Statistik II - Sitzung 2

Lena Masch

Institut für Politikwissenschaft

Sitzung 2

Statistik II - Sitzung 2

1. Organisatorisches
2. Wiederholung Kausalität
3. Wiederholung Zusammenhangsmaße
   - Die Kreuztabelle - Grundlegendes
   - Die Kreuztabelle als Indikator für den Zusammenhang
   - Zusammenhangsmaße für nominal skalierte Variablen

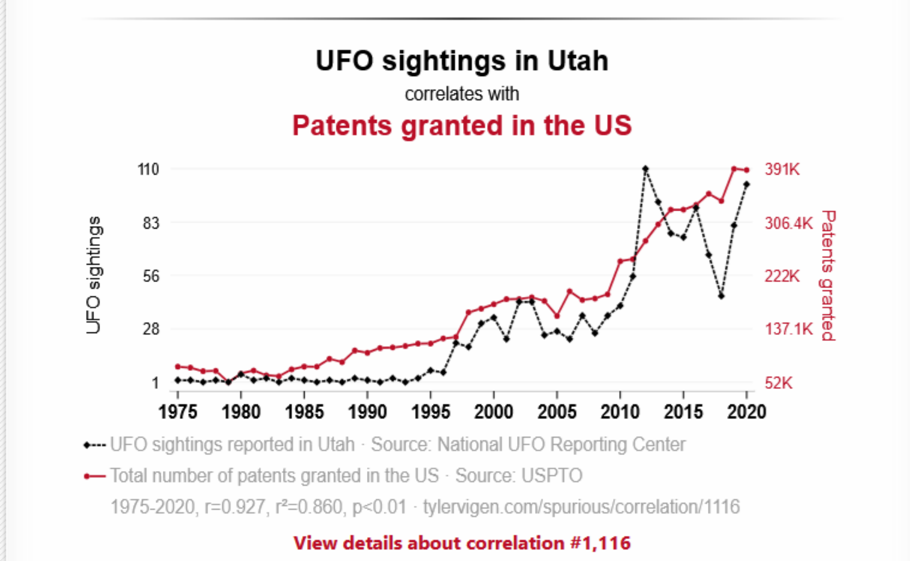
Organisatorisches

- Fachschaft Politik?
- Brandschutz
- 5min Pause nach 40-45min
- Fragen: grundsätzlich jederzeit
- kein Abfilmen, Teilen oder Speichern der Aufzeichnungen
- optionale Literatur und Materialien im Learnweb
- Tutorienvergabe: 15.10.-18.10.

Organisatorisches

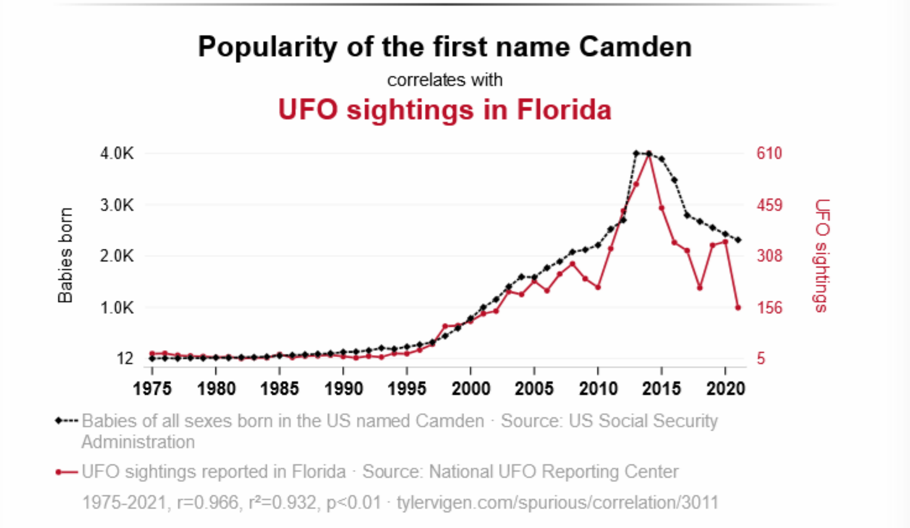
Das Veröffentlichen oder Teilen von Bild- und Tonaufzeichnungen dieser Lehrveranstaltung ist nicht gestattet.

