Willkommen zur Vorlesung Statistik

Thema dieser Vorlesung:
Analyse von Kreuztabellen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Inhaltsübersicht

Organisatorisches (WiSe 2019/20)

- Die Frist für Anmeldung der Studien- oder Prüfungsleistung läuft seit gestern!
  Achtung: Ich gewähre prinzipiell keine Möglichkeit zur Nachmeldung.
- ImJanuar (Beginn 8. 1.) findet jeweils Mittwochs von 14 bis 16 Uhr ein mündliches Tutorium statt, in dem der Stoff noch einmal durchgesprochen / wiederholt wird. Besonders empfehlenswert für jene, die noch Fragen haben – aber es dürfen alle kommen, die an Wiederholung Interesse haben.
  Raum: AR-HB 101/102.

Analyse von Zusammenhängen

Mit der heutigen Vorlesung beginnen wir, Zusammenhänge zwischen (zunächst) zwei Merkmalen (Variablen) zu analysieren. Dabei müssen wir unterscheiden:

- In manchen Fällen wollen wir nur wissen, ob zwei Merkmale miteinander zusammenhängen.
- In anderen Fällen gehen wir davon aus, dass ein Merkmal durch ein (oder mehrere) andere(s) beeinflusst wird. In diesem Fall ist das Ziel die Analyse von Kausalität (Verursachung).

Im zweiten Fall (Kausalanalyse) unterscheiden wir zwischen

- der abhängigen Variablen, dem Merkmal, dessen Ausprägungen erklärt werden sollen, und
- der oder den unabhängigen Variablen, dem Merkmal oder den Merkmalen, die die abhängige Variable beeinflussen.

Kausalität und Datenanalyse

Wenn wir Kausalanalysen durchführen wollen, müssen wir wissen: Die statistische Analyse selbst kann nicht entscheiden, ob Merkmal X einen Einfluss auf Y hat oder umgekehrt! Die Richtung der Kausalanalyse muss die Forscherin nach ihrem Wissen über das Forschungsdesign und die Datenerhebung festlegen.

Beispiele:

- Wir stellen fest, dass arbeitslose Personen häufiger krank sind als erwerbstätige Personen $\rightarrow$ Verursacht Arbeitslosigkeit die Krankheit, oder werden Kranke häufiger arbeitslos?
  Bei einer Querschnittserhebung (Messung von Arbeitslosigkeit und Gesundheit zum gleichen Zeitpunkt) kann dies nicht entschieden werden.
- Wir stellen fest, dass Kinder von Eltern mit niedriger Bildung in der Schule schlechter abschneiden als solche von Eltern mit höherer Bildung.
  $\rightarrow$ Da die Bildung der Eltern dem Schulerfolg der Kinder zeitlich vorangeht, ist hier die Richtung des Einflusses klar.
  Allerdings ist unklar, wie der Zusammenhang genau entsteht.

Kreuztabellen als Zusammenhangsanalyse

Kreuztabellen eignen sich zur Darstellung und Analyse von Zusammenhängen zwischen nominal- und ordinalskalierten Merkmalen (in allen Kombinationen, also nominal/nominal, nominal/ordinal, ordinal/ordinal). Charakteristischerweise (aber nicht zwingend) haben die Merkmale wenige Ausprägungen.

Metrische Merkmale (mit vielen Ausprägungen) müssen für eine Präsentation in Kreuztabellen in geeigneter Weise zusammengefasst werden. Dies sollte – wenn überhaupt – nur zum Zweck der übersichtlichen Präsentation von Daten geschehen, auf keinen Fall für die statistische Analyse (Informationsverlust!).

Ausgangsmaterial einer Kreuztabelle

Für eine Kreuztabelle zweier Merkmale erstellen wir gewissermaßen mehrere Häufigkeitstabellen, nämlich solche mit bedingten Häufigkeiten, d. h. Häufigkeitsverteilungen für verschiedene Bedingungen.

Beispiel (Zahlen fiktiv, Zusammenhang an sich nicht unrealistisch): Berufliche Stellung des Vaters (Arbeiter vs. Angestellter) und Schulbesuch des Kindes (Haupt-/Realschule vs. Gymnasium). Wir erhalten folgende Teiltabellen:

	Arbeiter		Angest.	
Schule	h_j	%	h_j	%
Haupt-/Realschule	360	90	270	45
Gymnasium	40	10	330	55

Darstellung von Zusammenhängen in Kreuztabellen

- Wenn möglich: Unabhängige (erklärende) Variable als Spaltenvariable
- Prozentuierung in den „Teiltabellen“ der Ausprägungen der unabhängigen Variablen (im Regelfall also in den Spalten)
- Absolute Zahlen nur in den Randverteilungen.

