STATISTIK FÜR DIE SOZIALWISSENSCHAFTEN

Zusammenhänge Dichotomer Variablen beschreiben

Organisatorisches; Symmetrische und Asymmetrische Zusammenhänge;

Zusammenhänge Dichotomer Variablen

Themen für heute

- Organisatorisches
- Additive Zusammenhänge an Vierfeldertafeln
- Multiplikative Zusammenhänge an Vierfeldertafeln
- Odds Ratios (schätzen & testen)
- Yules Q
- (Formalismus an) Kreuztabellen

Meine Notizen:

Beispiel: Religiosität in Deutschland

- Es wird vermutet, dass (durch die Geschichte) eine Wirkung der Region (Ost/West) auf die Religiosität (Religionsmitgliedschaft) der Region gegeben ist.

	Alte Bundesländer (x_1)	Neue Bundesländer (x_2)	Gesamt
Mitglied einer Religionsgemeinschaft (y_1)	800 0,64 0,8 0,914	75 0,06 0,3 0,086	875 0,7 0,7 1
Nicht Mitglied einer Religionsgemeinschaft (y_2)	200 0,16 0,2 0,533	175 0,14 0,7 0,467	375 0,3 0,3 1
Gesamt	1000 0,8 1 0,8	250 0,2 1 0,2	1250 1 1 1

Meine Notizen:

Prozentsatzdifferenz d_{YX}

- Die Prozentsatzdifferenz d_{YX} für eine unabhängige Variable X und die Abhängige Y berechnet sich mit einer der drei Varianten:

$$d_{YX}\% = 100 * (p_{1(1)} - p_{1(2)}) = 100 * \left(\frac{n_{11}}{n_{+1}} - \frac{n_{12}}{n_{+2}} \right) = 100 * \left(\frac{p_{11}}{p_{+1}} - \frac{p_{12}}{p_{+2}} \right)$$

- Die Einheit der Prozentsatzdifferenz sind Prozentpunkte.
- Die Stärke des asymmetrischen Zusammenhangs kann mit folgenden Faustregeln bewertet werden:

	$-5\% < d_{YX}\% < +5\%$	praktisch kein Zusammenhang
	$+5\% \leq d_{YX}\% < +20\%$	geringer Zusammenhang
bzw.	$-20\% < d_{YX}\% \leq -5\%$	
	$+20\% \leq d_{YX}\% < +50\%$	mittlerer Zusammenhang
bzw.	$-50\% < d_{YX}\% \leq -20\%$	
	$+50\% \leq d_{YX}\%$	starker Zusammenhang
bzw.	$-50\% \leq d_{YX}\%$	

Meine Notizen:

Beispiel: Religiosität in Deutschland

	Alte Bundesländer (x_1)	Neue Bundesländer (x_2)	Gesamt
Mitglied einer Religionsgemeinschaft (y_1)	800 0,64 0,8 0,914	75 0,06 0,3 0,086	875 0,7 0,7 1
Nicht Mitglied einer Religionsgemeinschaft (y_2)	200 0,16 0,2 0,533	175 0,14 0,7 0,467	375 0,3 0,3 1
Gesamt	1000 0,8 1 0,8	250 0,2 1 0,2	1250 1 1 1

$$d_{YX}\% = 100 * (p_{1(1)} - p_{1(2)}) = 100 * \left(\frac{n_{11}}{n_{+1}} - \frac{n_{12}}{n_{+2}} \right) = 100 * \left(\frac{p_{11}}{p_{+1}} - \frac{p_{12}}{p_{+2}} \right) =$$

$$100 * (0,8 - 0,3) = 100 * \left(\frac{800}{1000} - \frac{75}{250} \right) =$$

$$100 * \left(\frac{0,64}{0,8} - \frac{0,06}{0,2} \right) = 50$$

Meine Notizen:

Symmetrische Zusammenhänge Dichotomer Variablen

Meine Notizen:

Beispiel: Religiosität in Deutschland

- Neben der historischen Wirkung der Region auf die Religion könnte es heute eine Wirkung der Religiosität auf die Wohnortwahl geben. Beidseitige Wirkungen bedürfen wie unklare Wirkungen eines symmetrischen Zusammenhangsmaßes.