Korrelation vs. Kausalität



Quelle: www.tylervigen.com

Korrelation vs. Kausalität



Quelle: www.tylervigen.com

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 ($= Z, IntV$)

Abhängige Variable

- **Abhängige Variable** (AV, engl. dependent variable, DV) – sind jene Eigenschaften eines Falles, welche erklärt werden sollen (lat. explanandum = das zu Erklärende)
  - Bsp.: AfD-Wahlabsicht
  - Bsp.: Handeln von Staaten
  - Bsp.: Kostenentscheidungen in Kommunen in einem Bundesland

Unabhängige Variable

- **Unabhängige Variable** (UV, engl. independent variable, IV) – sind jene Eigenschaften eines Falles, welche die abhängige Variable erklären sollen (lat. explanans = das Erklärende)
  - Bsp.: UV = Einstellung gegenüber Immigration ⇒ AV = AfD-Wahlabsicht
  - Bsp.: UV = Bedrohungswahrnehmung ⇒ AV = Handeln von Staaten
  - Bsp.: UV = Politische Ziele ⇒ AV = Kostenentscheidungen in Kommunen in einem Bundesland

Intervenierende Variable

- **Intervenierende Variable** (IntV, engl. intervening / mediating variable) – sind Eigenschaften eines Falles, welche den Zusammenhang zwischen den unabhängigen und den abhängigen Variablen beeinflussen

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
- Deduktive / Induktive Forschung gehen unterschiedlich mit diesem Problem um

Kausalität und deduktives Vorgehen

Deduktives Vorgehen

1. Entwicklung einer **Theorie zum gerichteten Zusammenhang zweier Variablen** $\Rightarrow$
2. Aus der Theorie abgeleitete Hypothesen zum Zusammenhang zweier Variablen $\Rightarrow$
3. Statistische Überprüfung der Hypothesen

Kausalität und induktives Vorgehen

Induktives Vorgehen

1. Statistische Berechnung / Beobachtung eines **ungerichteten** Zusammenhangs zwischen zwei Variablen $\Rightarrow$
2. Nachdenken über eine / Entwicklung einer Theorie des **gerichteten Zusammenhangs** zwischen zwei Variablen

Kausalität und deduktives / induktives Vorgehen

Beispiel: Zusammenhang zwischen Demokratisierung und wirtschaftlicher Entwicklung

- Vor einigen Jahrzehnten wurde beobachtet, dass es einen (damals noch *ungerichteten*) Zusammenhang zwischen dem Grad an Demokratisierung und dem wirtschaftlichen Wohlergehen eines Staates gibt. D.h., unter den reichen Staaten gab es mehr Demokratien

Kausalität und deduktives / induktives Vorgehen

Beispiel: Zusammenhang zwischen Demokratisierung und wirtschaftlicher Entwicklung

- Die Frage war (und ist immer noch), ob es einen gerichteten Zusammenhang zwischen beiden Variablen (Grad an Demokratisierung, Reichtum des Staates) gibt – und wenn ja, in welche Richtung?

Kausalität und deduktives / induktives Vorgehen

Es existieren zwei Möglichkeiten eines gerichteten Zusammenhangs

- Entweder bieten Demokratien (= X) einen besseren Nährboden für die wirtschaftliche Entwicklung (= Y), so dass der Grad der Demokratisierung (= UV) den Grad an wirtschaftlicher Entwicklung (= AV) erklärt

Kausalität und deduktives / induktives Vorgehen

Es existieren zwei Möglichkeiten eines gerichteten Zusammenhangs

- Oder: Die wirtschaftliche Entwicklung eines Staates (= X) führt zu Liberalisierungs-Tendenzen, welche sich wiederum im Bedürfnis der Gesellschaft nach mehr (politischer) Freiheit widerspiegelt und damit zur Demokratisierung des Staates (= Y) führt.

Kausalität und induktives / deduktives Vorgehen

Das **induktive** Vorgehen endet nun da, wo aus der Beobachtung des ungerichteten Zusammenhangs eine theoretische Behauptung aufgestellt wird.

- Theorie I: Wirtschaftliche Entwicklung führt zu Demokratisierung (Klassiker der Modernisierungstheorie, siehe Lipset 1960)
- Theorie II: Demokratisierung führt zu wirtschaftlicher Entwicklung (Ersson/Lane 1996)

Kausalität und deduktives / induktives Vorgehen

Das **deduktive** Vorgehen startet mit diesen theoretischen Annahmen, welche eine Kausalitäts-Aussage treffen, und versucht, die aus der Theorie abzuleitenden Erwartungen (= Hypothesen) zu überprüfen

- In der quantitativen Auseinandersetzung geschieht diese Überprüfung mit statistischen Methoden

Kausalität und deduktives / induktives Vorgehen

Am Beispiel der Modernisierungstheorie müsste also anhand der vier Kriterien für Kausalität nach Diaz-Bone (2006) nachgewiesen werden, dass

1. Die wirtschaftliche Entwicklung der Demokratisierung zeitlich vorangeht
2. Der Zusammenhang zwischen Demokratisierung und wirtschaftlicher Entwicklung tatsächlich statistisch belegbar ist
3. Der Zusammenhang nicht durch das Einwirken einer dritten Variable (traditionelle politische Kultur, Druck von außen, ...) zustande gekommen ist
4. Der vorgeschlagene gerichtete Zusammenhang zwischen beiden Variablen theoretisch tatsächlich Sinn ergibt

Kausalität und deduktives / induktives Vorgehen

- 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!

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

	X1	**H2**
Y1	f_{11}	f_{12}
Y2	f_{21}	f_{22}

Die Kreuztabelle

- Hypothetisches, aber grundlegend realitätsnahes Beispiel: Der Zusammenhang zwischen Wahlabsicht und Studiengang

	Politikwissenschaft	**Rechtswissenschaft**
Partei A	10	3
Partei B	5	7

- Die Kombination der Ausprägungen „Politikwissenschaft“ und „Partei A“ kommt der Tabelle zufolge 10mal vor. Anders formuliert: 10 Fälle weisen die Kombination „Politikwissenschaft“ und „Partei A“ auf.

Die Kreuztabelle

- Hypothetisches, aber grundlegend realitätsnahes Beispiel: Der Zusammenhang zwischen Studiengang und Wahlverhalten

	Politikwissenschaft	**Rechtswissenschaft**
Partei A	10	3
Partei B	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.
- Spaltensumme oder Reihensumme ergeben jeweils N, d.h. die Fallzahl der Verteilung

	Politikwissenschaft	Rechtswissenschaft	$\sum$ **Reihensumme**
Partei A	10	3	**13**
Partei B	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 einer Zeilenzelle die Häufigkeit einer Ausprägung von Y unter der Bedingung steht, dass X einen bestimmten Wert (= Ausprägung) annimmt
- Das Auftreten der Y-Werte „Partei A" und „Partei B" wird also bedingt durch die Ausprägungen der Variable X (= Studiengang).

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

	Politikwissenschaft	Rechtswissenschaft	$\sum$ **Reihensumme**
Partei A	66.7% (= (10/15)*100)	30% (= (3/10)*100)	**52% = 13**
Partei B	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

	Politikwissenschaft	Rechtswissenschaft	$\sum$ **Reihensumme**
Partei A	66.7% (= (10/15)*100)	30% (= (3/10)*100)	**52% = 13**
Partei B	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: Die Zustimmung (Wahl) scheint vom gewählten Studiengang eines Individuums statistisch abhängig zu sein

Die Prozentsatzdifferenz

- 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

Die Prozentsatzdifferenz

- Die Berechnung der Prozentsatzdifferenz erfolgt über die Formel d$\% = y_{1|x1} - y_{1|x2}$ oder d$\% = y_{2|x1} - y_{2|x2}$

	Politikwissenschaft	Rechtswissenschaft	$\sum$ **Reihensumme**
Partei A	66.7% $(= Y_{1\|x1})$	30% $(= Y_{1\|x2})$	**52%**
Partei B	33.3% $(= Y_{2\|x1})$	70% $(= Y_{2\|x2})$	**48%**
$\sum$ **Spaltensumme**	**100%**	**100%**	**100%**

- Es gilt: Wenn d$\% = 0$, sind die beiden Variablen statistisch unabhängig voneinander
- Allerdings sagt die PD noch wenig 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?

