Statistik I - Sitzung 6

Bernd Schlipphak

Institut für Politikwissenschaft

Sitzung 6

Statistik I - Sitzung 6

1. Konzeption eines Kausalmodells

2. Bivariate Zusammenhangsmaße
   - Grundlegende Einführung
   - Die Kreuztabelle - Grundlegendes
   - Die Kreuztabelle als Indikator für den Zusammenhang

3. Zusammenhangsmaße für nominal skalierte Variablen
   - Prozentsatzdifferenz
   - Odds Ratio
   - Chi^2

Grundgedanke Kausalmodell

- In einem Kausalmodell (lat. *causa* = Ursache, Grund) gehen wir davon aus, dass die Ausprägung, die ein Fall auf der unabhängigen Variable (= X) einnimmt, die *Ursache* für die Ausprägung ist, die der Fall auf der abhängigen Variablen (=Y) einnimmt (= *Wirkung*)
- Daher sprechen wir auch von der *Kausalität* zwischen X und Y

Variablen in einem Kausalmodell

- **Abhängige** Variable ($= Y$, AV)
- **Unabhängige** Variable ($= X$, UV)
- **Intervenierende** Variable ($= Z$, $IntV$)

Grundgedanke Kausalmodell

- In einem Kausalmodell (lat. *causa* = Ursache, Grund) gehen wir davon aus, dass die Ausprägung, die ein Fall auf der unabhängigen Variable (= X) einnimmt, die *Ursache* für die Ausprägung ist, die der Fall auf der abhängigen Variablen (=Y) einnimmt (= *Wirkung*)
- Daher sprechen wir auch von der *Kausalität* zwischen X und Y

Bedingungen für Kausalität

- Eine solche kausale Ursache-Wirkungs-Beziehung zwischen X und Y liegt nach Diaz-Bone (2006: 64) jedoch nur dann vor, wenn
  1. Die Ursache X der Wirkung Y zeitlich vorangeht
  2. Der Zusammenhang zwischen beiden Variablen statistisch belegbar ist
  3. Der Zusammenhang NICHT durch Einfluss einer anderen, dritten Variable zustande gekommen ist
  4. Eine theoretische Erklärung für die Wirkung von X auf Y vorliegt

Bedingungen für Kausalität

- Eine solche kausale Ursache-Wirkungs-Beziehung zwischen X und Y liegt nach Diaz-Bone (2006: 64) jedoch nur dann vor, wenn
  1. Die Ursache X der Wirkung Y zeitlich vorangeht
  2. Der Zusammenhang zwischen beiden Variablen statistisch belegbar ist
  3. Der Zusammenhang NICHT durch Einfluss einer anderen, dritten Variable zustande gekommen ist
  4. Eine theoretische Erklärung für die Wirkung von X auf Y vorliegt

Bedingungen für Kausalität

- Statistik allein kann Kausalität nicht analysieren – theoretische Vorarbeit ist immer notwendig, bevor statistische Berechnungen ins Spiel kommen!
- Statistische Berechnungen – etwa zum Zusammenhang zwischen zwei Variablen – geben per se nur Auskunft über einen ungerichteten Zusammenhang

Bedingungen für Kausalität

- Vierter Punkt sollte stets der erste sein, den Forschende berücksichtigen
- Wichtigster Punkt in einem deduktiven Forschungsdesign ist also die überzeugende Herleitung theoretischer Erwartungen

Merksatz zu Kausalität und Statistik:

Nur nach einer **überzeugenden theoretischen Argumentation** ist die statistische Überprüfung eines gerichteten Zusammenhangs sinnvoll!

Einführung in die Zusammenhangsmaße

- In dieser und den nächsten beiden Sitzungen werden wir **bivariate Zusammenhangsmaße** - d.h., Maße für den Zusammenhang zweier Variablen - kennenlernen
- Diese Zusammenhangsmaße unterscheiden sich vor allem danach, auf welchen Skalenniveaus sie basieren
- Außerdem demonstrieren diese Zusammenhangsmaße - sofern nicht anders vermerkt - immer zunächst **ungerichtete Zusammenhänge**, d.h. Zusammenhänge, in denen wir nicht vorab zwischen unabhängiger und abhängiger Variable unterscheiden