	Alte Bundesländer (x_1)	Neue Bundesländer (x_2)	Gesamt
Mitglied einer Religionsgemeinschaft (y_1)	800 0,64 0,8 0,914	75 0,06 0,3 0,086	875 0,7 0,7 1
Nicht Mitglied einer Religionsgemeinschaft (y_2)	200 0,16 0,2 0,533	175 0,14 0,7 0,467	375 0,3 0,3 1
Gesamt	1000 0,8 1 0,8	250 0,2 1 0,2	1250 1 1 1

Meine Notizen:

Symmetrischer Zusammenhang Phi Φ

- Bekannter Weise sind Zusammenhänge auch dann als symmetrisch zu beurteilen, wenn beide Wirkrichtungen sinnvoll theoretisch angenommen werden können.
- Ausgehend davon lassen sich die asymmetrischen Prozentsatzdifferenzen nutzen, um ein symmetrisches Maß zu berechnen: als Mittel der beiden Prozentsatzdifferenzen
- Dieses symmetrische Maß Phi berechnet sich als geometrisches Mittel von d_{YX} und d_{XY} :

$$|\Phi| = \sqrt{\frac{d_{XY}\%}{100} * \frac{d_{YX}\%}{100}}$$

- Liegen keine berechneten Prozentsatzdifferenzen vor, lässt Φ sich direkt berechnen durch:

$$\Phi = \frac{n_{11} * n_{22} - n_{12} * n_{21}}{\sqrt{(n_{11} + n_{12}) * (n_{21} + n_{22}) * (n_{11} + n_{21}) * (n_{12} + n_{22})}}$$

Meine Notizen:

Beispiel: Religiosität in Deutschland

	Alte Bundesländer (x_1)	Neue Bundesländer (x_2)	Gesamt
Mitglied einer Religionsgemeinschaft (y_1)	800 0,64	75 0,06	875 0,7
Nicht Mitglied einer Religionsgemeinschaft (y_2)	200 0,16	175 0,14	375 0,3
Gesamt	1000 0,8	250 0,2	1250 1

$$\Phi = \frac{n_{11} * n_{22} - n_{12} * n_{21}}{\sqrt{(n_{11} + n_{12}) * (n_{21} + n_{22}) * (n_{11} + n_{21}) * (n_{12} + n_{22})}} =$$
$$\frac{800 * 175 - 75 * 200}{\sqrt{(800 + 75) * (200 + 175) * (800 + 200) * (75 + 175)}} = 0,4364$$

Die Interpretation von Φ folgt den Faustformeln der Prozentsatzdifferenz.

Meine Notizen:

Multiplikative Zusammenhänge Dichotomer Variablen

Meine Notizen:

Grenzen bisheriger Zusammenhangsmessung

- Maximale Zusammenhangswerte sind nur realisierbar, wenn entweder nur die Hauptdiagonale oder nur die Gegendiagonale besetzt ist.
- Zeilen- und Spaltenvariable müssen also absolut gleich verteilt sein.
- Bei Abweichungen von gleicher Verteilung: Je schiefer X und/oder Y sind, desto geringer ist maximal möglicher Zusammenhang.
- Beispiel:

	Alte Bundesländer (x_1)	Neue Bundesländer (x_2)	Gesamt
Mitglied einer Religionsgemeinschaft (y_1)	100 0,1	2 0,008	102 0,0816
Nicht Mitglied einer Religionsgemeinschaft (y_2)	900 0,9	248 0,992	1148 0,9184
Gesamt	1000 1	250 1	1250 1

Meine Notizen:

Grenzen bisheriger Zusammenhangsmessung

- Beispiel:

	Alte Bundesländer (x_1)	Neue Bundesländer (x_2)	Gesamt
Mitglied einer Religionsgemeinschaft (y_1)	100 0,1	2 0,008	102 0,0816
Nicht Mitglied einer Religionsgemeinschaft (y_2)	900 0,9	248 0,992	1148 0,9184
Gesamt	1000 1	250 1	1250 1

- Werte additiver asymmetrischer und symmetrischer Zusammenhangsmaße sind randverteilungsabhängig.
- Bei Anteils- und Prozentsatzdifferenzen gilt dies nur für die Randverteilung der abhängigen Variable.