	Beruf Vater		
Schule	Arbeiter	Angest.	n
Haupt-/Realschule	90	45	630
Gymnasium	10	55	370
n	400	600	1000

Eine Tabelle hat nicht mehr horizontale Linien als unbedingt nötig; vertikale Linien sind nur in Ausnahmefällen sinnvoll. Negativbeispiel: Folgende Folie.

Schlechte Gestaltung der Tabelle

Kreuztabellen dürfen nicht so, wie sie von Software erzeugt werden, in wissenschaftliche (oder andere) Berichte übernommen werden.

Beispiel eines weit verbreiteten Statistik-Paketes (SPSS) (mit absoluten Häufigkeiten und Spaltenprozent):

	Arbeiter	Angestellter	n
Haupt-/Realschule	360 90,0%	270 45,0%	630 63,0%
Gymnasium	40 10,0%	330 55,0%	370 37,0%
n	400	600	1000

Achtung: Auch weit verbreitete Textverarbeitungssoftware erzeugt Tabellen per Voreinstellung im „Gefängnis-Look“ (Blick auf Zahlen durch Gitter) $\rightarrow$ weg mit den vielen Linien (und wenn Sie es noch so oft in Texten anderer Leute sehen).

Weitere Hinweise zur Gestaltung

- Normalerweise (vor allem bei ohnehin fehlerbehafteten Stichprobendaten) ist die Angabe der Prozentwerte mit Nachkommastellen überflüssig.
- Auch genügt es meist, die absoluten Zahlen in den Randverteilungen auszuweisen.
- Ebenso wenig muss hinter jeder einzelnen Prozentzahl ein Prozentzeichen stehen.
- Dafür sollte möglichst in der Tabellenüberschrift (oder in der Legende) angegeben sein, ob es sich um Spalten- oder Zeilenprozent (oder evtl. beides) handelt.

Achtung: Prozentuierungsrichtung I

Ein anderes Beispiel: Berufliche Bildung oder Studium in Abhängigkeit vom Elternhaus – wiederum fiktive Zahlen, aber angelehnt an reale Verhältnisse (Tabelle zeigt absolute Zahlen).

	1990		2010	
	Arbeiter	Andere	Arbeiter	Andere
Berufliche Bildung	400	1000	200	1100
Studium	100	500	100	600
n	500	1500	300	1700

Errechnet man die Spaltenprozent, so zeigt sich: Unter den Arbeiterkindern hat der Anteil derer, die studieren, zugenommen – und zwar deutlicher als unter den Kindern aus anderen Elternhäusern.

Achtung: Prozentuierungsrichtung II

Hier die gleiche Tabelle mit Zeilenprozentwerten (innerhalb jeder Teiltabelle!).

	1990		2010	
	Arbeiter	Andere	Arbeiter	Andere
Berufliche Bildung	29	71	15	85
Studium	17	83	14	86
n	500	1500	300	1700

Die Aussage „Der Anteil der Arbeiterkinder unter den Studierenden ist zurückgegangen“ wird häufig so interpretiert, als hätten die Arbeiterkinder schlechtere Chancen, zu studieren. Tatsächlich ist aber einfach der Anteil der Arbeiter in der Bevölkerung zurückgegangen (während die Chancen der Arbeiterkinder stärker gestiegen sind als die der Kinder aus anderen Elternhäusern).

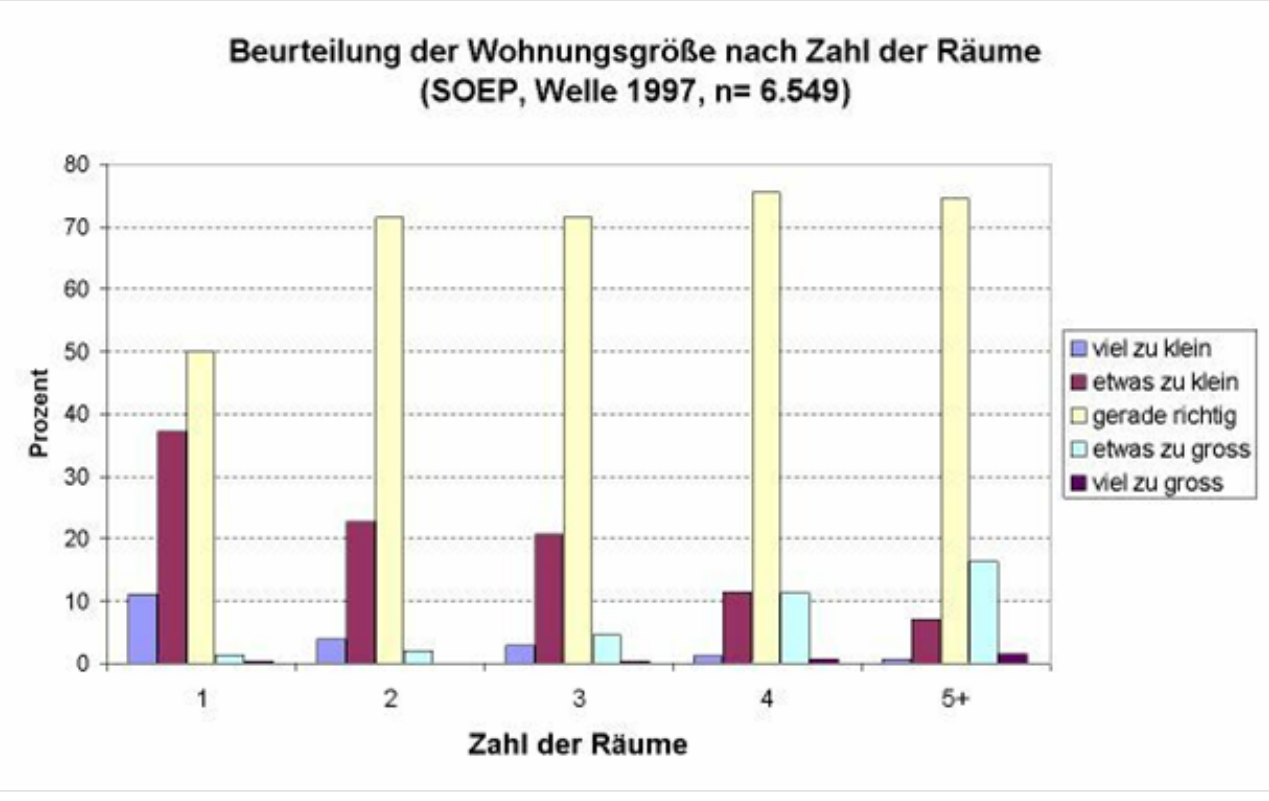
Achtung: Prozentuierungsrichtung III

Achtung: Es ist selbstverständlich erlaubt, die unabhängige Variable in den Zeilen (und die abhängige in den Spalten) abzutragen; dies empfiehlt sich vor allem, wenn die unabhängige Variable eine größere Zahl von Ausprägungen hat. Für Ursache-Wirkungs-Aussagen ist entscheidend:

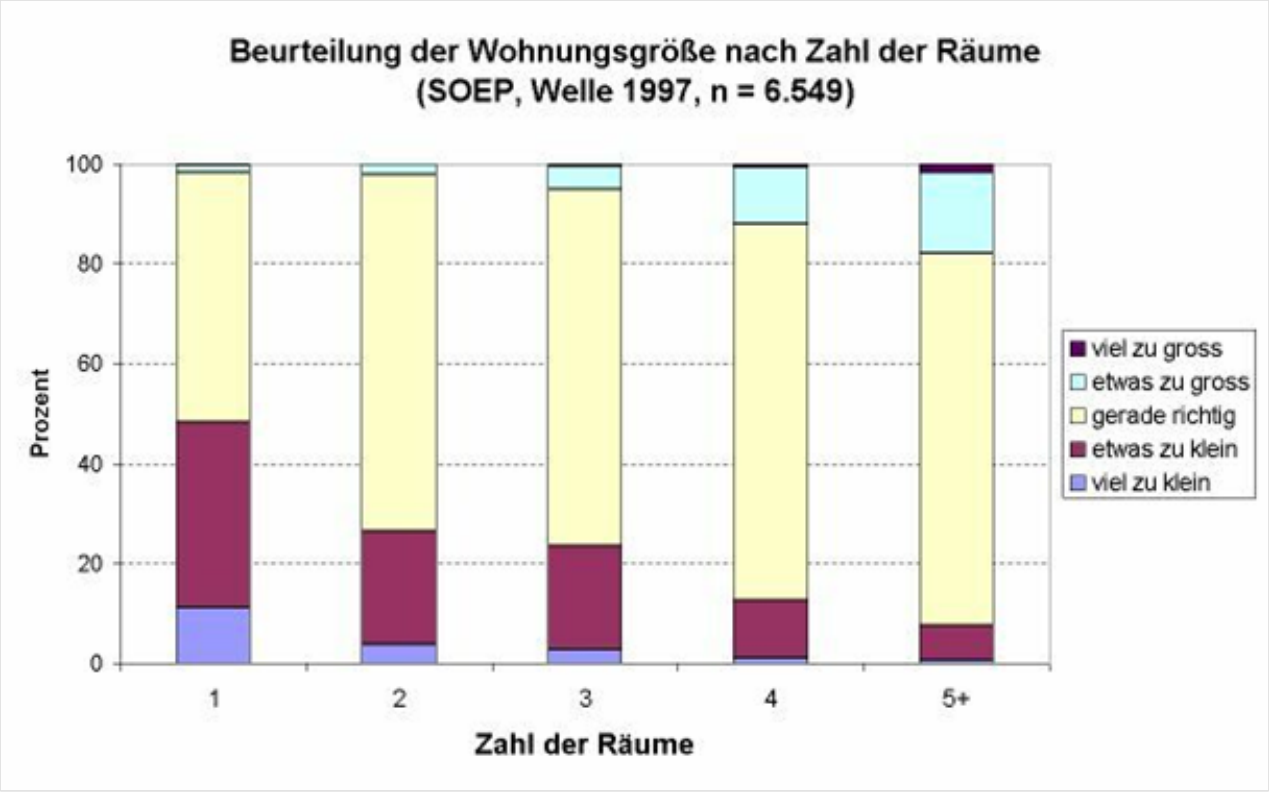
Prozentuiert wird innerhalb der einzelnen Ausprägungen der unabhängigen Variablen über die verschiedenen Ausprägungen der abhängigen Variablen – gleichgültig, ob die unabhängige Variable in den Spalten oder in den Zeilen steht.

Und: Auch die Prozentuierung innerhalb der Ausprägungen der abhängigen Variablen kann sinnvoll sein – sie führt aber zu anderen Aussagen, nämlich solchen über die Zusammensetzung der unterschiedlichen Gruppen. Beispiel: Unter den Personen mit beruflicher Ausbildung waren im Jahr 1990 29 Prozent Arbeiterkinder.

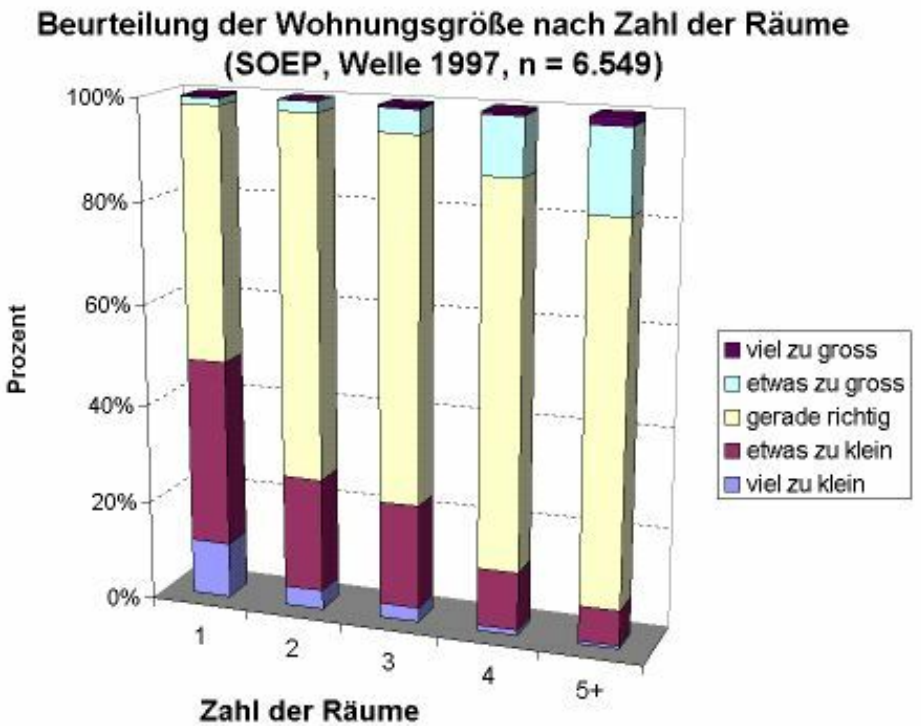
Graphische Darstellung I: Gruppiertes Säulendiagramm



