Karlsruher Institut für Technologie

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Ökonometrie und Statistik

Kapitel IX - Kontingenzkoeffizient

Deskriptive Statistik

Prof. Dr. W.-D. Heller
Hartwig Senska
Carlo Siebenschuh

Agenda

1. **Untersuchung der Abhängigkeit**
2. Chi-Quadrat
3. Kontingenzkoeffizient nach Pearson

Untersuchung der Abhängigkeit

Bei der Untersuchung des Zusammenhangs zweier nominalskalierter Merkmale ist es nicht möglich, formale Aussagen über die Art des Zusammenhangs zu machen. Man kann zunächst nur feststellen, dass eine Abhängigkeit besteht. Worin diese besteht, lässt sich jedoch nicht direkt feststellen. Man ist auf die Untersuchung weiterer Merkmale angewiesen. Ohne Kenntnis weiterer Merkmale ist die einzige Möglichkeit, die besteht, zu untersuchen, wie ausgeprägt die Abhängigkeit ist, etwa indem man feststellt, wie weit die Verteilung von der Unabhängigkeit abweicht.

Ziel:
Messung der Abweichung von Unabhängigkeit

Untersuchung der Abhängigkeit

Bei der Untersuchung des Zusammenhangs zweier nominalskalierter Merkmale ist es nicht möglich, formale Aussagen über die Art des Zusammenhangs zu machen. Man kann zunächst nur feststellen, dass eine Abhängigkeit besteht. Worin diese besteht, lässt sich jedoch nicht direkt feststellen. Man ist auf die Untersuchung weiterer Merkmale angewiesen. Ohne Kenntnis weiterer Merkmale ist die einzige Möglichkeit, die besteht, zu untersuchen, wie ausgeprägt die Abhängigkeit ist, etwa indem man feststellt, wie weit die Verteilung von der Unabhängigkeit abweicht.

Ziel:
Messung der Abweichung von Unabhängigkeit

Untersuchung der Abhängigkeit

Für unabhängige Merkmale ist die gemeinsame Häufigkeitsverteilung mit den Randhäufigkeiten festgelegt durch die Formel

$$p(a, b) = p(a) \cdot p(b).$$

Für die absoluten Häufigkeiten gilt also in diesem Fall

$$\begin{aligned} h(a, b) &= n \cdot p(a, b) = n \cdot p(a) \cdot p(b) \\ &= n \cdot \frac{h(a)}{n} \cdot \frac{h(b)}{n} = \frac{h(a) \cdot h(b)}{n}. \end{aligned}$$

Untersuchung der Abhängigkeit

Für unabhängige Merkmale ist die gemeinsame Häufigkeitsverteilung mit den Randhäufigkeiten festgelegt durch die Formel

$$p(a, b) = p(a) \cdot p(b).$$

Für die absoluten Häufigkeiten gilt also in diesem Fall

$$\begin{aligned} h(a, b) &= n \cdot p(a, b) = n \cdot p(a) \cdot p(b) \\ &= n \cdot \frac{h(a)}{n} \cdot \frac{h(b)}{n} = \frac{h(a) \cdot h(b)}{n}. \end{aligned}$$

Untersuchung der Abhängigkeit

Ausgehend von den beiden Randverteilungen der gemeinsamen Häufigkeitsverteilung zweier Merkmale kann also ermittelt werden, wie die Häufigkeitsverteilung aussehen müßte, falls die Merkmale unabhängig wären. Die Tabelle dieser fiktiven Werte nennt man auch **Indifferenztafel** oder **Indifferenztabelle**.

Der Unterschied der beiden Tabellen dokumentiert die Abweichung der tatsächlichen Häufigkeitsverteilung von der bei Unabhängigkeit. Es bietet sich damit an, in jedem Feld der Tabelle die tatsächliche absolute Häufigkeit mit dem theoretisch ermittelten Wert bei Unabhängigkeit zu vergleichen.
Die Abweichung ist im Feld (a, b)

$$d(a,b) = h(a,b) - \frac{h(a) \cdot h(b)}{n}.$$

Damit erhält man ein Maß für die Abhängigkeit, wenn man die Differenzen $d(a, b)$ zu einer Zahl zusammenfasst.