Meine Notizen:

Alternative deshalb: Odds-Ratios als ein multiplikatives Zusammenhangsmaß

- Statt einer Differenz (wie bei der Prozentsatzdifferenz und damit auch bei Phi) wird ein Quotient gebildet.
- Der Quotient heißt Odds-Ratio oder auch Kreuzproduktverhältnis α

$$\alpha = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} = \frac{p_{11} * p_{22}}{p_{12} * p_{21}} = \frac{\frac{n_{11}}{n_{1+}} * \frac{n_{22}}{n_{2+}}}{\frac{n_{21}}{n_{2+}} * \frac{n_{12}}{n_{1+}}} = \frac{\frac{n_{11}}{n_{+1}} * \frac{n_{22}}{n_{+2}}}{\frac{n_{21}}{n_{+2}} * \frac{n_{12}}{n_{+1}}}$$

Meine Notizen:

Odds-Ratios am Beispiel

	Alte Bundesländer (x ₁)	Neue Bundesländer (x ₂)	Gesamt
Mitglied einer Religionsgemeinschaft (y ₁)	100 0,1	2 0,008	102 0,0816
Nicht Mitglied einer Religionsgemeinschaft (y ₂)	900 0,9	248 0,992	1148 0,9184
Gesamt	1000 1	250 1	1250 1

$$\alpha = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} = \frac{p_{11} * p_{22}}{p_{12} * p_{21}} = \frac{\frac{n_{11}}{n_{1+}} * \frac{n_{22}}{n_{2+}}}{\frac{n_{21}}{n_{2+}} * \frac{n_{12}}{n_{1+}}} = \frac{\frac{n_{11}}{n_{+1}} * \frac{n_{22}}{n_{+2}}}{\frac{n_{21}}{n_{+2}} * \frac{n_{12}}{n_{+1}}} = \frac{100 * 248}{900 * 2} = 13,778$$

Meine Notizen:

Interpretation von Odds-Ratios α

- Der Wertebereich von α reicht von $0 \leq \alpha \leq \infty$. Negative Zusammenhänge werden durch $\alpha < 1$ ausgedrückt, Positive durch $\alpha > 1$. Unabhängigkeit zeigt sich in $\alpha = 1$.
- Der Wert von α gibt an, um welchen Faktor das Produkt der Hauptdiagonalelemente größer bzw. kleiner ist als das Produkt der Nebendiagonalelemente.
- Bei Interpretation als Kreuzproduktverhältnis in einer Vierfeldertabelle wird α als symmetrisches Zusammenhangsmaß interpretiert.
- Der Koeffizient kann aber auch als asymmetrische Maß interpretiert werden: α gibt an, um welchen Faktor sich das Häufigkeits-, Anteils- oder Wahrscheinlichkeitsverhältnis der beiden Ausprägungen der abhängigen Variable ändert, wenn statt der zweiten Ausprägung die erste Ausprägung der UV betrachtet wird.

Meine Notizen:

Interpretation von Odds-Ratios α

$$\alpha = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} = \frac{p_{11} * p_{22}}{p_{12} * p_{21}} = \frac{\frac{n_{11}}{n_{1+}} * \frac{n_{22}}{n_{2+}}}{\frac{n_{21}}{n_{2+}} * \frac{n_{12}}{n_{1+}}} = \frac{\frac{n_{11}}{n_{+1}} * \frac{n_{22}}{n_{+2}}}{\frac{n_{21}}{n_{+2}} * \frac{n_{12}}{n_{+1}}} = \frac{100 * 248}{900 * 2} = 13,778$$

- „Im Westen ist der Anteil der Religionsgemeinschaftsmitglieder 13,778-mal so hoch wie im Osten.“
- „Im Osten ist der Anteil der Religionsgemeinschaftsmitglieder 0,0726-mal so hoch wie im Westen.“ (Kehrwert)

