Themen: χ^2-basierte Zusammenhangsmaße, PRE-Maß Lambda

Prof. Dr. Elmar Schlüter
Justus-Liebig-Universität Giessen
Fachbereich Sozial- und Kulturwissenschaften
Institut für Soziologie
Wintersemester 2018/19

Übersicht

- Kurze Rückschau & Wiederholung

  χ^2

- Kontingenzkoeffizient C
  ϕ (Phi)
  Cramér's V

- Kovarianz & Korrelation

Kurze Rückschau & Wiederholung

Berechnung von Zusammenhängen für nominale Daten

- Zusammenhänge bei nominalen Daten
  - Ausprägungen treten häufiger bzw. seltener gemeinsam auf als bei zufälliger Verteilung zu erwarten wäre

- Ausgangsbeispiel:

	Erhebungsgebiet		
Partei	West	Ost	Gesamt
PDS	4	116	*120*
Nicht-PDS	1572	606	*2178*
Gesamt	*1576*	*722*	*2298*

  - Randsummen & Gesamtsumme

Berechnung von Zusammenhängen für nominale Daten

- Deutlich unterschiedliche Wahlpräferenzen für Ost- & Westdeutsche

  → Es besteht ein Zusammenhang

Chi-Quadrat

	Erhebungsgebiet		
Partei	West	Ost	Gesamt
PDS	4	116	*120*
Nicht-PDS	1572	606	*2178*
Gesamt	*1576*	*722*	*2298*

	Erhebungsgebiet		
Partei	West	Ost	Gesamt
PDS	82.3	37.7	*120*
Nicht-PDS	1493.7	684.3	*2178*
Gesamt	*1576*	*722*	*2298*

$$\chi^2 = \frac{(4-82.3)^2}{82.3} \ldots + \frac{(606-684.3)^2}{684.3}$$
$$= 250.16$$

Chi-Quadrat

- Maßzahl um Aussage über den Zusammenhang zwischen zwei Merkmalen zu treffen
- Quadrierte Residuen aller Zellen werden aufsummiert und an den erwarteten Häufigkeiten relativiert
- Kann Werte von 0 bis $+\infty$ annehmen
- 0 = kein Zusammenhang
- Aber: Abhängig von der Fallzahl (mehr dazu bei der nächsten Sitzung)

χ^2-basierte Zusammenhangsmaße
ϕ (Phi) - Kontingenzkoeffizient C - Cramér's V

χ^2-basierte Zusammenhangsmaße

- Problem#1:
  χ^2 ist abhängig von den absoluten Häufigkeiten in den Zellen

- z.B. Verdopplung der Häufigkeiten = Verdopplung χ^2

- Prozentuale Verteilung ändert sich jedoch nicht

- Alternative: Normierung des χ^2-Wertes

Chi-Quadrat

	Erhebungsgebiet		
Partei	West	Ost	Gesamt
PDS	4	116	*120*
Nicht-PDS	1572	606	*2178*
Gesamt	*1576*	*722*	*2298*

	Erhebungsgebiet		
Partei	West	Ost	Gesamt
PDS	82.3	37.7	*120*
Nicht-PDS	1493.7	684.3	*2178*
Gesamt	*1576*	*722*	*2298*

$$\chi^2 = \frac{(4-82.3)^2}{82.3} \ldots + \frac{(606-684.3)^2}{684.3}$$
$$= 250.16$$

χ^2-basierte Zusammenhangsmaße

- Für 2×2-Kreuztabellen: ϕ (Phi)

- Ziel ist Relativierung des χ^2-Wertes für die Anzahl der Beobachtungen
- Formal $$\phi = \sqrt{\frac{\chi^2}{n}}$$
- ϕ variiert zwischen 0 (min.) und 1 (max.)
- Beispiel: $$\phi = \sqrt{\frac{250.2}{2298}} = 0.33$$

χ^2-basierte Zusammenhangsmaße

- Problem#2:
  χ^2 ist abhängig von der Kategorienanzahl pro Tabelle

- Kontingenzkoeffizient C

- Formal:

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

- Variiert zwischen 0 (min.) und C_{max}

$$C_{max} = \sqrt{\frac{R-1}{R}}$$

χ^2-basierte Zusammenhangsmaße

- Kontingenzkoeffizient C

$$C_{max} = \sqrt{\frac{R-1}{R}}$$

mit R = min (I, m)

- R = Minimum der Zeilen- bzw. Spaltenzahl
- Beispiele (R):
  2×2: R = 2
  3×4: R = 3
  4×3: R = 3

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{250{,}2}{250{,}2 + 2298}} = 0{,}31$$

χ^2-basierte Zusammenhangsmaße

- Problem#3:
  Kontingenzkoeffizienten aus Tabellen unterschiedlicher Größe sind schwierig zu vergleichen

- Cramér's V

  - Formal: Cramér's V = $\sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \cdot (R-1)}}$