Die Kreuztabelle

- Ganz grundlegend lässt sich der Zusammenhang zwischen zwei Variablen in einer **Kontingenz- oder Kreuztabelle** darstellen
- In einer solchen Kreuztabelle tragen wir die absoluten oder relativen Häufigkeiten für beide Variablen ab
- Die Ausprägungen der beiden Variablen werden dann jeweils den Zeilen / Reihen (engl. rows) bzw. den Spalten (engl. columns) zugeordnet

Die Kreuztabelle

- Möchte man aus einer Kreuztabelle Aussagen zu einem **gerichteten** Zusammenhang machen, so gilt die folgende, aus dem angloamerikanischen Wissenschaftskontext übernommene Konvention
  - Zeilen / Reihen = Ausprägungen der abhängigen Variable (= Y)
  - Spalten = Ausprägungen der unabhängigen Variable (= X)
- Diese Konvention spielt dann vor allem für die später noch vorzustellende Prozentuierung der Zellen eine Rolle

Die Kreuztabelle

- Die Ausprägungen von X werden mit j indiziert, wobei der Index von $j=1,\ldots,s$ läuft und wobei s die Anzahl der Ausprägungen von X (und damit die Anzahl der Spalten) definiert
- Die Ausprägungen von Y werden mit i indiziert, wobei der Index von $i=1,\ldots,r$ läuft und wobei r die Anzahl der Ausprägungen von Y (und damit die Anzahl der Zeilen / Reihen) definiert

Die Kreuztabelle

- Damit stellt jede Zelle in einer Kreuztabelle eine Kombination jeweils einer Ausprägung von X und Y dar
- Inhalt der Zellen ist jeweils die Häufigkeit, mit der die Kombination dieser Ausprägungen in der Verteilung vorkommt

Die Kreuztabelle

- Hypothetisches, aber grundlegend realitätsnahes Beispiel: Der Zusammenhang zwischen Vertrauen in internationale Organisationen und Bildung

	Hohe Bildung	**Niedrige Bildung**
Vertrauen in internationale Organisationen	10	3
Kein Vertrauen in internationale Organisationen	5	7

- Die Kombination der Ausprägungen „Hohe Bildung“ und „Vertrauen in IO“ kommt der Tabelle zufolge 10mal vor. Anders formuliert: 10 Fälle weisen die Kombination „Hohe Bildung“ und „Vertrauen in IO“ auf.

Die Kreuztabelle

- Hypothetisches, aber grundlegend realitätsnahes Beispiel: Der Zusammenhang zwischen Vertrauen in internationale Organisationen und Bildung

	Hohe Bildung	**Niedrige Bildung**
Vertrauen in internationale Organisationen	10	3
Kein Vertrauen in internationale Organisationen	5	7

- In diesem Beispiel wären darüber hinaus s (= Ausprägungen X) = 2 und r (= Ausprägungen Y) = 2
- **Größe** oder **Format** der Kreuztabelle = r x s

Die Kreuztabelle

- Weiter unterscheidet man **Reihensummen** (= Summierung der Häufigkeiten der Reihen) und **Spaltensummen** (= Summierung der Häufigkeiten der Spalten)
- Beide zusammen werden **Randsummen** genannt.
- Spaltensummen oder Reihensummen ergeben addiert jeweils N, d.h. die Fallzahl der Verteilung

	Hohe Bildung	Niedrige Bildung	$\sum$ **Reihensumme**
Vertrauen in IO	10	3	**13**
Kein Vertrauen in IO	5	7	**12**
$\sum$ **Spaltensumme**	**15**	**10**	**25**

Die Kreuztabelle

- Reihensummen / Spaltensummen stellen die **Randverteilungen / marginale Verteilungen** von X und Y dar
  - Die Spaltensummen stellen die univariate Verteilung von X dar
  - Die Reihensummen stellen die univariate Verteilung von Y dar
- Man nennt diese Verteilungen auch **unbedingte** Verteilungen, weil sie *nicht* von der jeweils anderen Variable beeinflusst sind

Die Kreuztabelle

- In den Zellen der Kreuztabelle liegen demgegenüber die **bedingten** Verteilungen vor
- Man sagt dann auch, dass in jeder Zelle einer Zeile die Häufigkeit einer Ausprägung von Y (= der Zeile) unter der Bedingung steht, dass X einen bestimmten Wert (= der Spalte) annimmt
- Das Auftreten der Y-Werte „Vertrauen in IO“ und „Kein Vertrauen in IO“ wird also bedingt durch die Ausprägungen der Variable X (= Bildung).