Meine Notizen:

Folgen der Interpretation für die Schätzung

- Ein Odds-Ratio von 3 und eines von 6 liegen nicht gleich weit auseinander wie ein Odds-Ratio von 103 und eines von 106.
- Die multiplikative Logik sorgt deshalb dafür, dass Konfidenzintervalle von Odds-Ratios den Stichprobenwert nicht genau in der Mitte des Intervalls haben.
- Wie bekommen wir stattdessen die Intervallbreite heraus?
- Wir bedienen uns der Logik, die Odds-Ratios zu logarithmieren.
- Besonders der natürliche Logarithmus ist dafür geeignet, da er Kehrwerte in betragsmäßig gleiche Werte überführt.
- Wir übersetzen also unsere Odds-Ratio α in die Einheit $\ln(\alpha)$

Meine Notizen:

Schätzung von Odds-Ratios α

$$\alpha = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} \rightarrow \ln(\alpha) = \ln\left(\frac{n_{11} * n_{22}}{n_{12} * n_{21}}\right) = \ln(n_{11}) - \ln(n_{12}) - \ln(n_{21}) + \ln(n_{22})$$

- Es lässt sich dann herleiten, dass:

$$\hat{\sigma}^2(\ln(\alpha)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

- Der Standardfehler von $\ln(\alpha)$ entspricht also der Wurzel der Summe der Kehrwerte der vier Felder.

Meine Notizen:

Schätzung von Odds-Ratios am Beispiel

$$\alpha = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} = \frac{100 * 248}{900 * 2} = 13,778 \quad \rightarrow \ln(\alpha) = 2,623$$

$$\hat{\sigma}^2(\ln(\alpha)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} = \frac{1}{100} + \frac{1}{900} + \frac{1}{2} + \frac{1}{248} = 0,515 \quad \rightarrow \hat{\sigma}(\ln(\alpha)) = 0,718$$

$$90\% - c.i. (\ln(\alpha)) = 2,623 \pm 1,645 * 0,718 = [1,442 ; 3,804]$$

$$\rightarrow 90\% - c.i. (\alpha) = [e^{1,442} ; e^{3,804}] = [4,231 ; 44,864]$$

	Alte Bundesländer (x ₁)	Neue Bundesländer (x ₂)	Gesamt
Mitglied einer Religionsgemeinschaft (y ₁)	100 0,1	2 0,008	102 0,0816
Nicht Mitglied einer Religionsgemeinschaft (y ₂)	900 0,9	248 0,992	1148 0,9184
Gesamt	1000 1	250 1	1250 1

Meine Notizen:

Testen von Odds-Ratios: Die Hypothesenaufstellung

- Genau wie in der Schätzung von Odds-Ratios müssen für einen Test Odds-Ratios in die Einheit $\ln(\alpha)$ übersetzt werden.

- Soll gegen einen bestimmten Wert $\alpha_{\text{pop.}}$ getestet werden, funktioniert die Hypothesenaufstellung wie folgt:

- zweiseitiger Test: $H_0: \alpha_{\text{pop.}} = \alpha_0$ vs. $H_1: \alpha_{\text{pop.}} \neq \alpha_0 \rightarrow H_1: \alpha_{\text{pop.}}/\alpha_0 \neq 1$
- einseitiger Test nach oben: $H_0: \alpha_{\text{pop.}} \leq \alpha_0$ vs. $H_1: \alpha_{\text{pop.}} > \alpha_0 \rightarrow H_1: \alpha_{\text{pop.}}/\alpha_0 > 1$
- einseitiger Test nach unten: $H_0: \alpha_{\text{pop.}} \geq \alpha_0$ vs. $H_1: \alpha_{\text{pop.}} < \alpha_0 \rightarrow H_1: \alpha_{\text{pop.}}/\alpha_0 < 1$

- Durch natürliche Logarithmieren entsteht daraus:

- zweiseitiger Test: $H_0: \ln(\alpha_{\text{pop.}}) = \ln(\alpha_0)$ vs. $H_1: \ln(\alpha_{\text{pop.}}) \neq \ln(\alpha_0)$
- einseitiger Test nach oben: $H_0: \ln(\alpha_{\text{pop.}}) \leq \ln(\alpha_0)$ vs. $H_1: \ln(\alpha_{\text{pop.}}) > \ln(\alpha_0)$
- einseitiger Test nach unten: $H_0: \ln(\alpha_{\text{pop.}}) \geq \ln(\alpha_0)$ vs. $H_1: \ln(\alpha_{\text{pop.}}) < \ln(\alpha_0)$.

