JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

7. Einheit

Plan

- **Erstellung einer Indifferenztabelle**
- **Zusammenhangsmaße für nominalskalierte Variablen**

Lernziele

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

Indifferenztabelle, Hinführung

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

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

- Aber 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

Indifferenztabelle, Beispiel und Übung

Politisches Interesse und Geschlecht – Ergänzen Sie die Indifferenztabe 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

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 (Quadrierung der Residuen weil Summe der einfachen Residuen Null ist)

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

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

Erwartete H.

Parteiwahl	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, Cramers V und C

χ2-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 (Parteiwahl Ost/West) ebenfalls ϕ

χ2-basierte Zusammenhangsmaße: Kontingenzkoeffizient C

- Berücksichtigt die **Anzahl der Kategorien** einer Kreuztabelle

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

- hat im Prinzip ebenfalls Grenzen zwischen 0 und +1, aber der maximal erreichbare Wert ist abhängig von der Tabellengröße, d.h. variiert zwischen 0 und C_{max} (mit R = min (k,m)

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

Berechnung für Beispiel politisches Interesse und Geschlecht (Tafel, Tutorium)

χ2-basierte Zusammenhangsmaße: Cramer‘s V

- Weiterentwicklung von C, normiert den Kontingenzkoeffizienten und ermöglicht so auch Vergleiche über unterschiedliche große Kreuztabellen
- Variiert zwischen 0 und +1

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

Oder:

$$\sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n \cdot (R-1)}}$$

Berechnung für Beispiel (Tafel, Tutorium)

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

Interpretation des Zusammenhangs (siehe auch Lehrbrief, S. 58)

Wert von Cramer's V (V) bzw. Betrag von Phi ($\lvert\Phi\rvert$)	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

PRE-Maß Lambda λ

Weitere Möglichkeit den Zusammenhang zwischen 2 nominalskalierten Merkmalen zu untersuchen

Exkurs PRE-Maße

- PRE: „Proportional Reduction of Error“
- Verschiedene PRE-Maße in der Statistik
- Ausgangsfrage: Wie gut können die Werte einer abhängigen Variable durch die Werte einer unabhängigen Variable vorhergesagt werden?
- Folgen der gleichen Logik
- Setzen eine „gerichtete“ Hypothese voraus, d.h. wir kennen abhängige und unabhängige Variable

Exkurs PRE-Maße

Schrittweises Vorgehen:

1) Wie lautet die Prognose des Wertes der abhängigen Variable **ohne** Kenntnis der unabhängigen Variablen? (Vorhersagefehler E1)
2) Prognose des Wertes der abhängigen Variable **mit** Kenntnis der Verteilung der unabhängigen Variable (Vorhersagefehler E2)
3) Ermittlung des PRE-Maßes und Aussage, ob die Vorhersage durch die unabhängige Variable verbessert wurde. PRE = (E1 - E2) / E1

Je nach PRE-Maß werden die Fehler dabei unterschiedlich berechnet

Beispiel: PRE-Maß Lambda λ

Wie gut erklärt die Kanzlerpräferenz die Wahlabsicht einer Person?

Ohne Kenntnis der Kanzlerpräferenz ist der Modus die CDU-CSU-Wahlabsicht die beste Vorhersage

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

Schritt 1: Wie lautet die Prognose des Wertes der abhängigen Variable OHNE Kenntnis der unabhängigen Variablen?
→ Modus Wahlabsicht als bestmögliche Vorhersage (OHNE Kenntnis der uV)

Beispiel: PRE-Maß Lambda λ

Wie gut erklärt die Kanzlerpräferenz die Wahlabsicht einer Person?

Ohne Kenntnis der Kanzlerpräferenz ist der Modus die CDU-CSU-Wahlabsicht die beste Vorhersage

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

Schritt 1:
Vorhersagefehler E1:
345+186=531

Beispiel: PRE-Maß Lambda λ – Schritt 2

Wie gut erklärt die Kanzlerpräferenz die Wahlabsicht einer Person?

Mit Kenntnis der Kanzlerpräferenz

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

Schritt 2a CDU/Merkel: Wie lautet die Prognose des Wertes der abhängigen Variable MIT Kenntnis der unabhängigen Variablen?
→ Modus d. kombinierten Merkmalskombination
Nicht erklärt: 25+84=109

Beispiel: PRE-Maß Lambda λ – Schritt 2

Wie gut erklärt die Kanzlerpräferenz die Wahlabsicht einer Person?

Mit Kenntnis der Kanzlerpräferenz

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

Schritt 2b
SPD/Steinmeier: →
Modus der kombinierten Merkmalskombination
Steinmeier-Fans
Nicht erklärt:
15+102=117

Beispiel: PRE-Maß Lambda λ – Schritt 2 und Schritt 3

- Wie gut erklärt die Kanzlerpräferenz die Wahlabsicht einer Person?
- **Schritt 2 komplett**: Vorhersagefehler E2: 117+109=226
- **Schritt 3**: Ermittlung des PRE-Maßes und Aussage, ob die Vorhersage durch die unabhängige Variable verbessert wurde. PRE = (E1 - E2) / E1

$$\lambda = \frac{E_1 - E_2}{E_1} \qquad = \frac{(531 - 226)}{531} = 0.57$$

Beispiel: PRE-Maß Lambda λ – Interpretation

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{(531 - 226)}{531} = 0.57$$

- **λ kann Werte zwischen 0 und 1 annehmen (multipliziert mit 100 auch prozentuale Aussage möglich):**
  - 0: Berücksichtigung der uV bringt keine Verbesserung
  - 1: Mit Hilfe der uV lassen sich alle Fälle korrekt vorhersagen

→ **Interpretation:** Kenntnis der Kanzlerpräferenz verringert die Fehler bei der Vorhersage der Wahlabsicht um 57%

Übung: PRE-Maß Lambda λ

Wie gut erklärt das Geschlecht die Haarlänge einer Person? Berechnen Sie Lambda.

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

E1:
E2:
Lambda?