JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

8. Vorlesung am 14.12.23 (Fortsetzung bivariate Zusammenhänge)

Plan

- **Abschluss Kreuztabelle, Teil 1 (Achtung siehe Foliensatz vom letzten Mal, letzte 11 Folien)**
- **Kreuztabelle, Teil 2: Erstellung einer Indifferenztabelle**
- **Zusammenhangsmaße für nominalskalierte Variablen: Chi-Quadrat, Phi und Cramer‘s V**

Plan

- Abschluss Kreuztabelle, Teil 1 (Foliensatz vom letzten Mal)
- **Kreuztabelle, Teil 2: Erstellung einer Indifferenztabelle**
- **Zusammenhangsmaße für nominalskalierte Variablen: Chi-Quadrat und Cramer‘s V**

Lernziele

- **Sie wissen was eine Indifferenztabelle ist und können eine erstellen**
- **Sie kennen Zusammenhangsmaße für nominalskalierte Variablen und können diese berechnen**

Kreuztabelle Politisches Interesse und Geschlecht – Wiederholung beobachtete Häufigkeiten

Politisches Interesse \ Geschlecht	Männliche Befragte	Weibliche Befragte	Gesamt
Sehr stark	311 17,6%	116 6,7%	427 12,2%
Stark	537 30,3%	345 20,1%	882 25,3%
Mittel	634 35,8%	795 46,2%	1429 40,9%
Wenig	207 11,7%	349 20,3%	556 15,9%
Überhaupt nicht	81 4,6%	115 6,7%	196 5,6%
Gesamt	1770 100,0%	1720 100,0%	3490 100,0%

Daten: ALLBUS 2016. Eigene Berechnungen

Indifferenztabelle, Hinführung

- Weiteres „feature“ für Rückschlüsse auf mögliche Zusammenhänge zwischen den untersuchten Merkmalen
- Bildet die sog „**erwarteten Häufigkeiten**“ ab
- **Definition erwartete Häufigkeiten**: Kombinierte Verteilung zweier Variablen, die erwartet wird, wenn es **statistische Unabhängigkeit** gibt
- Beobachtete vs. Erwartete Häufigkeiten

Hinführung Indifferenztabelle

- **Unterschiede** im politischen Interesse zwischen Männern und Frauen auf Basis des **Vergleichs der Spaltenprozente →** Vermutung: Es besteht ein Zusammenhang

- Wie würde die Verteilung bei statistischer Unabhängigkeit aussehen? → Grundlage erwarteter Häufigkeiten

- Berechnung **erwartete Häufigkeiten** durch Einbezug der Randhäufigkeiten:

$$f_{e(ij)} = \frac{Zeilensumme \times Spaltensumme}{n}$$

Indifferenztabelle, Beispiel und Übung

Politisches Interesse und Geschlecht – Ergänzen Sie die Indifferenztabelle und vergleichen Sie beobachtete und erwartete Häufigkeiten

Politisches Interesse \ Geschlecht	Männliche Befragte	Weibliche Befragte	Gesamt
Sehr stark	1770 * 427 / 3490 = 216,56		427
Stark			882
Mittel		1720 * 1429 / 3490 = 704,26	1429
Wenig			556
Überhaupt nicht			196
Gesamt	1770	1720	3490

Daten: ALLBUS 2016 (n = 3490). Eigene Berechnungen

Politisches Interesse und Geschlecht – Ergänzen Sie die Indifferenztabelle und vergleichen Sie beobachtete und erwartete Häufigkeiten

Tabelle 28: Politisches Interesse und Geschlecht (erwartete Häufigkeiten) – Indifferenztabelle

Politisches Interesse / Geschlecht	Männliche Befragte	Weibliche Befragte	Gesamt
Sehr stark	216,56	210,44	427
Stark	447,32	434,68	882
Mittel	724,74	704,26	1429
Wenig	281,98	274,02	556
Überhaupt nicht	99,40	96,60	196
Gesamt	1770	1720	3490

Politisches Interesse / Geschlecht	Männliche Befragte	Weibliche Befragte	Gesamt
Sehr stark	311 17,6%	116 6,7%	427 12,2%
Stark	537 30,3%	345 20,1%	882 25,3%
Mittel	634 35,8%	795 46,2%	1429 40,9%
Wenig	207 11,7%	349 20,3%	556 15,9%
Überhaupt nicht	81 4,6%	115 6,7%	196 5,6%
Gesamt	1770 100,0%	1720 100,0%	3490 100,0%

Indifferenztabelle

- Je stärker sich **erwartete und beobachtete Häufigkeiten unterscheiden**, desto stärker ist der Zusammenhang zwischen den beiden Merkmalen
- **Residuum**: Abweichung bzw. Differenz der beobachteten und erwarteten Werte

Beispiel:

- Beobachtet wurden 311 Männer, die ein starkes politisches Interesse bekunden
- Bei statistischer Unabhängigkeit wären 217 Männer zu erwarten gewesen
- Es haben also 94 mehr Männer ein starkes politisches Interesse bekundet, als zu erwarten gewesen wäre

→ Berechnung eines **nominalen Zusammenhangsmaßes**: Für alle Zellen stellt die Indifferenztabelle die Basis für die Berechnung von Chi-Quadrat ($\chi 2$) bereit