Testen von Odds-Ratios: Die Teststatistik

- Als Teststatistik für $\ln(\alpha)$ gilt:

$$Z = \frac{\ln(\alpha) - \ln(\alpha_0)}{\hat{\sigma}(\ln(\alpha))} = \frac{\ln\left(\frac{n_{11} * n_{22}}{n_{12} * n_{21}}\right) - \ln(\alpha_0)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

Meine Notizen:

Hypothesentest: Schritt 1: Hypothesen formulieren

- Forschungshypothese:

„Im Westen ist die Chance der Religionsgemeinschaftsmitgliedschaft mehr als fünfmal so hoch wie im Osten.“

- Gegenhypothese:

„Im Westen ist die Chance der Religionsgemeinschaftsmitgliedschaft nicht mehr als fünfmal so hoch wie im Osten.“

- Mathematisierung:

$$H_0: \alpha_{\text{Pop.}} \leq 5 \text{ vs. } H_1: \alpha_{\text{Pop.}} > 5 \quad \rightarrow \quad H_1: \alpha_{\text{Pop.}}/5 > 1$$

$$H_0: \ln(\alpha_{\text{Pop.}}) \leq \ln(5) \quad \text{vs.} \quad H_1: \ln(\alpha_{\text{Pop.}}) > \ln(5)$$

Meine Notizen:

Hypothesentest: Schritt 2: Teststatistik berechnen

- Die Teststatistik berechnet sich als

$$Z = \frac{\ln(\alpha) - \ln(\alpha_0)}{\hat{\sigma}(\ln(\alpha))} = \frac{\ln\left(\frac{n_{11} * n_{22}}{n_{12} * n_{21}}\right) - \ln(\alpha_0)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} = \frac{2,623 - \ln(5)}{0,718} = 1,412$$

Meine Notizen:

Schritt 3: Kritischen Wert bestimmen

(1) Irrtumswahrscheinlichkeit festlegen:	$\alpha = 0,05$
(2) Verteilung:	Z-Verteilung
(3) Art des Testes berücksichtigen:	Rechtsseitiger Test
(4) Kritischen Wert aus der Tabelle ablesen.	$Z_{0,95} = 1,645$

Meine Notizen:

Hypothesentest: Schritt 4: Entscheidung treffen

- Da es sich um einen rechtsseitigen Test handelt, muss die Teststatistik rechts des kritischen Wertes liegen, um die H_0 ablehnen zu können.
- Im Beispiel ist die Teststatistik nicht rechts des kritischen Wertes, da $1,412 < 1,645$.
- Die Nullhypothese kann damit nicht verworfen werden. Die Stichprobe ist bei $\alpha = 5\%$ nicht geeignet zu zeigen, dass der Einfluss der Region auf die Religionsmitgliedschaft größer als eine fünffache Chance ist.

Meine Notizen:

Eine Alternative: Yules Q

Meine Notizen:

Das Maß Yules Q

- Eine leichter auszurechnende Alternative zum symmetrischen Zusammenhangsmaß Phi ist Yules Q.

- Es berechnet sich nach:

$$Q = \frac{n_{11} * n_{22} - n_{21} * n_{12}}{n_{11} * n_{22} + n_{21} * n_{12}}$$

- Die Werte von Yules Q liegen zwischen -1 und 1.
- Null steht für keinen Zusammenhang, 1 bzw. -1 für einen perfekten positiven bzw. negativen Zusammenhang.