Untersuchung der Abhängigkeit

Ausgehend von den beiden Randverteilungen der gemeinsamen Häufigkeitsverteilung zweier Merkmale kann also ermittelt werden, wie die Häufigkeitsverteilung aussehen müßte, falls die Merkmale unabhängig wären. Die Tabelle dieser fiktiven Werte nennt man auch **Indifferenztafel** oder **Indifferenztabelle**.

Der Unterschied der beiden Tabellen dokumentiert die Abweichung der tatsächlichen Häufigkeitsverteilung von der bei Unabhängigkeit. Es bietet sich damit an, in jedem Feld der Tabelle die tatsächliche absolute Häufigkeit mit dem theoretisch ermittelten Wert bei Unabhängigkeit zu vergleichen.
Die Abweichung ist im Feld (a, b)

$$d(a,b) = h(a,b) - \frac{h(a) \cdot h(b)}{n}.$$

Damit erhält man ein Maß für die Abhängigkeit, wenn man die Differenzen $d(a, b)$ zu einer Zahl zusammenfasst.

Untersuchung der Abhängigkeit

Ausgehend von den beiden Randverteilungen der gemeinsamen Häufigkeitsverteilung zweier Merkmale kann also ermittelt werden, wie die Häufigkeitsverteilung aussehen müßte, falls die Merkmale unabhängig wären. Die Tabelle dieser fiktiven Werte nennt man auch **Indifferenztafel** oder **Indifferenztabelle**.

Der Unterschied der beiden Tabellen dokumentiert die Abweichung der tatsächlichen Häufigkeitsverteilung von der bei Unabhängigkeit. Es bietet sich damit an, in jedem Feld der Tabelle die tatsächliche absolute Häufigkeit mit dem theoretisch ermittelten Wert bei Unabhängigkeit zu vergleichen.
Die Abweichung ist im Feld (a, b)

$$d(a,b) = h(a,b) - \frac{h(a) \cdot h(b)}{n}.$$

Damit erhält man ein Maß für die Abhängigkeit, wenn man die Differenzen $d(a, b)$ zu einer Zahl zusammenfasst.

Agenda

1. Untersuchung der Abhängigkeit
2. **Chi-Quadrat**
3. Kontingenzkoeffizient nach Pearson

Chi-Quadrat

Maßzahl für die Abweichung von Unabhängigkeit, die durch Summation der relativen quadrierten Abweichungen der beobachteten Merkmalsausprägungen von den Werten bei Unabhängigkeit entsteht:

$$\chi^2 = \sum_{\substack{a \in M_1 \\ h(a) \neq 0}} \sum_{\substack{b \in M_2 \\ h(b) \neq 0}} \frac{\left(h(a,b) - \frac{h(a) \cdot h(b)}{n} \right)^2}{\frac{h(a) \cdot h(b)}{n}}$$

Offensichtlich gilt:

- $\chi^2 \geq 0$.
- $\chi^2 = 0$ genau dann, wenn die Merkmale unabhängig sind.

Chi-Quadrat

Maßzahl für die Abweichung von Unabhängigkeit, die durch Summation der relativen quadrierten Abweichungen der beobachteten Merkmalsausprägungen von den Werten bei Unabhängigkeit entsteht:

$$\chi^2 = \sum_{\substack{a \in M_1 \\ h(a) \neq 0}} \sum_{\substack{b \in M_2 \\ h(b) \neq 0}} \frac{\left(h(a,b) - \frac{h(a) \cdot h(b)}{n} \right)^2}{\frac{h(a) \cdot h(b)}{n}}$$

Offensichtlich gilt:

- $\chi^2 \geq 0$.
- $\chi^2 = 0$ genau dann, wenn die Merkmale unabhängig sind.

Chi-Quadrat

Interpretation von χ^2

Je größer χ^2 ist, desto größer sind die relativen Abweichungen in den einzelnen Feldern, desto größer der Unterschied zwischen Häufigkeitstabelle und Indifferenztafel, desto größer also die quadrierten Abweichungen von Unabhängigkeit.