Plan

- Abschluss Kreuztabelle, Teil 1 (Foliensatz vom letzten Mal)
- Kreuztabelle, Teil 2: Erstellung einer Indifferenztabelle
- **Zusammenhangsmaße für nominalskalierte Variablen: Chi-Quadrat, Phi und Cramer's V**
- **(Falls noch Zeit: Einstieg in bivariate Zusammenhänge von metrisch skalierten Variablen)**

Chi-Quadrat ($\chi 2$)

- **Zusammenhangsmaß** für nominale Merkmale (2 Merkmale)
- Nutzt alle Residuen einer Indifferenztabelle, um eine „globale" Aussage über den Zusammenhang zwischen zwei Merkmalen (über die gesamte Kreuztabelle hinweg)
- Formal:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(f_{b_{ij}} - f_{e_{ij}})^2}{f_{e_{ij}}}$$

$f_{b(ij)}$ = beobachtete Häufigkeit in der i-ten Zeile und j-ten Spalte

$f_{e(ij)}$ = erwartete Häufigkeit in der i-ten Zeile und j-ten Spalte

k = Anzahl der Zeilen

m = Anzahl der Spalten

- Residuen aller Zellen werden aufsummiert und quadriert (dadurch erhalten alle Werte ein positives Vorzeichen)

Arbeitstabelle zur Berechnung Chi-Quadrat ($\chi 2$)

f_b	f_e	$f_b - f_e$	$(f_b - f_e)^2$	$\frac{(fb - fe)^2}{fe}$
311	216,56	94,44	8918,91	41,18
537	447,32	89,68	8042,50	17,98
634	724,74	–90,74	8233,75	11,36
207	281,98	–74,98	5622,00	19,94
81	99,40	–18,40	338,56	3,41
116	210,44	–94,44	8918,91	42,38
345	434,68	–89,68	8042,50	18,50
795	704,26	90,74	8233,75	11,69
349	274,02	74,98	5622,00	20,52
115	96,60	18,40	338,56	3,50
			Chi-Quadrat	190,46

Quelle: Eigene Darstellung

Chi-Quadrat ist von n und der Anzahl der Zellen abhängig

Chi-Quadrat ($\chi 2$)

- Quadrierte Residuen aller Zellen werden aufsummiert und an den erwarteten Häufigkeiten relativiert
- Kann **Werte von 0 bis +∞ annehmen** (nicht-standardisiertes Maß)
- 0 = kein Zusammenhang
- Je größer der Wert, desto größer der Zusammenhang
- **Aber**: Abhängig von n und der „Größe“ der Kreuztabelle

Übung Chi-Quadrat, fiktives Beispiel!

- Bitte berechnen Sie auf Basis der fiktiven Daten $\chi 2$!

Beobachtete Häufigkeiten

-

Partei-identifikation	Erhebungsgebiet		Gesamt
	West	Ost	
AfD	20	130	150
Andere Partei	1572	606	2178
Gesamt	1592	736	2328

Erwartete Häufigkeiten

Partei-identifikation	Erhebungsgebiet		Gesamt
	West	Ost	
AfD	103	47	150
Andere Partei	1489	689	2178
Gesamt	1592	736	2328

$\chi 2$-basierte Zusammenhangsmaße

- **Weiterentwicklung** von Chi-Quadrat, um Limitationen zu beseitigen
- Chi-Quadrat ist abhängig von den absoluten Häufigkeiten in den Zellen
- Verdopplung der Häufigkeiten = Verdopplung von $\chi 2$ (Prozentuale Verteilung ändert sich jedoch nicht!)

- Lösung: „**Normierung**“ des $\chi 2$-Wertes mit Phi und Cramer‘s V

$\chi 2$-basierte Zusammenhangsmaße: ϕ (Phi)

- **Für 2x2-Kreuztabellen!**

- Ziel: Relativierung des χ2-Wertes für die Anzahl der Beobachtungen

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

- ϕ variiert zwischen 0 (min.) und 1 (max.)
- Übung: Bitte berechnen Sie für die letzte Übung (Parteiidentifikation Ost/West) ebenfalls ϕ

$\chi 2$-basierte Zusammenhangsmaße: Cramer‘s V

- Weiterentwicklung, die auch Vergleiche über unterschiedliche große Kreuztabellen hinweg ermöglicht
- Variiert zwischen 0 und +1

$$Cramer'sV = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n * (min(k,m) - 1)}}$$

Berechnung für Beispiel?

(sowie entsprechendes Lernmodul 3-Video)

Interpretation des Zusammenhangs aus sozialwissenschaftlicher Sicht (siehe auch Lehrbrief, S. 58 bzw. Lernmodul 3 im WBT)

Wert von Cramer's V (V) bzw. Betrag von Phi (|Φ|)	Interpretation
≤ 0,05	kein Zusammenhang
> 0,05 bis ≤ 0,10	sehr schwacher Zusammenhang
> 0,10 bis ≤ 0,20	schwacher Zusammenhang
> 0,20 bis ≤ 0,40	mittelstarker Zusammenhang
> 0,40 bis ≤ 0,60	starker Zusammenhang
> 0,60	sehr starker Zusammenhang

Quelle: Eigene Darstellung

$$V \in \begin{cases} [0{,}00; 0{,}10[\rightarrow \textit{kein Zusammenhang} \\ [0{,}10; 0{,}30[\rightarrow \textit{schwacher Zusammenhang} \\ [0{,}30; 0{,}60[\rightarrow \textit{mittlerer Zusammenhang} \\ [0{,}60; 1{,}00[\rightarrow \textit{starker Zusammenhang} \end{cases}$$

Quelle: Cleff 2011: 92.