  - χ^2-Wert wird durch maximal erreichbaren χ^2-Wert dividiert
  - Für Mehrfeldertabellen: $\chi^2_{max} = n\,(R\text{-}1)$
    R = minimale Spalten- bzw. Zeilenanzahl
  - Für unser Beispiel: $\sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{250{,}2}{2298 \cdot (2-1)}} = 0{,}33$

PRE-Maß λ (Lambda)

PRE-Maß λ (Lambda)

- Ausgangspunkt: „Wie gut können die Werte einer abhängigen Variablen durch die Werte einer unabhängigen Variablen vorhergesagt werden?“
  für nominale Daten
- Schritt 1: Prognose des Wertes der abhängigen Variablen ohne Kenntnis der unabhängigen Variablen
  Modus als bestmögliche Vorhersage
- Schritt 2: Prognose des Wertes der abhängigen Variablen mit Kenntnis der unabhängigen Variablen
- Schritt 3: Ermittlung des PRE-Maßes (inwieweit wurde die Vorhersage durch Einbezug der unabhängigen Variable verbessert?)
- Formal:

$$\lambda = \frac{(\text{Fehler}_1 - \text{Fehler}_2)}{\text{Fehler}_1}$$

PRE-Maß λ (Lambda)

- Beispiel (nach Gehring/Weins 2009, S. 154):

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Gesamt
CDU/CSU	335	15	*350*
SPD	25	320	*345*
Andere	84	102	*186*
Gesamt	*444*	*437*	*881*

- Modus Wahlabsicht: CDU/CSU
- Also: Ohne Kenntnis der Kanzlerpräferenz ist die bestmögliche Vorhersage der Wahlabsicht CDU/CSU

PRE-Maß λ (Lambda)

- Beispiel (nach Gehring/Weins 2009, S. 154):

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Gesamt
CDU/CSU	335	15	*350*
SPD	25	320	*345*
Andere	84	102	*186*
Gesamt	*444*	*437*	*881*

- Modus Wahlabsicht: CDU/CSU
- Also: Ohne Kenntnis der Kanzlerpräferenz ist die bestmögliche Vorhersage der Wahlabsicht CDU/CSU
- Vorhersagefehler 1: 345+186 = 531

PRE-Maß λ (Lambda)

- Beispiel (nach Gehring/Weins 2009, S. 154):

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Gesamt
CDU/CSU	335	15	*350*
SPD	25	320	*345*
Andere	84	102	*186*
Gesamt	*444*	*437*	*881*

- Bei Kenntnis der Kanzlerpräferenz:
- CDU/CSU für Merkel-Anhänger
- Vorhersagefehler: 25+84 =109

PRE-Maß λ (Lambda)

- Beispiel (nach Gehring/Weins 2009, S. 154):

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Gesamt
CDU/CSU	335	15	*350*
SPD	25	320	*345*
Andere	84	102	*186*
Gesamt	*444*	*437*	*881*

- Bei Kenntnis der Kanzlerpräferenz:
- SPD für Steinmeier-Anhänger
- Vorhersagefehler : 15+102 =117
- Zusammen: 109 + 117 = 226 Vorhersagefehler 2

PRE-Maß λ (Lambda)

- Beispiel (nach Gehring/Weins 2009, S. 154):

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Gesamt
CDU/CSU	335	15	*350*
SPD	25	320	*345*
Andere	84	102	*186*
Gesamt	*444*	*437*	*881*

- Formal: $$\lambda = \frac{(\text{Fehler}_1 - \text{Fehler}_2)}{\text{Fehler}_1} = \frac{(531 - 226)}{531} = 0.57$$

- Die Kenntnis der Kanzlerpräferenz verringert die Fehler bei der Prognose der Wahlabsicht um 57%

PRE-Maß λ (Lambda)

- Mini-Übung: Bitte berechnen sie Lambda!

Frisur	Geschlecht		
	Frau	Mann	Gesamt
Lang	60	30	*90*
Kurz	40	70	*110*
Gesamt	*100*	*100*	*200*

Unabhängige Variable:

Formal:

Bestmögliche Vorhersage ohne UV:

1. Vorhersagefehler:

2. Vorhersagefehler:

Lambda: = 22%

PRE-Maß λ (Lambda)

- Mini-Übung: Bitte berechnen sie Lambda!

	Geschlecht		
Frisur	Frau	Mann	Gesamt
Lang	60	30	*90*
Kurz	40	70	*110*
Gesamt	*100*	*100*	*200*

Unabhängige Variable: Geschlecht

Formal: (Fehler 1- Fehler 2)/Fehler 1

Bestmögliche Vorhersage ohne UV: 110

1. Vorhersagefehler: 90

2. Vorhersagefehler: 70

Lambda: (90 – 70)/90 = 20/90 = 0.22 = 22%

Chi-Quadrat

- Formal