Graphische Darstellung II: Gestapeltes Säulendiagramm



Graphische Darstellung III: Chart-Junk



Formale Notation zur Bezeichnung der Zellen

In Formeln werden die Zellen der Tabelle mit Indices versehen. Der erste Index bezieht sich auf die Zeile, der zweite auf die Spalte. Das Zeichen $\bullet$ steht für die gesamte Zeile bzw. Spalte.

Am Beispiel einer Vier-Felder-Kreuztabelle:

	X=1	X=2	$\sum$
Y=1	n_{11}	n_{12}	$n_{1\bullet}$
Y=2	n_{21}	n_{22}	$n_{2\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Lesebeispiele: n_{21} heißt: Häufigkeit in der zweiten Zeile und der ersten Spalte. $n_{\bullet 2}$ heißt: Häufigkeiten über alle Zeilen in der zweiten Spalte.

Formale Notation: Bedingte Anteilswerte

Um die Spaltenprozent (oder Zeilenprozent) kennzuzeichen, verwenden wir ein Symbol für „bedingte Anteilswerte“ (Anteils- bzw. Prozentwert unter der Bedingung ...).

Die folgende Tabelle zeigt ein Beispiel für Spaltenprozente:

	X=1	X=2
Y=1	$p_{1\|X=1}$	$p_{1\|X=2}$
Y=2	$p_{2\|X=1}$	$p_{2\|X=2}$

Lesebeispiel: $p_{1|X=2}$ heißt: Anteilswert (oder Prozentwert) in der ersten Zeile unter der Bedingung $X = 2$ (X hat den Wert 2).

(Eine kürzere Schreibweise wäre: $p_{1|2}$. Die hier verwendete Schreibweise dient dazu, den Unterschied zwischen bedingten und nicht bedingten Anteilswerten klarer erkennen zu lassen.)

Stärke des Zusammenhangs I: Die Prozentsatzdifferenz

„Bei spaltenbezogener Prozentuierung kann die Differenz der Spaltenprozentwerte innerhalb einer Zeile als Maß für die Stärke des Zusammenhangs herangezogen werden.“ (K & K, S. 317 [1. Auflage]).

Am Beispiel der ersten Zeile:

$$d_{XY}\% = 100 \cdot \left(\frac{n_{11}}{n_{\bullet 1}} - \frac{n_{12}}{n_{\bullet 2}} \right) = 100 \cdot \left(p_{1|X=1} - p_{1|X=2} \right)$$

oder einfacher: $\text{Spaltenprozent}_{11} - \text{Spaltenprozent}_{12}$.

Ebenso kann man $\text{Spaltenprozent}_{21} - \text{Spaltenprozent}_{22}$ (oder auch $\text{Spaltenprozent}_{12} - \text{Spaltenprozent}_{11}$ usw.) berechnen; entscheidend ist anzugeben, was man berechnet.

Prozentsatzdifferenz im Beispiel

In unserer Beispieltabelle ergibt sich bspw. eine Prozentsatzdifferenz zwischen Arbeiterkindern und Angestelltenkindern hinsichtlich des Haupt-/Realschulbesuchs von 45.

Wir können sagen: Der Anteil der Haupt- und Realschüler unter den Arbeiterkindern ist um 45 Prozentpunkte (!) höher als der Anteil der Haupt-/Realschüler unter den Angestelltenkindern.

Falsch: Der Anteil der Haupt-/Realschüler unter den Arbeiterkindern liegt um 45 Prozent über dem Anteil der Haupt-/Realschüler unter den Angestelltenkindern.