Meine Notizen:

Nachteile von Yules Q

- Yules Q nimmt leichter Werte weiter weg von 0 an als Phi. Die Interpretation nach gegebener Faustregeltabelle ist daher kritisch.
- Außerdem ist Yules Q sehr anfällig für eine Null in der Vierfeldertabelle. Eine einzige Null produziert sofort ein Q von +1 oder -1.
- Yules Q kann daher die folgenden Tabellen nicht unterscheiden:

	Arbeiter	Angestellte	Arbeiter	Angestellte
Haupt-/Realschule	400	590	400	0
Gymnasium	0	10	0	600

Meine Notizen:

Zusammenhänge Nominaler Variablen beschreiben

Meine Notizen:

Beispiel

	katholisch (x_1)	Lutherisch (x_2)	Ohne Konfession (x_3)	Gesamt
Linke (y_1)	14	18	91	123
	0,0148	0,0190	0,0961	0,1299
	0,0569	0,0520	0,2563	0,1299
	0,1138	0,1463	0,7398	1
Grüne (y_2)	41	55	49	145
	0,0433	0,0581	0,0517	0,1531
	0,1667	0,1590	0,1380	0,1531
	0,2828	0,3793	0,3379	1
SPD (y_3)	64	128	104	296
	0,0676	0,1352	0,1098	0,3126
	0,2602	0,3699	0,2930	0,3126
	0,2162	0,4324	0,3514	1
CDU/CSU (y_4)	127	145	111	383
	0,1341	0,1531	0,1172	0,4044
	0,5163	0,4191	0,3127	0,4044
	0,3316	0,3786	0,2898	1
Gesamt	246	346	355	947
	0,2598	0,3654	0,3749	1
	1	1	1	1
	0,2598	0,3654	0,3749	1

Formalismus

	katholisch (x_1)	Lutherisch (x_2)	Ohne Konfession (x_3)	Gesamt
Linke (y_1)	n_{11} p_{11} $p_{1(1)}$ $p_{(1)1}$	n_{12} p_{12} $p_{1(2)}$ $p_{(1)2}$	n_{13} p_{13} $p_{1(3)}$ $p_{(1)3}$	n_{1+} p_{1+} $p_{1(+)}$ 1
Grüne (y_2)	n_{21} p_{21} $p_{2(1)}$ $p_{(2)1}$	n_{22} p_{22} $p_{2(2)}$ $p_{(2)2}$	n_{23} p_{23} $p_{2(3)}$ $p_{(2)3}$	n_{2+} p_{2+} $p_{2(+)}$ 1
SPD (y_3)	n_{31} p_{31} $p_{3(1)}$ $p_{(3)1}$	n_{32} p_{32} $p_{3(2)}$ $p_{(3)2}$	n_{33} p_{33} $p_{3(3)}$ $p_{(3)3}$	n_{3+} p_{3+} $p_{3(+)}$ 1
CDU/CSU (y_4)	n_{41} p_{41} $p_{4(1)}$ $p_{(4)1}$	n_{42} p_{42} $p_{4(2)}$ $p_{(4)2}$	n_{43} p_{43} $p_{4(3)}$ $p_{(4)3}$	n_{4+} p_{4+} $p_{4(+)}$ 1
Gesamt	n_{+1} p_{+1} 1 $p_{(+)1}$	n_{+2} p_{+2} 1 $p_{(+)2}$	n_{+3} p_{+3} 1 $p_{(+)3}$	$n=n_{++}$ 1 1 1

Was Sie am Ende der Woche können sollten

- **Kern:** Sie analysieren in Vierfeldertafeln die Notwendigkeit einer multiplikativen Betrachtung und bestimmen Odds-Ratios zur Interpretation dieser.
- Sie erklären Schwächen additiver Zusammenhänge bei schiefen Randverteilungen.
- Sie evaluieren den Nutzen multiplikativer Betrachtungen, wie auch deren Gefahren.
- Sie berechnen Odds-Ratios und Yules Q.
- Sie schätzen ein Odds-Ratio oder testen es, um multiplikative Aussagen über die soziale Welt zu treffen.
- Sie erklären, warum das Konfidenzintervall von Odds-Ratios nicht dieselbe Symmetrie hat wie andere Konfidenzintervalle.

Meine Notizen: