Zusammenhangsmaße I

Statistik I

Sommersemester 2009

Nächste Woche: Maße für ordinale, nominal/intervallskalierte und intervallskalierte Daten

$\chi^2 =?!?$

Zum Nachlesen

- Agresti/Finlay: Kapitel 8.1-8.4
- Gehring/Weins: Kapitel 7.1
- Schumann: Schumann Kapitel 7

Verteilungen

- Verteilungen von Variablen haben ...
- Eine Streuung um diese Mitte
- Sie sind symmetrisch oder schief
- Und haben eine mehr oder minder breiten Gipfel
- **Beschreibung der Eigenschaften durch Maßzahlen**

Univariate vs. bivariate Verteilungen

- Letzte Woche: *univariate* Verteilungen – eine einzige Variable
- Aber: LRS in drei Ländern: *konditionale* Verteilung – Zusammenhang?
- *Gemeinsame* Verteilung von zwei Variablen: *bivariate* Verteilung – zeigt Zusammenhang (oder auch nicht)

Was ist ein Zusammenhang?

- Allgemein: gemeinsames „Muster“ in der Verteilung zweier Variablen (kausal?)
- Beispiele
  - Arbeiter wählen häufiger die SPD als andere Gruppen
  - Hochgebildete vertreten häufiger postmaterialistische Werte als Niedriggebildete
  - Männer haben ein höheres Durchschnittsgehalt als Frauen
  - Je älter ein Befragter ist, desto höher ist auch sein Wert auf einer Konservatismusskala

Warum sind Zusammenhänge interessant?

- Zusammenhänge erlauben Prognosen
- Zusammenhänge = empirische Entsprechung von Hypothesen

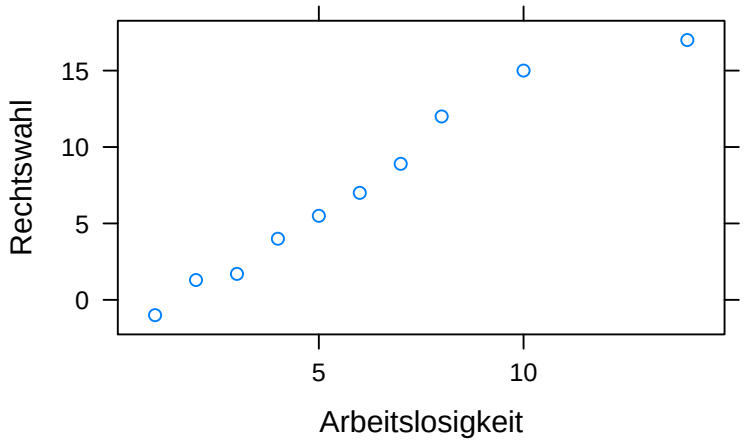
Welche Arten von Zusammenhängen gibt es?

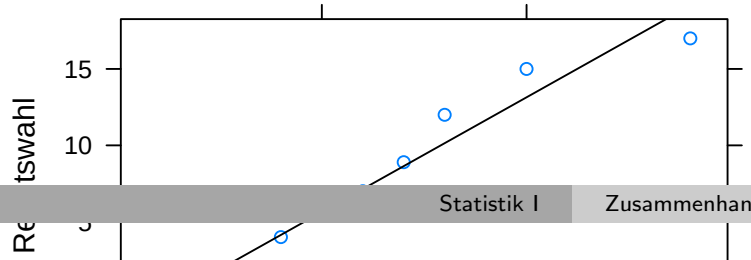
- Unterschieds- vs. Zusammenhangshypothesen – strukturell identisch
- Durchschnittseinkommen von Männern/Frauen unterscheidet sich = Zusammenhang zwischen Geschlecht und Einkommen
- Gerichtete vs. ungerichtete Hypothesen $\rightarrow$ gerichtete vs. ungerichtete Zusammenhänge
  - Einkommen von Männern ist höher vs.
  - Einkommen unterscheidet sich
- Gerichtete Zusammenhänge
  - positiv: mehr von x, mehr von y; weniger von x ... ?
  - negativ: mehr von x, weniger von y; weniger von x mehr von y
- Starke (perfekte) vs. schwache Zusammenhänge

Welche Arten von Zusammenhängen gibt es? II

1. (Näherungsweise) lineare Zusammenhänge
2. (Monotone Zusammenhänge)
3. Kurvilineare und andere nicht-lineare Zusammehänge

Was sind lineare Zusammenhänge?