Wegen

$$\chi^2 = n \cdot \sum_{\substack{a \in M_1 \\ p(a)>0}} \sum_{\substack{b \in M_2 \\ p(b)>0}} \frac{\Big(p(a,b) - p(a) \cdot p(b)\Big)^2}{p(a) \cdot p(b)}$$

verdoppelt sich bei Verdoppelung der absoluten Häufigkeiten (also bei Verdoppelung von n bei konstanten $p(\cdot)$) auch die Zahl χ^2.

Chi-Quadrat

Interpretation von χ^2
Je größer χ^2 ist, desto größer sind die relativen Abweichungen in den einzelnen Feldern, desto größer der Unterschied zwischen Häufigkeitstabelle und Indifferenztafel, desto größer also die quadrierten Abweichungen von Unabhängigkeit.

Wegen

$$\chi^2 = n \cdot \sum_{\substack{a \in M_1 \\ p(a)>0}} \sum_{\substack{b \in M_2 \\ p(b)>0}} \frac{\Big(p(a,b) - p(a) \cdot p(b)\Big)^2}{p(a) \cdot p(b)}$$

verdoppelt sich bei Verdoppelung der absoluten Häufigkeiten (also bei Verdoppelung von n bei konstanten $p(\cdot)$) auch die Zahl χ^2.

Agenda

1. Untersuchung der Abhängigkeit
2. Chi-Quadrat
3. **Kontingenzkoeffizient nach Pearson**

Kontingenzkoeffizient nach Pearson

Die Zahl χ^2 hat nicht die Eigenschaft, als Maximalwert den Wert 1 zu haben, vielmehr kann χ^2 auch Werte größer als 1 annehmen, wobei der Maximalwert mit n ansteigt. Die häufigste Methode, dies zu korrigieren, ist der Kontingenzkoeffizient nach Pearson.

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{\frac{\chi^2}{n}}{1 + \frac{\chi^2}{n}}}$$

Es gilt:

$0 \leq C < 1 \quad$ und $\quad C = 0 \Leftrightarrow \chi^2 = 0 \Leftrightarrow$ Merkmale unabhängig

Je größer C, desto stärker die Abweichung von Unabhängigkeit. Der Maximalwert von C ergibt sich bei "absoluter Abhängigkeit".

Kontingenzkoeffizient nach Pearson

Die Zahl χ^2 hat nicht die Eigenschaft, als Maximalwert den Wert 1 zu haben, vielmehr kann χ^2 auch Werte größer als 1 annehmen, wobei der Maximalwert mit n ansteigt. Die häufigste Methode, dies zu korrigieren, ist der Kontingenzkoeffizient nach Pearson.

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{\frac{\chi^2}{n}}{1 + \frac{\chi^2}{n}}}$$

Es gilt:

$0 \leq C < 1 \quad$ und $\quad C = 0 \Leftrightarrow \chi^2 = 0 \Leftrightarrow$ Merkmale unabhängig

Je größer C, desto stärker die Abweichung von Unabhängigkeit. Der Maximalwert von C ergibt sich bei "absoluter Abhängigkeit".

Kontingenzkoeffizient nach Pearson

Die Zahl χ^2 hat nicht die Eigenschaft, als Maximalwert den Wert 1 zu haben, vielmehr kann χ^2 auch Werte größer als 1 annehmen, wobei der Maximalwert mit n ansteigt. Die häufigste Methode, dies zu korrigieren, ist der Kontingenzkoeffizient nach Pearson.

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{\frac{\chi^2}{n}}{1 + \frac{\chi^2}{n}}}$$

Es gilt:

$0 \leq C < 1 \quad$ und $\quad C = 0 \Leftrightarrow \chi^2 = 0 \Leftrightarrow$ Merkmale unabhängig

Je größer C, desto stärker die Abweichung von Unabhängigkeit. Der Maximalwert von C ergibt sich bei "absoluter Abhängigkeit".