Die Kreuztabelle

- Ob es tatsächlich einen **statistischen** Zusammenhang zwischen X und Y gibt, lässt sich anhand der absoluten Häufigkeiten nur schwer ablesen
- Daher arbeitet man in Kreuztabellen meist mit **relativen Häufigkeiten** (in Prozent)
- Für gerichtete Zusammenhänge interessieren dabei vor allem die relativen Häufigkeiten der Ausprägungen von Y unter den verschiedenen Bedingungen der Ausprägungen von X. Diese relativen Häufigkeiten erhält man durch die **Spaltenprozentuierung**

Die Kreuztabelle

- Spaltenprozentuierung = absolute Häufigkeit einer Ausprägung von Y unter einer X-Ausprägung durch Summe der Häufigkeiten dieser X-Ausprägung (= Spaltensumme der X-Ausprägung) * 100

	Hohe Bildung	Niedrige Bildung	$\sum$ **Reihensumme**
Vertrauen in IO	66.7% (= (10/15)*100)	30% (= (3/10)*100)	**52% = 13**
Kein Vertrauen in IO	33.3% (= (5/15)*100)	70% (= (7/10)*100)	**48% = 12**
$\sum$ **Spaltensumme**	**100% = 15**	**100% = 10**	**100% = 25 = N**

Die Kreuztabelle

- Spaltenprozentuierung wird auch für Reihensummen durchgeführt
- Prozentuierte Reihensummen als marginale/unbedingte Verteilung für Y, mit der die bedingten Y-Verteilungen für die X-Ausprägungen verglichen werden

	Hohe Bildung	Niedrige Bildung	$\sum$ **Reihensumme**
Vertrauen in IO	66.7% (= (10/15)*100)	30% (= (3/10)*100)	**52% = 13**
Kein Vertrauen in IO	33.3% (= (5/15)*100)	70% (= (7/10)*100)	**48% = 12**
$\sum$ **Spaltensumme**	**100% = 15**	**100% = 10**	**100% = 25 = N**

Die Kreuztabelle

- Sind die prozentuierten bedingten Verteilungen in jeder Spalte genau gleich der prozentuierten Reihensumme (d.h., der unbedingten relativen Verteilung von Y), dann besteht kein Zusammenhang zwischen X und Y
- Man spricht dann auch davon, dass Y von X **statistisch unabhängig** ist
- In unserem Beispiel ist dies aber nicht der Fall: Das Vertrauen in IO scheint damit von der Bildung statistisch abhängig zu sein

Ausblick

- Mit der Kreuztabelle haben wir einen ersten Überblick über die Darstellung bivariater Zusammenhänger erhalten
- Durch die Darstellung relativer Häufigkeiten (= prozentuierter Verteilungen) können wir in Kreuztabellen einen ersten Einblick in die Abhängigkeit oder Unabhängigkeit zweier Variablen voneinander gewinnen
- Über die Größe des Zusammenhangs - ganz zu schweigen von der statistischen Signifikanz ($\Rightarrow$ *Statistik II*) - können wir jedoch noch keine Aussage treffen.

Kurze Einführung

- Die Zusammenhangsmaße für nominal skalierte Variablen können für jedes Skalenniveau eingesetzt werden
- Die drei grundsätzlichen Maße sind dabei die Prozentsatzdifferenz, Odds Ratio und Chi^2
- Für die weiterführende Statistik sind vor allem die Maße Odds Ratio ($\Rightarrow$ Logistische Regression) und Chi^2 ($\Rightarrow$ Hypothesentest) wichtig.