	PW ($=X_1$)	RW ($=X_2$)	$\sum$
Partei A ($=Y_1$)	f_{11}	f_{12}	$f_{1.}$
Partei B ($=Y_2$)	f_{21}	f_{22}	$f_{2.}$
$\sum$	$f_{.1}$	$f_{.2}$	

Odds Ratio (OR) - Formalisierung

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

Odds Ratio (OR) und Yule's Q

- Normierung des Zusammenhangs über **Yule's Q**: $Q = \dfrac{OR - 1}{OR + 1}$
- **Yule's Q** gibt dann den Zusammenhang mit Werten zwischen $\pm$ 1 wieder
- Je stärker sich Yule's Q $\pm$ 1 annähert, desto stärker der 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: $e_{ij} = \dfrac{f_{i.} * f_{.j}}{n}$

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 $\Rightarrow \dfrac{(f_{ij} - e_{ij})^2}{e_{ij}}$
- Anschließend werden die Resultate für alle Zellen zusammengezählt $\Rightarrow \sum_{Zellen}$ = Summe aller Zellen
- $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \dfrac{(f_{ij} - e_{ij})^2}{e_{ij}} = \sum_{Zellen} \dfrac{(f_{ij} - e_{ij})^2}{e_{ij}}$

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

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

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

- Hypothetische Beispielumfrage - Sollte der Name der Universität geändert werden?

	RW	**PW**	$\sum = f_{i.}$
Dafür	68	92	160
Unentschieden	22	18	40
Dagegen	110	90	200
$\sum = f_{.j}$	200	200	400 (= 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

	RW	**PW**	$\sum = f_{i.}$
Dafür	68 / 80 $(= \frac{f_{dafuer} * f_{RW}}{400})$	92	160
Unentschieden	22	18	40
Dagegen	110	90	200
$\sum = f_{.j}$	200	200	400 (= 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

	RW	**PW**	$\sum = f_{i.}$
Dafür	68 / 80	92 / 80	160
Unentschieden	22 / 20	18 / 20	40
Dagegen	110 / 100	90 / 100	200
$\sum = f_{.j}$	200	200	400 (= 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 = Studium**	f_{ij}	e_{ij}	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
Dafür	RW	68	80	144	1.8
Dafür	PW	92	80	144	1.8
Unentschieden	RW	22	20	4	0.2
Unentschieden	PW	18	20	4	0.2
Dagegen	RW	110	100	100	1
Dagegen	PW	90	100	100	1
					.

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 = Studium**	f_{ij}	e_{ij}	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
Dafür	RW	68	80	144	1.8
Dafür	PW	92	80	144	1.8
Unentschieden	RW	22	20	4	0.2
Unentschieden	PW	18	20	4	0.2
Dagegen	RW	110	100	100	1
Dagegen	PW	90	100	100	1
	$\sum$	400	400		**6**

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

- $\chi^2 = 6 \Rightarrow$ Die Variablen hängen also zusammen bzw. sind NICHT statistisch unabhängig voneinander
- Dazu wird ein das beobachtete χ^2 mit einem kritischen Wert verglichen (dazu mehr in späteren Sitzungen)
- Eine Irrtumswahrscheinlichkeit bleibt bestehen
- In diesem Beispiel scheint der gewählte Studiengang einer/s Befragten und seine/ihre Einstellung zur Namensänderung 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 $V = \sqrt{\dfrac{\chi^2}{N * (q-1)}}$
- 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 = 3 und s = 2, also ist q = s = 2
- Cramers $V = \sqrt{\frac{6}{400 * (2-1)}} = 0.122$
- Cramers $V = 0.12$ entspricht einem Zusammenhang schwacher 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

- In der nächsten Sitzung wiederhole ich zunächst das Konzept der bivariaten Regression. In den dann folgenden beiden Sitzungen werden wir uns der Verbindung zwischen Kausalitätsüberlegungen und Zusammenhangsmaßen widmen
- Dafür gehen wir auf den zweiten Bereich der schließenden oder **inferentiellen Statistik** ein: das **Testen von Hypothesen**
- Darin spielt das **Konzept der Signifikanz** eine entscheidende Rolle - mehr dazu in Sitzung 3!