Kontingenzkoeffizient nach Pearson

Absolute Abhängigkeit
liegt vor, wenn jede Merkmalsausprägung a von Merkmal 1 nur in Kombination mit einer ganz bestimmten Merkmalsausprägung b von Merkmal 2 beobachtet wurde und umgekehrt. In diesem Fall ist zu vermuten, dass C maximal wird. (Es gilt dann $\chi^2_{max} = n(k-1)$)

Für den Kontingenzkoeffizienten nach Pearson gilt dann:

$$0 \leq C \leq \sqrt{\frac{k-1}{k}}.$$

Kontingenzkoeffizient nach Pearson

Absolute Abhängigkeit
liegt vor, wenn jede Merkmalsausprägung a von Merkmal 1 nur in Kombination mit einer ganz bestimmten Merkmalsausprägung b von Merkmal 2 beobachtet wurde und umgekehrt. In diesem Fall ist zu vermuten, dass C maximal wird. (Es gilt dann $\chi^2_{max} = n(k-1)$)

Für den Kontingenzkoeffizienten nach Pearson gilt dann:

$$0 \leq C \leq \sqrt{\frac{k-1}{k}}.$$

Kontingenzkoeffizient nach Pearson

Mit $k = \min\{r, s\}$, wobei r, s die Anzahlen der beobachteten Merkmalsausprägungen von Merkmal 1 bzw. 2 sind.

Korrigierter Kontingenzkoeffizient nach Pearson

$$C_{\text{corr}} = \sqrt{\frac{k}{k-1}} \cdot C,$$

Es gilt:

$$0 \leq C_{\text{corr}} \leq 1.$$

Kontingenzkoeffizient nach Pearson

Mit $k = \min\{r, s\}$, wobei r, s die Anzahlen der beobachteten Merkmalsausprägungen von Merkmal 1 bzw. 2 sind.

Korrigierter Kontingenzkoeffizient nach Pearson

$$C_{\text{corr}} = \sqrt{\frac{k}{k-1}} \cdot C,$$

Es gilt:

$$0 \leq C_{\text{corr}} \leq 1.$$

Kontingenzkoeffizient nach Pearson

Beispiel 9.1

Bei einer Untersuchung von 100 statistischen Einheiten hat sich die folgende zweidimensionale relative Häufigkeitsverteilung ergeben:

	b_1	b_2	b_3	$p(a_i)$
a_1	0.02	0.25	0.03	0.30
a_2	0.03	0.04	0.33	0.40
a_3	0.15	0.11	0.04	0.30
$p(b_i)$	0.20	0.40	0.40	1

Kontingenzkoeffizient nach Pearson

Beispiel 9.1
Die Berechnung von χ^2 erfolgt übersichtlich mit folgendem Arbeitsschema für das Feld (a, b):

$p(a,b)$	$(p(a,b) - p(a)p(b))^2$
$p(a,b) - p(a)p(b)$	$p(a)p(b)$

Das Vorzeichen für die Differenz links unten kann vernachlässigt werden, da die Zahl ohnehin quadriert wird.

Kontingenzkoeffizient nach Pearson

Beispiel 9.1

	b_1		b_2		b_3		$p(a_1)$
a_1	0.02	0.0016	0.25	0.0169	0.03	0.0081	0.3
	0.04	0.06	0.13	0.12	0.09	0.12	
a_2	0.03	0.0025	0.04	0.0144	0.33	0.0289	0.4
	0.05	0.08	0.12	0.16	0.17	0.16	
a_3	0.15	0.0081	0.11	0.0001	0.04	0.0064	0.3
	0.09	0.06	0.01	0.12	0.08	0.12	
	0.2		0.4		0.4		

Kontingenzkoeffizient nach Pearson

Beispiel 9.1

$$\frac{1}{n} \cdot \chi^2 = 0.02\bar{6} + 0.1408 + 0.0675 + 0.03125 + 0.09 + 0.180625 + 0.135 + 0.0008\bar{3} + 0.05\bar{3} = 0.726$$

und daraus mit $n = 100$:

$$\chi^2 = 72.6; \quad C = \sqrt{\frac{72.6}{172.6}} = 0.65; \quad C_{\text{corr}} = \sqrt{\frac{3}{2}} \cdot 0.65 = 0.79$$