Die Prozentsatzdifferenz (PD)

- Grundlage der Berechnung ist die Kreuztabelle
- Angewendet werden kann die PD jedoch nur auf die Kreuztabelle mit zwei Variablen mit zwei Ausprägungen (= Vierfeld-Tabelle)
- Für größere Kreuztabellen ist die PD nicht anwendbar

Die Prozentsatzdifferenz (PD)

- Die PD - formal = d% - erfasst, wie stark die Spaltenprozente voneinander abweichen
- Die Maßeinheit sind daher PP = Prozentpunkte
- Grundsätzlich kann d% damit Werte zwischen -100 und +100 PP annehmen
- Es gilt: Wenn d% = 0, sind die beiden Variablen statistisch unabhängig voneinander

Die Prozentsatzdifferenz (PD)

- Die Berechnung der Prozentsatzdifferenz erfolgt über die Formel $\text{d}\% = y_{1|x_1} - y_{1|x_2}$ oder $\text{d}\% = y_{2|x_1} - y_{2|x_2}$
- Dabei ist $y_{1|x_1}$ die (relative) Häufigkeit der ersten Ausprägung von Y für die erste Ausprägung von X, $y_{1|x_2}$ ist die Häufigkeit der ersten Ausprägung von Y für die zweite Ausprägung von X etc.

	Hohe Bildung (x_1)	Niedrige Bildung (x_2)	$\sum$ **Reihensumme**
Vertrauen in IO	$y_{1\|x_1}$	$y_{1\|x_2}$	y_1
Kein Vertrauen in IO	$y_{2\|x_1}$	$y_{2\|x_2}$	y_2

Die Prozentsatzdifferenz am Beispiel

	Hohe Bildung	Niedrige Bildung	$\sum$ **Reihensumme**
Vertrauen in IO	66.7% $(= y_{1\|x_1})$	30% $(= y_{1\|x_2})$	**52**%
Kein Vertrauen in IO	33.3% $(= y_{2\|x_1})$	70% $(= y_{2\|x_2})$	**48**%
$\sum$ **Spaltensumme**	**100**%	**100**%	**100**%

- $\text{d}\% = y_{1|x_1} - y_{1|x_2} = 66.7\% - 30\% = 36.7$ PP
- $\text{d}\% = y_{2|x_1} - y_{2|x_2} = 33.3\% - 70\% = -36.7$ PP

Die Prozentsatzdifferenz am Beispiel

	Hohe Bildung	Niedrige Bildung	$\sum$ **Reihensumme**
Vertrauen in IO	66.7% $(= y_{1\|x_1})$	30% $(= y_{1\|x_2})$	**52**%
Kein Vertrauen in IO	33.3% $(= y_{2\|x_1})$	70% $(= y_{2\|x_2})$	**48**%
$\sum$ **Spaltensumme**	**100**%	**100**%	**100**%

- Die Prozentsatzdifferenz bestätigt, was der erste Eindruck schon vermuten lässt. Die beiden Variablen sind statistisch NICHT unabhängig voneinander. D.h., sie hängen miteinander zusammen.
- Allerdings sagt die PD auch noch nicht viel über die Stärke oder das Ausmaß des Zusammenhangs zwischen beiden Variablen aus.

Odds Ratio (OR)

- Grundlegend misst die Odds Ratio Ähnliches wie die Prozentsatzdifferenz, die Berechnung und die Interpretation sind jedoch etwas komplizierter
- OR misst nicht das Verhältnis zwischen Prozenten, sondern zwischen Odds (= Gewinnchancen oder einfach 'Chancen')
- Odds Ratio wird daher manchmal auch als **Verhältnis der Verhältnisse** oder als **Kreuzproduktverhältnis** bezeichnet

Odds Ratio (OR)

- Das erste Verhältnis bezeichnet dabei die Odds, d.h. das Verhältnis zwischen den Ausprägungen der einen Variable (=Y)
  - Odds = Wie oft tritt y_1 im Vergleich zu y_2 auf?
- Das zweite Verhältnis bezeichnet dann das jeweilige Verhältnis der Odds für die beiden Ausprägungen von X
  - Odds Ratio = Wie oft tritt y_1 im Vergleich zu y_2 unter der Ausprägung x_1 im Verhältnis zu x_2 auf?
- An der Formel wird bereits deutlich, dass auch Odds Ratio nur für Vierfeldtabellen anwendbar ist. Für größere Kreuztabellen kann OR nicht berechnet werden!

