Was sind Zusammenhangsmaße?

- Zusammenhangsmaße beschreiben einen Zusammenhang zwischen zwei Variablen
- Beispiele für Zusammenhänge:
  - 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

Was ist ein Zusammenhang?

- Ein Zusammenhang zwischen zwei Variablen besteht dann, wenn bestimmte Merkmale häufiger gemeinsam auftreten, als bei einer *zufälligen* Verteilung zu erwarten wäre
- Die Erwartung für das gemeinsame Auftreten zweier Merkmale bei zufälliger Verteilung ergibt sich nach den Regeln der Kombinatorik, indem die Wahrscheinlichkeiten für das Auftreten der einzelnen Merkmale miteinander multipliziert werden
- Die Wahrscheinlichkeit für das Auftreten eines Merkmals ist identisch mit der relativen Häufigkeit des Merkmals in der untersuchten Gruppe

Beispiel: „Überzufällig häufig“

- Der Anteil von Männern/Frauen in einer Gesellschaft (=Grundgesamtheit) beträgt 50%
- Der Anteil von Protestanten in einer Gesellschaft beträgt 60% (Katholiken 40%)
- Wenn zwischen beiden Merkmalen kein Zusammenhang besteht, hängt es nur vom Zufall ab, ob ein männlicher Bürger katholisch oder protestantisch ist
- In diesem Fall beträgt die Wahrscheinlichkeit, daß ein beliebiger Bürger männlich und katholisch ist, 0,5*0,4=0,2 und der Anteil der männlichen Katholiken in dieser Gesellschaft beträgt dementsprechend 20%

Beispiel (II)

- Wird der Wert von 20% in der Realität über- oder unterschritten, dann treten männliche Katholiken in dieser „überzufällig“ häufig bzw. selten auf
- In diesem Fall besteht ein statistischer Zusammenhang (Korrelation) zwischen den Variablen Konfession und Geschlecht
- Dieser Zusammenhang kann nicht ohne weiteres *kausal* interpretiert werden
- Die Logik des überzufälligen gemeinsamen Auftretens gilt für alle Zusammenhangsmaße, unabhängig von ihrem Datenniveau
- Zusammenhänge, die in Stichproben gemessen werden, stellen eine Schätzung für den Zusammenhang in der GG dar, die mit einem entsprechenden Schätz(=Stichproben)-Fehler behaftet sind

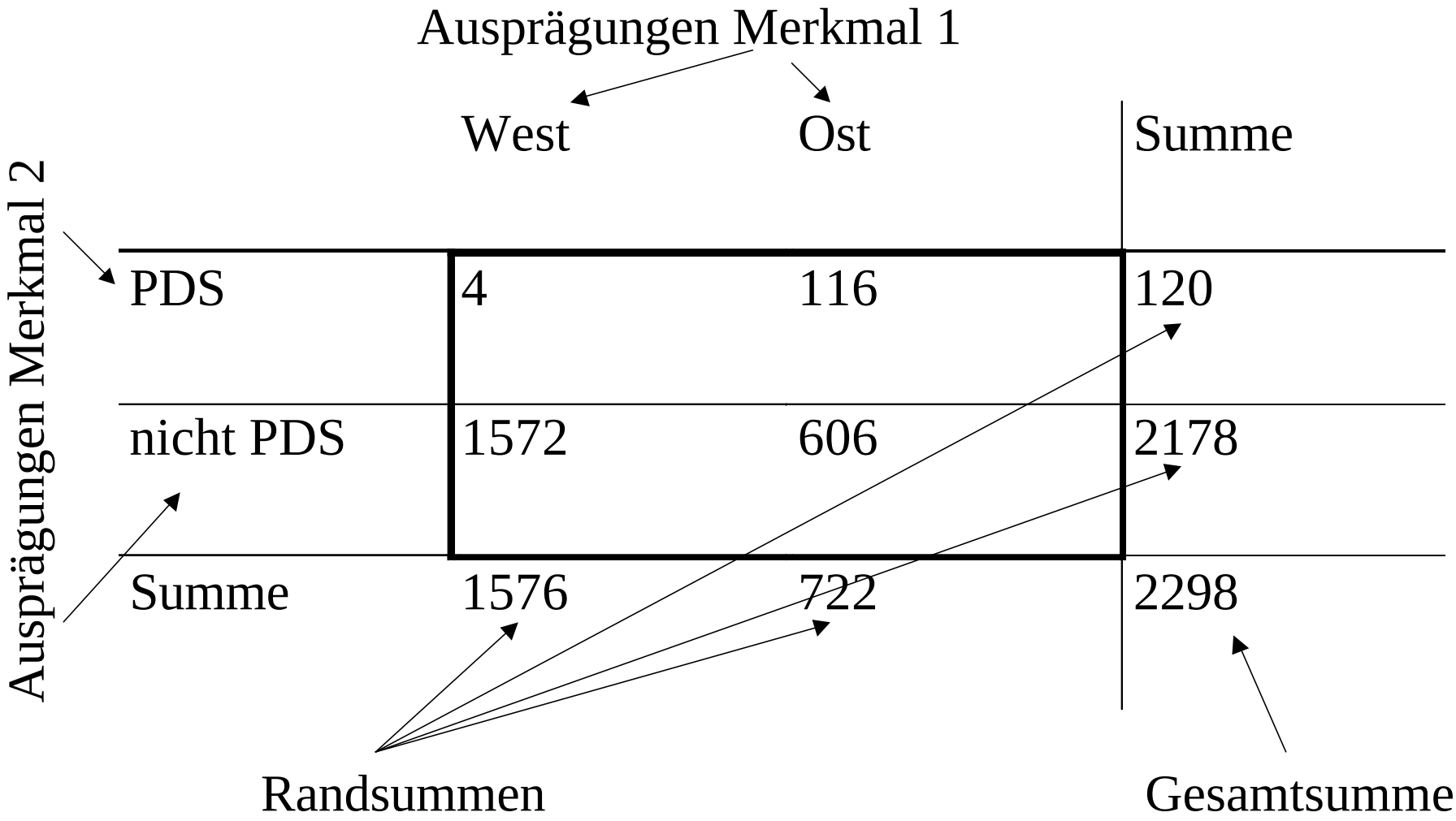
Warum Zusammenhangsmaße?

- Zusammenhangsmaße weisen einen definierten Wertebereich auf
- Mit Hilfe von Zusammenhangsmaßen kann die Stärke verschiedener Zusammenhänge miteinander verglichen werden
- Zusammenhangsmaße sollten einen Wertebereich von 0 bis 1 bzw. von -1 bis +1 aufweisen
- Die Wahl des richtigen Zusammenhangsmaßes hängt hauptsächlich vom Skalenniveau der Variablen ab

Zwei nominalskalierte Variablen: Maße auf der Basis von χ^2

- Bei Maßen auf der Basis von χ^2 ist besonders leicht zu erkennen, daß Zusammenhänge sich auf das überzufällig häufige gemeinsame Auftreten von Merkmalen beziehen
- Maße auf der Basis von χ^2 vergleichen eine empirische Kreuztabelle mit einer Tabelle, in der die Häufigkeiten eingetragen sind, die zu erwarten wäre, wenn kein Zusammenhang zwischen den Merkmalen bestünde (Indifferenztabelle)

Was ist eine Kreuztabelle?



Welche Prozentuierungsarten gibt es in einer 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?
  - Wieviel Prozent der Wähler insgesamt stimmen für die PDS
- *Totalprozente* setzen den Inhalt einer Zelle zur Gesamtsumme ins Verhältnis
  - Wie hoch ist der Anteil der westdeutschen PDS-Wähler an allen Wählern?

Wie wird die Indifferenztabelle konstruiert?

- Die Wahrscheinlichkeit für das Auftreten der Ausprägungen des ersten Merkmals wird bestimmt, indem in der Summenzeile am unteren Rand der Tabelle die Zeilenprozente bestimmt werden
- Die Wahrscheinlichkeiten für das zweite Merkmal werden analog durch die Berechnung der Spaltenprozente in der Summenspalte ermittelt

Zeilenprozente

	West	Ost	Summe
PDS	4/120 =3,3%	116/120 =96,6%	120 =100%
nicht PDS	1572/2178= =72,1%	606/2178 =27,8%	2178 =100%
Summe	1576/2298 =68,6%	722/2298 =31,4%	2298 =100%

Konstruktion der Indifferenztabelle II

- Die Multiplikation beider Wahrscheinlichkeiten ergibt die Gesamtwahrscheinlichkeit für die Kombination beider Merkmale bei Unabhängigkeit (Indifferenz) der Variablen. Multipliziert man diese mit der Gesamtzahl der Fälle, erhält man die erwartete absolute Häufigkeit für diese Kombination
- In der Praxis läßt sich die Formel durch kürzen vereinfachen:

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

- Diese Berechnung wird für alle Zellen vorgenommen
- Randsummen und Gesamtsumme ändern sich nicht!

Indifferenztabelle

	West	Ost	
PDS	82,3 =(120*1576)/2298	37,7	120
nicht PDS	1493,7	684,3	2178
	1576	722	2298

Berechnung von χ^2

- Für jede Zelle wird die Differenz zwischen beobachtetem und erwartetem Wert ermittelt (einfache Abweichung)
- Diese Werte werden quadriert
  - damit das Vorzeichen verschwindet (die einfachen Abweichungen addieren sich zu 0)
  - um größere Abweichungen stärker zu gewichten
- Und durch den erwarteten Wert geteilt
  - bei größeren erwarteten Werten ist mit größeren Schwankungen zu rechnen
  - quadrierte Abweichungen werden so auf eine Art gemeinsame Skala gebracht
- Diese Wert werden aufsummiert. Das Ergebnis ist die Maßzahl χ^2

Rechenbeispiel

$$c^2 = \frac{(4-82{,}3)^2}{82{,}3} + \frac{(116-37{,}7)^2}{37{,}7} + \frac{(1572-1493{,}7)^2}{1493{,}7} + \frac{(606-684{,}3)^2}{684{,}3}$$

$$= 74{,}49 + 162{,}62 + 4{,}10 + 8{,}96 = 250{,}16$$

- χ^2 hat keine Dimension
- Es kann Werte zwischen 0 und $+\infty$ annehmen
- Sein Wert von der Stärke des Zusammenhanges, der Zahl der Kategorien und der Zahl der Fälle beeinflußt
- χ^2 ist deshalb als Zusammenhangsmaß ungeeignet, dient aber als Basis für eine Reihe von Koeffizienten

Cramer‘s V

- Koeffizient C berücksichtigt nur die Zahl der Fälle ➔maximaler Wert meist < 1
- V berücksichtigt die Zahl der Fälle *und* die Zahl der Kategorien
- Kann für alle Tabellen Werte zwischen 0 und 1 annehmen
- Das „R“ in der Formel bezeichnet das Minimum der Anzahl von Zeilen/Spalten. D.h., R ist gleich der Zahl der Kategorien derjenigen Variablen, die weniger Ausprägungen aufweist
- V ist ein symmetrisches Maß

$$V = \sqrt{\frac{c^2}{n \times (R-1)}} = \sqrt{\frac{250{,}16}{2298 \times (2-1)}} = \sqrt{\frac{250{,}16}{2298}} = 0{,}33$$

PRE-Maße

- Drücken aus, um wieviel besser die Ausprägung einer abhängigen Variable für einen beliebigen Befragten *vorhergesagt* werden kann, wenn eine unabhängige Variable bekannt ist, d.h., um wieviel Prozent sich der Vorhersagefehler reduziert (*P*roportional *R*eduction of *E*rror)
- Ein PRE-Maß für nominalskalierte Variablen ist λ
- Die Logik von λ basiert darauf, daß wir ohne Kenntnis einer anderen Variablen das Auftreten der am häufigsten besetzte Kategorie (Modalkategorie) prognostizieren würden

λ

	Kanzlerpräferenz		
Wabs	Kohl	Scharping	Summe
Union	335	15	350
SPD	25	320	345
Andere	84	102	186
Summe	444	437	881

- Ohne Kenntnis der Kanzlerpräferenz würde man für die Wahlabsicht ein Votum zugunsten der Union vorhersagen und sich in 345+186=531 Fällen irren

λ II

- Wenn man die Kanzlerpräferenz kennt, würde man für Anhänger Kohls die Wahl der Union und für Scharping-Fans die Wahl der SPD (jeweils häufigste Kategorie) vorhersagen
- Dabei macht man in (84+25)+(15+102)=226 Fällen einen Fehler

$$l = \frac{(Fehler_1 - Fehler_2)}{Fehler_1} = \frac{531 - 226}{531} = 0{,}57$$

- Der Vorhersagefehler für die Wahlabsicht reduziert sich um 57%

λ III

- λ ist ein asymmetrisches Maß, d.h., es wird vorab theoretisch festgelegt, welches die abhängige Variable ist
- λ kann unter Umständen auch dann den Wert 0 annehmen, wenn ein Zusammenhang besteht
- Das ist dann der Fall, wenn die Modalkategorie der abhängigen Variablen über die Kategorien der unabhängigen Variablen hinweg gleich ist - wie im Beispiel mit dem Wahlverhalten zugunsten der PDS als abhängiger und der Region als unabhängiger Variablen

„Versagen“ von λ

	Region		
Wabs	West	Ost	Summe
PDS	4	116	120
andere	1572	606	2178
Summe	1576	722	2298

- $\text{Fehler}_1 = 120$
- $\text{Fehler}_2 = 4+116=120$
- $\lambda = 0$

Zwei ordinalskalierte Merkmale: γ

- Kann ebenfalls als PRE-Maß interpretiert werden
- Ist ein symmetrisches Maß, d.h. der Wert des Koeffizienten hängt nicht davon ab, welches die abhängige und welches die unabhängige Variable ist
- Zusammenhang zwischen zwei ordinalen Variablen bedeutet:
  - „mehr“ von der einen Variablen - „mehr“ von der anderen Variablen (positiver Zusammenhang)
  - „mehr“ von der einen Variablen - „weniger“ von der anderen Variablen (negativer Zusammenhang)
- γ beruht auf der Logik des Paarvergleiches

Die Idee des Paarvergleichs

- Für die Berechnung von γ werden Paare von Befragten gebildet
- Konkordantes Paar aus den Befragten A und B: Befragter B weist für beide Variablen (z.B. Bildung und politisches Interesse) höhere Werte auf als Befragter A
- Diskonkordantes Paar aus den Befragten A und B: Befragter B weist für eine Variable einen höheren Wert und für die andere Variable einen niedrigeren Wert auf als Befragter A
- Verhältnis konkordante und diskonkordante Paare:
  - Konkordante Paare überwiegen: positiver Zusammenhang
  - Diskonkordante Paare überwiegen: negativer Zusammenhang
- Paare, die für eine oder beide Variablen identische Werte aufweisen (ties) werden bei der Berechnung von γ nicht berücksichtigt

Berechnung der Anzahl der Paare

- Berechnung der entsprechenden Kreuztabelle.
- Anordnung der Kategorien: niedrigste oben/links, höchste unten/rechts
- Konkordante Paare können die Befragten mit allen Befragten in den Zellen rechts und unterhalb der eigenen Zelle bilden
- Diskonkordante Paare können die Befragten mit allen Befragten in den Zellen links und unterhalb der eigenen Zelle bilden
- Die Zahl der möglichen Paare für zwei Zellen ist das Produkt der Häufigkeiten beider Zellen
- Auf diese Weise werden alle konkordanten und dann alle diskonkordanten Paare zusammenaddiert

γ: Bildung und Nationalstolz

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

Zahl der Paare

NC=	115*(267+209+463+260+138+58)=	160425
	+128*(209+260+58)=	67456
	+267*(463+260+138+58)=	245373
	+267*(260+58)=	84906
	+731*(138+58)=	143276
	+463*58=	26854
	=	**728290**
ND=	128*(267+731+416)=	180992
	181*(267+267+731+463+416+138)=	413042
	267*(731+416)=	306249
	209*(731+463+416+138)=	365332
	463*416=	192608
	260*(416+138)=	144040
		1602263

Ergebnis für das Beispiel

$$g = \frac{N_c - N_d}{N_c + N_d}$$

$$= \frac{728.290 - 1.602.263}{728.290 + 1.602.263}$$

$$= -0{,}375$$