Die Prozentsatzdifferenz zwischen Arbeiterkindern und Angestelltenkindern hinsichtlich des Besuchs eines Gymnasiums beträgt dagegen -45.

Stärke des Zusammenhangs II: Das relative Risiko

Das relative Risiko ist das Verhältnis zweier bedingter Anteils- oder Prozentwerte. Im Beispiel: Das Risiko von Arbeiterkindern, nur die Haupt-/Realschule zu besuchen ist mit 90 % / 45 % = 2 doppelt so groß wie das von Angestelltenkindern. Allgemein:

$$RR = \frac{p_{1|X=1}}{p_{1|X=2}}$$

wenn die unabhängige Variable als Spaltenvariable abgetragen ist.

Statt der Prozentwerte in der ersten Zeile können auch die Werte aus anderen Zeilen zu einander in Beziehung gesetzt werden.

Ist dagegen die unabhängige Variable als Zeilenvariable abgetragen (und ist entsprechend zeilenweise prozentuiert), muss $RR = \frac{p_{\mathbf{1}|X=\mathbf{1}}}{p_{\mathbf{2}|X=\mathbf{1}}}$ verwendet werden.

Stärke des Zusammenhangs III: Die (das) Odds Ratio I

Odds = „Gewinnchancen“; Ratio = Verhältnis

Die Odds lassen sich anhand von bedingten Prozentwerten oder von absoluten Werten berechnen. Aus Gründen der Genauigkeit sollten die absoluten Zahlen verwendet werden, wenn die Prozentwerte gerundet vorliegen. Die Odds der Arbeiterkinder, eine Haupt-/Realschule zu besuchen, sind:

$$\text{Odds}(1{,}2|\text{Arbeiter}) = \frac{360}{40} = \frac{90\,\%}{10\,\%} = 9{,}0$$

Man kann aber auch berechnen:

$$\text{Odds}(2{,}1|\text{Arbeiter}) = \frac{40}{360} = \frac{10\,\%}{90\,\%} = 0,\overline{11} = \frac{1}{9}$$

Stärke des Zusammenhangs III: Die (das) Odds Ratio II

Ähnlich lassen sich die Odds für die Angestelltenkinder berechen, z. B.

$$\text{Odds}(1{,}2|\text{Angestellte}) = \frac{270}{330} = \frac{45\,\%}{55\,\%} = 0,\overline{8181}$$

Die Odds Ratio setzt nun die beiden Odds zu einander ins Verhältnis, z. B.:

$$OR_{1,2|\text{Arbeiter/Angestellte}} = \frac{360/40}{270/330} = \frac{360 \cdot 330}{40 \cdot 270} = \frac{9}{0,\overline{8181}} = 11$$

Die „Chancen“ der Arbeiterkinder auf den Besuch der Haupt-/Realschule sind also 11 mal so groß wie die der Angestelltenkinder. Ähnlich lassen sich auch OR für die Odds Gymnasium/andere Schulformen bzw. für die Angestellten- im Vergleich zu den Arbeiterkindern berechnen.

Stärke des Zusammenhangs III: Die (das) Odds Ratio III

Allgemein formuliert gilt:

$$\text{Odds}(1,2|X=1) = \frac{n_{11}}{n_{21}} = \frac{p_{1|X=1}}{p_{2|X=1}}$$

$$\text{Odds}(1,2|X=2) = \frac{n_{12}}{n_{22}} = \frac{p_{1|X=2}}{p_{2|X=2}}$$

$$\text{OR} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Wegen des letzten Ausdrucks wird die OR gelegentlich auch als Kreuzproduktverhältnis bezeichnet. Zum Zweck des Nachvollziehens ist es aber immer zweckmäßig, den Berechnungsweg über die Odds zu wählen!

Prozentsatzdifferenz, RR und OR im Vergleich

- Prozentsatzdifferenz: 0 = kein, > 0 = positiver, < 0 = negativer Zusammenhang
- RR: 1 = kein, > 1 = positiver, < 1 = negativer Zusammenhang
- OR: 1 = kein, > 1 = positiver, < 1 = negativer Zusammenhang

Aber: Richtung des Zusammenhang hängt davon ab, welche Ausprägungen wie zueinander in Beziehung gesetzt werden.

Prozentsatzdifferenz, RR oder OR?

Setzen wir unser (fiktives, aber nicht ganz unrealistisches) Beispiel fort: Vergleich von 1960 und 2000.

	1960		2000	
	Arbeiter	Angestellter	Arbeiter	Angestellter
Haupt-/Realschule	95	60	90	45
Gymnasium	5	40	10	55

$$d_{XY}\%_{(1960)} = 35; \quad RR_{(1960)} = \frac{95}{60} = 1{,}58; \quad \mathrm{OR}_{(1960)} = \frac{95/5}{60/40} = 12{,}66$$

$$d_{XY}\%_{(2000)} = 45; \quad RR_{(2000)} = \frac{90}{45} = 2{,}00; \quad \mathrm{OR}_{(2000)} = \frac{90/10}{45/55} = 11{,}00$$

Prozentsatzdifferenz, RR oder OR?

Der Nachteil des relativen Risikos gegenüber den anderen beiden Maßen ist, dass RR sich bei Betrachtung der zweiten Zeile nicht spiegelbildlich verhält.

Vergleichen wir die Berechnung der Maße für die Beispieltabelle der vorigen Folie (nur das Jahr 2000):

Berechnung ausgehend von der ersten Zeile (wie letzte Folie):

$$d_{XY}\% = 45; \quad rr = \frac{90}{45} = 2{,}00; \quad \text{OR} = \frac{90/10}{45/55} = 11{,}00$$

Berechnung ausgehend von der zweiten Zeile:

$$d_{XY}\% = -45; \quad RR = \frac{10}{55} = 0{,}1818; \quad \text{OR} = \frac{10/90}{55/45} = 0{,}09 = 1/11$$

Abschließendes zu Prozentsatzdifferenz, RR oder OR?

Bei Berechnung von Prozentsatzdifferenz, relativem Risiko oder Odds Ratio muss immer angegeben werden, welche Ausprägungen zu einander in Beziehung gesetzt werden (z. B.: relatives Risiko der Arbeiterkinder im Vergleich zu den Angestelltenkindern hinsichtlich Haupt-/Realschulbesuch vs. Besuch eines Gymnasiums).

Das gilt erst recht bei größeren Tabellen (mehr Zeilen und/oder mehr Spalten). Für die Berechnung der hier diskutierten Größen muss man auch bei solchen Tabellen zwei Zeilen und zwei Spalten auswählen, die genau zu bezeichnen sind.

Auf der Suche nach Kausalität: Mehrdimensionale Tabellen

Bivariate Tabellen können über den wahren Ursache-Wirkungs-Zusammenhang täuschen.

Ein fiktives Beispiel: Zusammenhang zwischen Aufsuchen eines Arztes bei Erkältung und Geschwindigkeit der Heilung.

	absolute Zahlen			Spaltenprozent	
Heilung	Arzt	kein Arzt	n	Arzt	kein Arzt
Rasch	262	232	494	44	59
Langsam	328	159	487	56	41
n	590	391	981	100	100

Nach Benninghaus, Hans: Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler. 9. überarbeitete Auflage. Wiesbaden: Westdeutscher Verlag, 2002, S. 263.

Auf der Suche nach Kausalität: Drittvariablenkontrolle

Leichte Erkältung

	absolute Zahlen			Spaltenprozent	
Heilung	Arzt	kein Arzt	n	Arzt	kein Arzt
Rasch	126	188	314	85	83
Langsam	23	38	61	15	17
n	149	226	375	100	100

Schwere Erkältung

	absolute Zahlen			Spaltenprozent	
Heilung	Arzt	kein Arzt	n	Arzt	kein Arzt
Rasch	136	44	180	31	27
Langsam	305	121	426	69	73
n	441	165	606	100	100

Mehr zu Drittvariablenkontrolle

Das Beispiel der vorstehenden Folien zeigt die Aufdeckung von nur scheinbar bestehenden Zusammenhängen („Scheinkorrelation“).