Odds Ratio (OR) - Formalisierung

	Hohe Bildung ($=x_1$)	Niedrige Bildung($=x_2$)	$\sum$
Vertrauen in IO ($=y_1$)	f_{11}	f_{12}	$f_{1.}$
Kein Vertrauen in IO ($=y_2$)	f_{21}	f_{22}	$f_{2.}$
$\sum$	$f_{.1}$	$f_{.2}$	

- f_{11} = Häufigkeit von y_1 unter Bedingung von x_1
- $Odds_{y_1|x_1} = \frac{f_{11}}{f_{21}}$ und $Odds_{y_1|x_2} = \frac{f_{12}}{f_{22}}$
- Odds Ratio (OR) ist dann das Verhältnis dieser beiden Odds zueinander
  - $OR = \frac{Odds_{y_1|x_1}}{Odds_{y_1|x_2}}$

Odds Ratio (OR) - Interpretation

- OR $= 1 \Rightarrow$ Variablen statistisch unabhängig voneinander
- OR $< 1 \Rightarrow Y_1$ tritt im Verhältnis zu Y_2 unter der Bedingung X_1 **seltener** auf als unter der Bedingung X_2
- OR $> 1 \Rightarrow Y_1$ tritt im Verhältnis zu Y_2 unter der Bedingung X_1 **häufiger** auf als unter der Bedingung X_2
- OR < 1 oder OR $> 1 \Rightarrow$ Beide Variablen sind statistisch NICHT unabhängig voneinander, sondern hängen zusammen

Odds Ratio (OR) - Neues Beispiel

- Hängen die Einstellungen zu liberaler Marktwirtschaft (=X) mit der Wahl der FDP (=Y) zusammen?

	Liberale MW ja ($=x_1$)	Liberale MW nein ($=x_2$)
FDP-Wahl ($=y_1$)	20 ($=f_{11}$)	10 ($=f_{12}$)
Keine FDP-Wahl ($=y_2$)	5 ($=f_{21}$)	40 ($=f_{22}$)

- $Odds_{y_1|x_1} = \frac{f_{11}}{f_{21}} = \frac{20}{5} = 4$
- $Odds_{y_1|x_2} = \frac{f_{12}}{f_{22}} = \frac{10}{40} \approx 0.3$
- $OR = \frac{Odds_{y_1|x_1}}{Odds_{y_1|x_2}} = \frac{4}{0.3} \approx 13.3$

Odds Ratio (OR) - Neues Beispiel

- Hängen die Einstellungen zu liberaler Marktwirtschaft (=X) mit der Wahl der FDP (=Y) zusammen?
- $OR = 13.3 \Rightarrow OR > 1$ Die Chance, die FDP zu wählen ($=y_1$) im Verhältnis die FDP nicht zu wählen ($=y_2$) ist angesichts einer positiven Einstellung zur Marktwirtschaft ($= x_1$) 13.3 mal höher als im Falle einer negativen Einstellung ($=x_2$)
- Vereinfacht: Die Wahrscheinlichkeit der FDP-Wahl steigt, wenn jemand eine positive Einstellung zur Marktwirtschaft hat!

Odds Ratio (OR) und Yule's Q

- Hängen die Einstellungen zu liberaler Marktwirtschaft (=X) mit der Wahl der FDP (=Y) zusammen?
- Ja, die beiden Variablen hängen zusammen. Aber wie **stark** ist der Zusammenhang?
- Normierung des Zusammenhangs über **Yule's Q**: $Q = \frac{OR - 1}{OR + 1}$
- **Yule's Q** gibt dann den Zusammenhang mit Werten zwischen ± 1 wieder
- Je stärker sich Yule's Q ± 1 annähert, desto stärker der Zusammenhang

Odds Ratio (OR) und Yule's Q

- Hängen die Einstellungen zu liberaler Marktwirtschaft (=X) mit der Wahl der FDP (=Y) zusammen?
- Am FDP-Beispiel:Yule's $Q = \frac{OR-1}{OR+1} = \frac{12.3}{14.3} = 0.86$
- Am FDP-Beispiel:Yule's $Q \approx 0.9 \Rightarrow$ starker Zusammenhang!