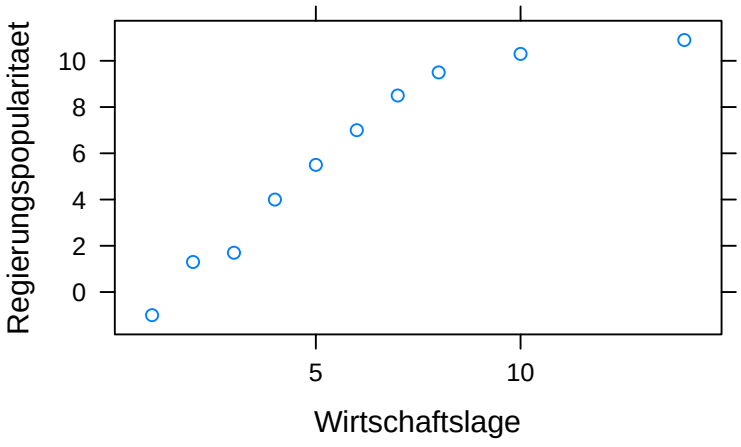


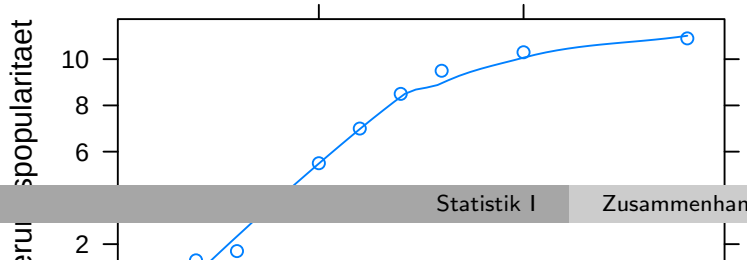


Was sind monotone Zusammenhänge

- Ändern ihre Richtung nicht
- Unabhängig vom Niveau von x Zunahme von x immer mit Zunahme (oder Abnahme) von y verbunden
- Linear oder nicht-linear

Monotoner nicht-linearer Zusammenhang

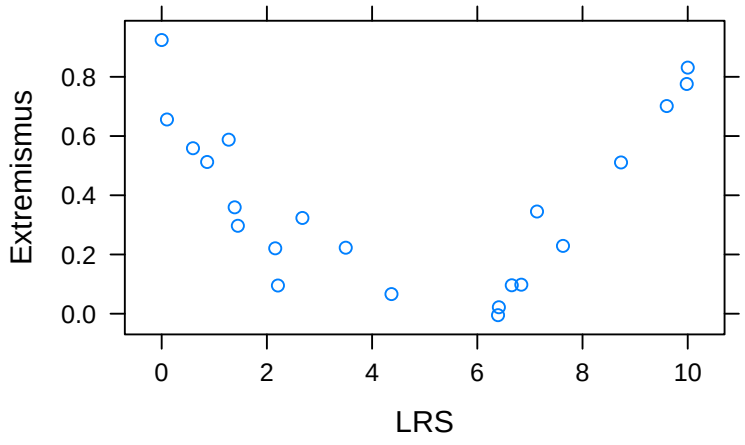


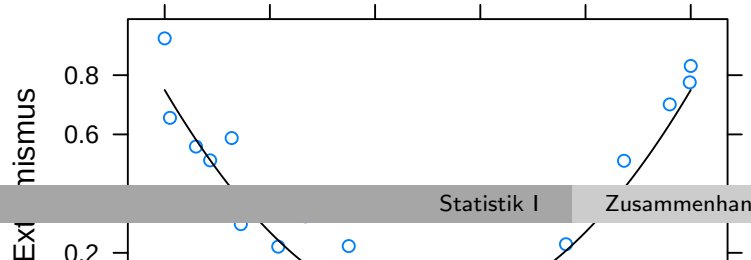


Was sind nicht lineare Zusammenhänge?