Kontingenzkoeffizient nach Pearson

Beispiel 9.1
Man sieht, dass Merkmalsausprägung b_2 am häufigsten zusammen mit a_1, b_3 am häufigsten zusammen mit a_2 und b_1 am häufigsten zusammen mit a_3 auftritt. Betrachtet man dazu die bedingten relativen Häufigkeiten, so wird dies besonders deutlich:

$$p(b_1|a_1) = 0.07; \quad p(b_2|a_1) = 0.83; \quad p(b_3|a_1) = 0.1;$$

$$p(b_1|a_2) = 0.075; \quad p(b_2|a_2) = 0.1; \quad p(b_3|a_2) = 0.825;$$

$$p(b_1|a_3) = 0.5; \quad p(b_2|a_3) = 0.37; \quad p(b_3|a_3) = 0.13.$$

Bsp. zu χ^2 und Kontingenzkoeffizient C

Zwei Merkmale A und B mit aufgetretenen Ausprägungen $a_1, \ldots, a_3$ und $b_1, \ldots, b_3$. Somit $r = 3$ und $s = 3$. Die Merkmale können aber Ausprägungen haben, die nicht beobachtet wurden.

Fall 1: sei $n = k = \min\{r, s\} = r = s = 3$

	b_1	b_2	b_3	$h(b)$
a_1	1			1
a_2			1	1
a_3		1		1
h(a)	1	1	1	$\sum = 3$

Bsp. zu χ^2 und Kontingenzkoeffizient C

Ansatz: $(n = 3)$

$$\chi^2 = 3 \left[\frac{(1-\frac{1}{3})^2}{1 \cdot 1} + \frac{(0-\frac{1}{3})^2}{1 \cdot 1} + \frac{(0-\frac{1}{3})^2}{1 \cdot 1} \right.$$
$$\frac{(0-\frac{1}{3})^2}{1 \cdot 1} + \frac{(0-\frac{1}{3})^2}{1 \cdot 1} + \frac{(1-\frac{1}{3})^2}{1 \cdot 1}$$
$$\left. \frac{(0-\frac{1}{3})^2}{1 \cdot 1} + \frac{(1-\frac{1}{3})^2}{1 \cdot 1} + \frac{(0-\frac{1}{3})^2}{1 \cdot 1} \right] = \dots$$

(Regelmäßigkeiten finden!)

Bsp. zu χ^2 und Kontingenzkoeffizient C

Zwei Merkmale A und B mit aufgetretenen Ausprägungen $a_1, \ldots, a_3$ und $b_1, \ldots, b_3$. Somit $r = 3$ und $s = 3$. Die Merkmale können aber Ausprägungen haben, die nicht beobachtet wurden.

Fall 2: sei $6 = n > k = \min\{r, s\} = r = s = 3$

	b_1	b_2	b_3	$h(b)$
a_1	2			2
a_2			3	3
a_3		1		1
$h(a)$	2	1	3	$\sum = 6$

Bsp. zu χ^2 und Kontingenzkoeffizient C

Ansatz ($n = 6$)**:**

$$\chi^2 = 6 \cdot \left[\frac{(2 - \frac{2 \cdot 2}{6})^2}{2 \cdot 2} + \frac{(0 - \frac{2 \cdot 1}{6})^2}{2 \cdot 1} + \frac{(0 - \frac{2 \cdot 3}{6})^2}{2 \cdot 3} \right.$$
$$\frac{(0 - \frac{3 \cdot 2}{6})^2}{3 \cdot 2} + \frac{(0 - \frac{3 \cdot 1}{6})^2}{3 \cdot 1} + \frac{(3 - \frac{3 \cdot 3}{6})^2}{3 \cdot 3}$$
$$\left. \frac{(0 - \frac{1 \cdot 2}{6})^2}{1 \cdot 2} + \frac{(1 - \frac{1 \cdot 1}{6})^2}{1 \cdot 1} + \frac{(0 - \frac{1 \cdot 3}{6})^2}{1 \cdot 3} \right] = \ldots$$

(Regelmäßigkeiten finden!)