Chi^2-Koeffizient (χ^2)

- Grundlegend erfasst der Chi^2-Koeffizient (vereinfacht: Chi^2) die Stärke eines *ungerichteten* Zusammenhangs zwischen zwei nominal skalierten Variablen
- Berücksichtigt man bei der Interpretation die Richtung des Zusammenhangs, so lässt er sich auch für die Überprüfung eines gerichteten Zusammenhangs einsetzen
- Chi^2 kann auch für größere Kreuztabellen (mit r oder s > 2) berechnet werden

Chi^2-Koeffizient (χ^2)

- Chi^2 vergleicht die **beobachteten Häufigkeiten** mit den durch die Randverteilungen einer Variablen **vorhergesagten Häufigkeiten**
  - Die beobachteten Häufigkeiten bezeichnen wir mit f_{ij}
  - Die erwarteten Häufigkeiten bezeichnen wir mit e_{ij} (= erwartet / expected)
- Die erwarteten Häufigkeiten stellen die Häufigkeiten für die Zellen dar, die auftreten würden, wenn beide Variablen *unabhängig* voneinander wären $\Rightarrow$ Indifferenz-Tabelle
- Berechnung der erwarteten Häufigkeiten: $e_{ij} = \frac{f_{i.} * f_{.j}}{n}$

Chi^2-Koeffizient (χ^2)

	x_1	x_2	x_s	$\sum$
y_1	f_{11} / $e_{11} = \frac{f_{1.} * f_{.1}}{n}$	f_{12} / e_{12}	f_{1s} / e_{1s}	$f_{1.}$
y_2	f_{21} / e_{21}	f_{22} / e_{22}	f_{2s} / e_{2s}	$f_{2.}$
y_r	f_{r1} / e_{r1}	f_{r2} / e_{r2}	f_{rs} / e_{rs}	$f_{r.}$
$\sum$	$f_{.1}$	$f_{.2}$	$f_{.s}$	**N**

- Wir erhalten also genau zwei gleich große Tabellen - eine für die beobachteten Häufigkeiten und eine für die erwarteten Häufigkeiten

Chi^2-Koeffizient (χ^2)

- Chi^2
  - bestimmt nun für jede Zelle die Differenz zwischen beobachteter und erwarteter Häufigkeit $\Rightarrow (f_{ij} - e_{ij})$
  - quadriert diese Differenz $\Rightarrow (f_{ij} - e_{ij})^2$
  - und teilt diese quadrierte Differenz durch die erwartete Häufigkeit $\Rightarrow \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
- Anschließend werden die Resultate für alle Zellen zusammengezählt $\Rightarrow \sum_{Zellen} =$ Summe aller Zellen
- $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \sum_{Zellen} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

Chi^2-Koeffizient (χ^2) - Interpretation

- $\chi^2 = 0$, wenn erwartete und beobachtete Häufigkeiten gleich sind $\Rightarrow$ Variablen statistisch unabhängig = kein Zusammenhang zwischen beiden Variablen
- Je größer χ^2, desto stärker der Zusammenhang. Dazu aber später mehr!

Chi^2-Koeffizient (χ^2) - Interpretation

- Hypothetische Beispielumfrage - Sollten die Sanktionen gegen Russland verschärft werden?

	West	**Ost**	$\sum = f_{i.}$
Sehr dafür	400	100	500
Eher dafür	100	40	140
Eher dagegen	200	160	360
Sehr dagegen	300	200	500
$\sum = f_{.j}$	1000	500	1500 (= N)

Chi^2-Koeffizient (χ^2) - Interpretation

- Wir berechnen nun die erwarteten Häufigkeiten aufgrund der Formel und tragen sie in roter Farbe in die Tabelle ein

	West	**Ost**	$\sum = f_{i.}$
Sehr dafür	400 / 333.3 ($= \frac{f_{sehrdafuer} * f_{West}}{1500}$)	100	500
Eher dafür	100	40	140
Eher dagegen	200	160	360
Sehr dagegen	300	200	500
$\sum = f_{.j}$	1000	500	1500 (= N)

Chi^2-Koeffizient (χ^2) - Interpretation

- Wir berechnen nun die erwarteten Häufigkeiten aufgrund der Formel und tragen sie in roter Farbe in die Tabelle ein

	West	**Ost**	$\sum = f_{i.}$
Sehr dafür	400 / 333.3	100 / 166.7	500
Eher dafür	100 / 93.3	40 / 46.7	140
Eher dagegen	200 / 240	160 / 120	360
Sehr dagegen	300 / 333.3	200 / 166.7	500
$\sum = f_{.j}$	1000	500	1500 (= N)

Chi^2-Koeffizient (χ^2) - Interpretation

- Dann berechnen wir die quadrierten Differenzen für jede Zelle und teilen sie durch die erwarteten Häufigkeiten. Dazu tragen wir die Daten aus der Originaltabelle in eine andere Übersichtstabelle ein

i Wert von Y = Einstellung	**j Wert von X = West/Ost**	f_{ij}	e_{ij}	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
Sehr dafür	West	400	333.3	4448.9	13.3
Sehr dafür	Ost				
Eher dafür	West				
Eher dafür	Ost				
Eher dagegen	West				
Eher dagegen	Ost				
Sehr dagegen	West				
Sehr dagegen	Ost				
					.