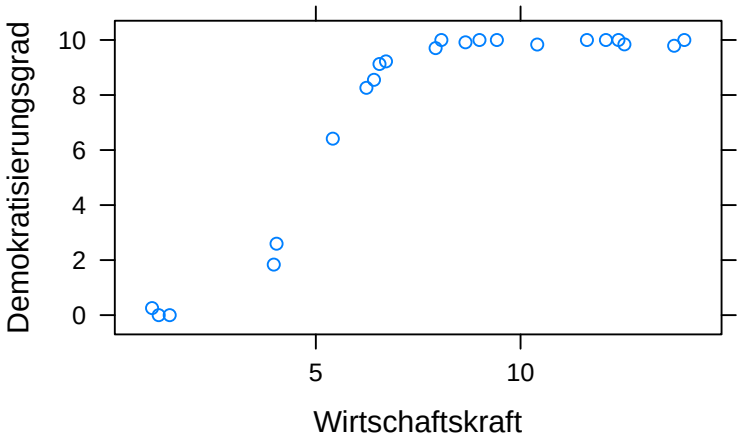
- Zusammenhang nicht in Form einer Linie (Zusammenhang nicht über Wertebereich konstant)
- Monoton/nicht monton
- Einige wichtige Formen:
  1. Kurvilinear (U-förmig, nicht-monoton)
  2. Exponential (monoton)
  3. Sigmoidal (S-förmig, monoton)

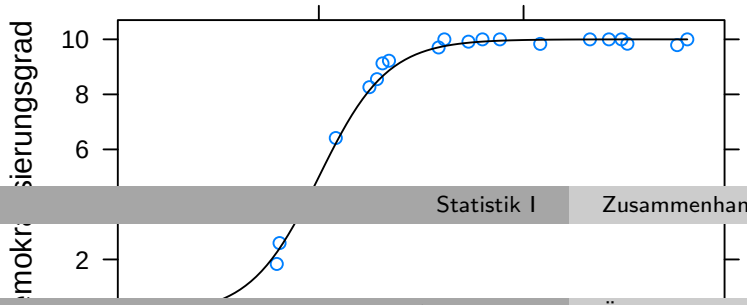
Beispiel für einen kurvilinearen Zusammenhang





Beispiel für einen sigmoidalen Zusammenhang





Wiederholung/Einführung
Zusammenhangsmaße
Zusammenfassung
Überblick
Nominale Daten: χ^2
Nominale Daten: PRE-Maß

Zusammenhänge und Skalenniveaus

- Gerichtete Zusammenhänge – geordnete Kategorien
- Für jedes Niveau passende Zusammenhangsmaße + Hierarchie
- Richtung im engeren Sinn setzt ordinales Skalenniveau voraus

Keine Richtung bei nominalen Daten?

- Katholiken sollten häufiger CDU wählen als Protestanten
- Protestanten sollten häufiger CDU wählen als Konfessionslose
- Warum?

Wenn Daten nominal sind . . .

- Besteht der Zusammenhang darin, daß es Unterschiede gibt
- Bei wenigen Kategorien und klaren Hypothesen: Annahmen *wo* Unterschiede auftreten
- Können Daten für diesen Zweck dieser Hypothese als ordinal betrachtet werden?

Was sind Zusammenhangsmaße?

- Quantifizieren Stärke (und ggf. Richtung) eines Zusammenhanges
- Definierter Wertebereich $\rightarrow$ Vergleich von zwei Zusammenhänge
- Idealerweise zwischen 0 und 1 bzw. -1 und 1
- Vielzahl von Zusammenhangsmaßen für verschiedene Skalenniveaus
- Basieren auf *gemeinsamer Verteilung zweier Variablen* – besonders bei nominalen Variablen gut zu erkennen

Was ist ein Zusammenhang zwischen zwei nominalen Variablen?

Nominaler Zusammenhang

- Ausprägungen treten
- Häufiger/seltener gemeinsam auf
- Als bei *zufälliger* Verteilung zu erwarten wäre

- Verteilungen nicht unabhängig voneinander $\rightarrow$ konditionale $\neq$ marginale Verteilung
- Verteilung unter Annahme von Unabhängigkeit: Wahrscheinlichkeiten für Ausprägungen miteinander multiplizieren
- Wahrscheinlichkeit für Ausprägung: relative Häufigkeit des Merkmals (mehr dazu in einigen Wochen)

Ein einfaches Beispiel: Konfession und Geschlecht

- Der Anteil von Männern/Frauen in einer Gesellschaft (GG) beträgt 50%
- Der Anteil der Katholiken beträgt 40%
- *Wenn* kein Zusammenhang besteht/beide Variablen unabhängig sind:
  - Wahrscheinlichkeit $p(m)$ daß eine zufällig ausgewählte Person männlich ist: ?
  - Wahrscheinlichkeit $p(k)$ daß eine zufällig ausgewählte Person katholisch ist: ?
  - Wahrscheinlichkeit $p(k)$ daß eine zufällig ausgewählte Person männlich **und** katholisch ist: $0.5 \times 0.4 = 0.2 = 20\%$

Was weiter?

