienfach ↓	Fahrrad	Auto	Öffentliche	
Geisteswissenschaft	5	5	9	19
	(4,22)	(8,02)	(6,76)	
	0,144	1,137	0,742	
Sozialwissenschaft	6	6	8	20
	(4,44)	(8,44)	(7,11)	
	0,548	0,705	0,111	
Naturwissenschaft	5	9	9	23
	(5,11)	(9,71)	(8,18)	
	0,002	0,052	0,082	
Ingenieurswissenschaft	4	18	6	28
	(6,22)	(11,82)	(9,96)	
	0,792	3,231	1,574	
	20	38	32	90

Der χ^2-Wert ergibt sich wieder aus der Summe (s. Gleichung 2):

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\\ &\approx 0{,}144 + 1{,}137 + 0{,}742 + 0{,}548 + 0{,}705 + 0{,}111\\ &\quad + 0{,}002 + 0{,}052 + 0{,}082 + 0{,}792 + 3{,}231 + 1{,}574\\ &= 9{,}120\end{aligned}$$

Mit diesem Wert kann der Cramér-Index anhand von Gleichung 5 berechnet werden.

Die Zeilenanzahl ist $k = 4$ und die Spaltenanzahl $\ell = 3$. Der Ausdruck $\min(k,\ell)$ ergibt den kleineren dieser Werte, also 3:

$$\begin{aligned}CI &= \sqrt{\frac{\chi^2}{n\cdot(\min(k,\ell)-1)}}\\ &\approx \sqrt{\frac{9{,}122}{90\cdot(3-1)}}\\ &\approx 0{,}225\end{aligned}$$

Dieser Wert ist größer als der oben berechnete ϕ-Koeffizient. Das ist nicht besonders überraschend: Eine detailliertere Erfassung der Variablen führt zu einem deutlicheren Zusammenhang.

8 Aufgaben

8.1 Aufgabe 1

Sie fragen sich, wie die Wohnumgebung einer Person (Stadt oder Land) damit zusammenhängt, ob die Person ein eigenes Auto besitzt. Sie erheben die folgende Messreihe:

Wohnort	Autobesitz
Land	Ja
Land	Ja
Stadt	Nein
Stadt	Nein
Stadt	Ja
Stadt	Nein
Land	Ja
Land	Nein
Land	Ja
Land	Ja
Stadt	Nein
Land	Ja
Land	Ja
Land	Ja
Stadt	Nein
Land	Ja
Stadt	Nein
Land	Nein
Stadt	Ja
Stadt	Nein

a) Überführen Sie die Daten in eine Kreuztabelle.

b) Berechnen Sie die Erwartungswerte für jedes Tabellenfeld.

c) Berechnen Sie χ^2.

d) Berechnen Sie den ϕ-Koeffizienten.

e) Besteht eine Korrelation? In welche Richtung?

8.2 Aufgabe 2

Sie interessieren sich dafür, ob zwei „Ja/Nein“-Fragen auf einem Fragebogen korrelieren.

Sie ermittlen folgende Häufigkeiten:

	Frage 2	
Frage 1 ↓	Ja	Nein
Ja	5	28
Nein	40	72

a) Vervollständigen Sie die Kreuztabelle um ihre Summen und die Erwartungswerte.

b) Berechnen Sie χ^2 und den ϕ-Koeffizienten.

c) Wie würden Sie den Zusammenhang beschreiben?

8.3 Aufgabe 3

Sie möchten überprüfen, ob auf dem Arbeitsmarkt anhand von Namen diskriminiert wird, die auf einen Migrationshintergrund schließen lassen. Sie antworten als fiktive Bewerber*innen mit vergleichbaren Qualifikationen auf zufällige Stellenanzeigen und halten fest, ob die jeweilige Bewerbung in einer Einladung zum Vorstellungsgespräch resultiert.

Sie erheben diese Daten:

	Ergebnis	
Herkunft des Namens ↓	eingeladen	nicht eingeladen
deutsch	36	64
italienisch	23	77
slawisch	9	91
türkisch	11	89

Können Sie einen Zusammenhang zwischen Namensherkunft und Erfolg der Bewerbung feststellen? Begründen Sie Ihre Antwort.

9 Tipps zur Vertiefung

- Kapitel 9.1, 10.3.4 und 10.3.7 in Bortz und Schuster (2010)
- Kapitel 6.7.2 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 2.3 in Klemm (2002)

Quellen

Bahrenberg, Gerhard, Ernst Giese und Josef Nipper. 2010. *Statistische Methoden in der Geographie*. 5. Aufl. Bd. 1. Univariate und bivariate Statistik. Stuttgart: Bornträger.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.

Klemm, Elmar. 2002. *Einführung in die Statistik. Für die Sozialwissenschaften*. Wiesbaden: Westdeutscher Verlag.