Chi^2-Koeffizient (χ^2) - Interpretation

- Dann berechnen wir die quadrierten Differenzen für jede Zelle und teilen sie durch die erwarteten Häufigkeiten. Dazu tragen wir die Daten aus der Originaltabelle in eine andere Übersichtstabelle ein

i Wert von Y = Einstellung	**j Wert von X = West/Ost**	f_{ij}	e_{ij}	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
Sehr dafür	West	400	333.3	4448.9	13.3
Sehr dafür	Ost	100	166.7	4448.9	26.7
Eher dafür	West	100	93.3	44.9	0.5
Eher dafür	Ost	40	46.7	44.9	1.0
Eher dagegen	West	200	240	1600	6.7
Eher dagegen	Ost	160	120	1600	13.3
Sehr dagegen	West	300	333.3	1108.9	3.3
Sehr dagegen	Ost	200	166.7	1108.9	6.7
	$\sum$	1500	1500		**71.5**

Chi^2-Koeffizient (χ^2) - Interpretation

- $\chi^2 = 71.5 \Rightarrow$ Die Variablen hängen also zusammen bzw. sind NICHT statistisch unabhängig voneinander
- Damit scheint die Herkunft einer/s Befragten und seine/ihre Einstellung zur Verschärfung von Sanktionen zusammen zu hängen
- ACHTUNG - Dies ist ein hypothetisches (und zugegebenermaßen polemisches) Beispiel!
- ACHTUNG - Selbst wenn die Ergebnisse korrekt wären, müssten wir zusätzlich noch für Drittvariablen kontrollieren (Alter, Geschlecht, etc.)!

Chi^2-Koeffizient (χ^2) - Interpretation

- Da χ^2 Werte zwischen 0 und ∞ aufweisen kann, kann es bei großen Fallzahlen zu Werten kommen, die nicht mehr sinnvoll interpretierbar sind
- Auch für χ^2 existiert daher ein Normierungsmaß: **Cramers V**, das χ^2 auf einen Werteraum zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang) normiert
- Cramers $V = \sqrt{\frac{\chi^2}{N * (q-1)}}$
- q ist definiert als Minimum der Anzahl an Reihen und Spalten: $q = min(r, s)$. Wenn r (= Anzahl Reihen) $<$ s (= Anzahl Spalten) ist, verwendet man die Anzahl der Reihen und umgekehrt.

Chi^2-Koeffizient (χ^2) - Interpretation

- In unserem Beispiel ist r $= 4$ und s $= 2$, also ist q $=$ s $= 2$
- Cramers $V = \sqrt{\frac{71.5}{1500 * (2-1)}} = 0.2(2)$
- Cramers $V = 0.2$ entspricht einem Zusammenhang mittlerer Stärke
  $\Rightarrow$ Warum?

Chi^2-Koeffizient (χ^2) - Interpretation

- Nach Diaz-Bone (2006: 91) Faustformel zur Interpretation des metrischen Korrelationskoeffizienten Pearsons r ($\Rightarrow$ nächste Sitzung!)

$0{,}00 \leq r \leq 0{,}05$	keine Korrelation
$0{,}05 < r < 0{,}20$	schwache Korrelation
$0{,}20 < r < 0{,}50$	mittlere Korrelation
$0{,}50 < r < 0{,}70$	starke Korrelation
$0{,}70 < r < 1{,}00$	sehr starke Korrelation

- Diese Tabelle lässt sich auch auf Cramers V, Yules Q sowie auf andere Koeffizienten - die von -1 bis +1 reichen - anwenden.

Ausblick / Aufgaben

- Im Tutorium werden die Berechnungen der Zusammenhangsmaße vertieft.
- Dazu finden Sie auf Learnweb entsprechende Übungsaufgaben, die in der Tutoriumssitzung diskutiert werden. Bitte machen Sie sich bereits vorab mit diesen Aufgaben vertraut.