- Wenn realer Anteil $\neq 20\% \rightarrow$ männliche Katholiken „überzufällig“ häufig/selten $\rightarrow$ zusammenhang
- Statistischer Zusammenhang („Korrelation“)
  - Zusammenhang auch in GG vorhanden (Schätzung für GG, Stichprobenfehler)?
  - Starker/schwacher Zusammenhang?
- Logik des überzufällig häufigen gemeinsamen Auftretens (Verteilungen nicht unabhängig voneinander) unabhängig vom Skalenniveau

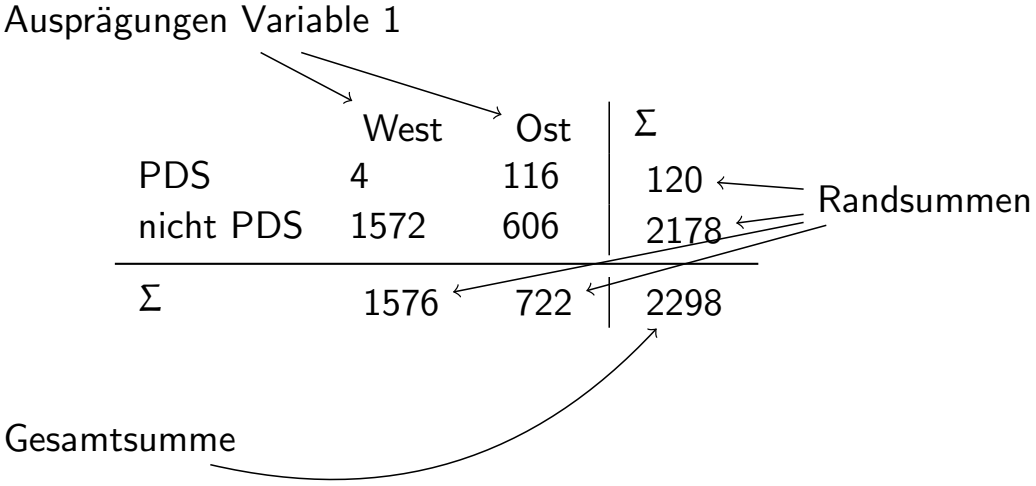
Maße auf der Basis von χ^2

- Verbindung zum „überzufällig häufigen“ gemeinsamen Auftreten besonders klar
- Vergleichen empirische Kreuztabelle
- Mit Tabelle die Häufigkeiten (gemeinsame Verteilung) unter Annahme zeigt
- Daß kein Zusammenhang besteht (Indifferenztablle)

Was ist eine Kreuztabelle?

- „Kreuzt“ zwei kategoriale Variablen miteinander
- Gemeinsame Häufigkeitstabelle für zwei Variablen
- Am Rand Rand-/Gesamtverteilungen (marginale Verteilung)
- In der Mitte gemeinsame (konditionale) Verteilung(en)

Was ist eine Kreuztabelle II



Welche Prozentuierungsarten gibt es in der Kreuztabelle?

- Zeilenprozente setzen den Inhalt einer Zelle zur Zeilensumme ins Verhältnis
  - Wieviel Prozent aller PDS-Wähler sind Westdeutsche?
  - Wieviel Prozent aller Wähler insgesamt sind Westdeutsche
- Spaltenprozente setzen den Inhalt einer Zelle zur Spaltensumme ins Verhältnis
  - Wieviel Prozent der westdeutschen Wähler stimmen für die PDS?
  - Wie hoch ist der Anteil der PDS-Wähler insgesamt?
- Totalprozente setzen den Inhalt einer Zelle zur Gesamtsumme ins Verhältnis
  - Wie hoch ist der Anteil der westdeutschen PDS-Wähler an allen Wählern

Zeilenprozente $\rightarrow$

	West	Ost	Σ
PDS	4/120=3.3%	116/120=96.6%	120/120=100%
nicht PDS	1572/2178=72.1%	606/2178=27.8%	2178/2178=100%
Σ	1576/2298=68.6%	722/2298=31.4%	2298/2298=100%

Spaltenprozente $\downarrow$

	West	Ost	Σ
PDS	4/1576=0.3%	116/722=16.1%	120/2298=5.2%
nicht PDS	1572/1576=99.7%	606/722=83.9%	2178/2298=94.8%
Σ	1576/1576=100%	722/722=100%	2298/2298=100%

- Zweimal dieselbe Information:
- Verteilungen der beiden Variablen sind nicht unabhängig voneinander
- Bzw. weichen von zufälligem Muster (Indifferenztabelle) ab

Wie wird die Indifferenztabelle konstruiert?

	West	Ost	Σ	$\downarrow$
PDS	4	116	120	5.2%
nicht PDS	1572	606	2178	94.8%
Σ	1576	722	2298	
$\rightarrow$	72.1%	31.4%	2298	

- Verteilung des ersten Merkmals: **Zeilensummen** am unteren Rand $\rightarrow$ Wahrscheinlichkeit ostdeutsch/westdeutsch
- Verteilung des zweiten Merkmals: **Spaltensummen** am rechten Rand $\rightarrow$ Wahrscheinlichkeit PDS-Wahl/nicht-PDS-Wahl
- Wahrscheinlichkeit ostdeutsch PDS-Wähler (bei Unabhängigkeit der Variablen): $0.314 \times 0.052 = 0.016328 \approx \frac{1}{3} \times \frac{1}{20}$
- Absolute zu erwartende Zahl der ostdeutschen PDS-Wähler: $0.016328 \times 2298 \approx 37.5$

Wie wird die Indifferenztablle konstruiert? II

- Allgemein: Multiplikation der Randwahrscheinlichkeiten $\rightarrow$ gemeinsame Wahrscheinlichkeit
- Gemeinsame Wahrscheinlichkeit mal Gesamtzahl Beobachtungen $\rightarrow$ erwartete Häufigkeit
- Vereinfacht:

$$\frac{\text{Spaltensumme}}{\text{Gesamtsumme}} \times \frac{\text{Zeilensumme}}{\text{Gesamtsumme}} \times \text{Gesamtsumme} = \frac{\text{Spaltensumme} \times \text{Zeilensumme}}{\text{Gesamtsumme}}$$

- Berechnung wird für alle Zellen im Inneren der Tabelle vorgenommen
- Randsummen und Gesamtsummen bleiben konstant

Wie wird die Indifferenztabelle konstruiert? III

	West	Ost	Σ
PDS	$\frac{1576 \times 120}{2298} = 82.3$	$\frac{722 \times 120}{2298} = 37.7$	120
nicht PDS	$\frac{1576 \times 2178}{2298} = 1493.7$	$\frac{722 \times 2178}{2298} = 684.3$	2178
Σ	1576	722	2298

Wie kommt man zum Wert χ^2?

- Für jede Zelle der 2×2 Tabelle Differenz zwischen beobachteten/erwarteten Werten ermitteln $\rightarrow$ einfache Abweichungen
- Summe der einfachen Abweichungen $= 0$
- Quadrieren
  - Damit das Vorzeichen verschwindet
  - Um größere Abweichungen stärker zu gewichten
- Quadrierte Abweichungen durch erwarteten Werte teilen
  - Größere erwartete Werte $\rightarrow$ größere zufällige Schwankungen
  - Quadrierte Abweichungen werden auf eine Art gemeinsame Skala gebracht
- Standardisierte quadrierte Abweichungen in jeder Zelle (Beitrag)
- Summe der Beiträge: χ^2

Wie kommt man zum Wert χ^2? II

beobachtet

	West	Ost	Σ
PDS	4	116	120
nicht PDS	1572	606	2178
Σ	1576	722	2298

erwartet

	West	Ost	Σ
PDS	82.3	37.7	120
nicht PDS	1493.7	684.3	2178
Σ	1576	722	2298

$$\chi^2 = \frac{(4-82.3)^2}{82.3} \cdots + \frac{(116-37.7)^2}{37.7} + \frac{(1572-1483.7)^2}{1493.7} + \frac{(606-684.3)^2}{684.3}$$
$$= 74.49 \cdots + 162.62 + 4.10 + 8.96 = 250.16$$

Eigenschaften von χ^2

χ^2

- ... ist eine Maßzahl
- ... hat keine Dimension
- Sein Wert hängt ab von
  1. **Der Stärke des Zusammenhangs**
  2. Der Zahl der Kategorien
  3. Der Zahl der Fälle
- Als Zusammenhangsmaß ungeeignet, aber *Basis* für eine Reihe von Zusammenhangsmaßen

Maße auf der Basis von χ^2

1. ϕ
2. Kontingenzkoeffizient C
3. **Cramer's** V

- Versuchen χ^2 zu standardisieren
- Korrektur für Zahl der Fälle:
  $\phi = \sqrt{\chi^2/n} = \sqrt{250.2/2298} = 0.33$
- Korrektur von ϕ für große Tabellen (viele Kategorien):
  $C = \sqrt{\chi^2/(\chi^2 + n)} = \sqrt{250.2/(250.2 + 2298)} = 0.31$
- Cramer's V

Wie berechnet man Cramer's V?

Cramer's V

$$
\begin{aligned}
V &= \sqrt{\frac{\chi^2}{n \times (R-1)}} \\
&= \sqrt{\frac{250.2}{2298 \times (2-1)}} \\
&= \sqrt{\frac{250.2}{2298}} \\
&= 0.33
\end{aligned}
$$

- R = Minimum der Zeilen/Spalten
- V berücksichtigt n und R (vs. C)
- Wertebereich [0;1]
- Symmetrisch
- Für 2×2 Tabellen mit ϕ identisch

Was ist ein PRE-Maß?

- Logik: Wieviel besser können wir eine abhängige Variable vorhersagen, wenn Wert der unabhängigen Variablen bekannt ist?
- (Bzw.: Wie unterscheiden sich marginale und konditionale Verteilung?)
- Wie stark reduziert sich der Vorhersagefehler – **P**roportionate **R**eduction in **E**rror
- PRE-Maß für nominalskalierte Daten: λ
- Logik von λ:
  - Ohne Kenntnis unabhängiger Variablen ist Modalkategorie beste Vorhersage $\rightarrow$ Vorhersagefehler
  - Reduziert sich Zahl der Fehler durch Kenntnis unabhängiger Variablen?

Wahlabsicht und Kanzlerpräferenz

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Σ
Union	**335**	**15**	**350**
SPD	**25**	**320**	**345**
Andere	**84**	**102**	**186**
Σ	444	437	881

- Beste Vorhersage ohne Kenntnis der Kanzlerpräferenz: Union
- **345+186=531** Vorhersagefehler
- Bei Kenntnis der Präferenz
  - Vorhersage „Union“ für Merkel-Anhänger
  - **25+84 = 109** Vorhersagefehler
  - Vorhersage „SPD“ für Steinmeier-Anhänger
  - **15+102= 117** Vorhersagefehler
- (84+25)+(15+102) = 226 Vorhersagefehler
- $\lambda = \frac{\text{Fehler1-Fehler2}}{\text{Fehler1}} = \frac{531-226}{531} = 0.57$
- Fehler reduziert sich um 57%

Mehr zu λ

- Asymmetrisches Maß; theoretische Unterscheidung zwischen abhängiger/unabhängiger Variable
- Kann Wert 0 annehmen obwohl Zusammenhang besteht
- Wenn Modalkategorie der abhängigen Variablen über Kategorien der unabhängigen Variablen hinweg gleich
- Beispiel Region $\rightarrow$ PDS-Wahl

„Versagen“ von λ

	Region		
Wahlabsicht	West	Ost	Σ
PDS	**4**	**116**	**120**
Andere	1572	606	2178
Σ	1576	772	2298

- Vorhersage gesamt: Andere $\rightarrow$ 120 Fehler
- Vorhersage West: Andere $\rightarrow$ 4 Fehler
- Vorhersage Ost: Andere $\rightarrow$ 116 Fehler
- $\lambda = \frac{120-(116+4)}{120} = 0$ (vs. $V = \phi = 0.33$)
- Anschaulich, aber nicht unproblematisch

Zusammenfassung

- Zusammenhang = Verteilungen zweier Variablen nicht unabhängig voneinander
- Kategorien treten häufiger gemeinsam auf als bei rein zufälliger Verteilung zu erwarten
- Zusammenhangs*maße* quantifizieren Stärke und ggf. Richtung des Zusammenhangs
- χ^2-basierte Maße betrachten direkt Abweichungen von zufälliger Verteilung und sind symmetrisch
- PRE-Maß λ betrachtet Verbesserung gegenüber naiver Vorhersage und ist asymmetrisch