Drittvariablenkontrolle dient auch der Aufdeckung nur scheinbarer Nicht-Zusammenhänge bzw. dem Nachweis unterschiedlicher Wirkungen. Dazu noch ein (wiederum fiktives, aber nicht unrealistisches) Beispiel: Zusammenhang zwischen Besuch des Tutoriums und Bestehen der Statistik-Klausur.

	absolute Zahlen			Spaltenprozent	
Klausur	Tutorium	kein Tut.	n	Tutorium	kein Tut.
Nicht bestanden	20	40	60	20	20
Bestanden	80	160	240	80	80
n	100	200	300	100	100

Drittvariablenkontrolle

Angst vor Statistik

	absolute Zahlen			Spaltenprozent	
Klausur	Tutorium	kein Tut.	n	Tutorium	kein Tut.
Nicht bestanden	15	16	31	25	44
Bestanden	45	20	65	75	56
n	60	36	96	100	100

Keine Angst vor Statistik

	absolute Zahlen			Spaltenprozent	
Klausur	Tutorium	kein Tut.	n	Tutorium	kein Tut.
Nicht bestanden	5	24	29	12,5	15
Bestanden	35	140	175	87,5	85
n	40	164	204	100	100

Grenzen der Drittvariablenkontrolle

Die Drittvariablenkontrolle durch Kreuztabellenanalyse ist beschränkt:

- Ein viertes Merkmal ist nur mehr sehr schwer kontrollierbar, weitere praktisch gar nicht mehr.
- Es fehlt eine Möglichkeit, den relativen Einfluss der Merkmale zu beurteilen.
- Metrische Merkmale (vor allem mit vielen Merkmalen) müssen (u.U. stark) zusammengefasst werden (Informationsverlust!).

Die Lösung: Multivariate Modellierung $\rightarrow$ Auf Wiedersehen im Master-Studium!

Einen ersten Einblick gibt es aber in der Vorlesung über lineare Regressionsanalyse (beschränkt auf metrische abhängige Variable).

Der χ^2-Test nach Karl Pearson

Liegt in Stichprobendaten ein Zusammenhang vor, wollen wir in der Regel prüfen, ob wir annehmen können, dass auch in der Grundgesamtheit ein Zusammenhang besteht. Im Falle von Kreuztabellen kann hierfür der χ^2-Test nach Karl Pearson geeignet sein.

Zur Erinnerung die vier Schritte bei Signifikanztests:

1. Formulierung der Hypothesen
2. Festlegung der Teststatistik
3. Festlegung des Signifikanzniveaus
4. Berechnung der Teststatistik und Entscheidung

Der χ^2-Test: Schritt 1

Unsere Frage: Können wir annehmen, dass der Unterschied in den Anteilswerten in unserer Stichprobe auch in der Grundgesamtheit besteht?

Schritt 1: Formulierung der Hypothesen.

- Nullhypothese: Es besteht *kein* Zusammenhang zwischen den beiden Merkmalen.

  Präziser: Die Verteilung der abhängigen Variablen ist über alle Ausprägungen der unabhängigen Variablen identisch.
- Alternativhypothese: Es besteht (irgend)ein Zusammenhang zwischen den beiden Merkmalen.

  Präziser: Die Verteilung der abhängigen Variablen unterscheidet sich je nach Ausprägung der unabhängigen Variablen.

Es wird also keine gerichtete Hypothese formuliert!

Der χ^2-Test: Schritt 2

Die Teststatistik zur Prüfung der Nullhypothese bezieht sich auf die Abweichung der absoluten Häufigkeiten, die unter der Nullhypothese zu erwarten wären, von den beobachteten Häufigkeiten. Je größer die Abweichung, desto unplausibler ist die Nullhypothese.

Werte, die bei Gültigkeit der Nullhypothese (Unabhängigkeit der Merkmale) zu erwarten wären:

	Arbeiter	Angestellter	n
Haupt-/Realschule	252	378	630
	63 %	63 %	63 %
Gymnasium	148	222	370
	37 %	37 %	37 %
n	400	600	1000

Die Verteilung der Werte in den Spalten entspricht der prozentualen Randverteilung über alle Gruppen.

Der χ^2-Test: Schritt 2

Einfache Berechnung der unter der H_0 erwarteten absoluten Häufigkeiten aus der Randverteilung:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \quad \text{z. B.} \quad e_{21} = \frac{n_{2\bullet} \cdot n_{\bullet 1}}{n} = \frac{370 \cdot 400}{1000} = 148$$

	Arbeiter	Angestellter	n
Haupt-/Realschule	252	378	630
	63 %	63 %	63 %
Gymnasium	148	222	370
	37 %	37 %	37 %
n	400	600	1000

Der χ^2-Test: Schritt 2

Die genaue Formel zur Berechnung der Teststatistik finden Sie in Schritt 4. Diese Teststatistik folgt der sog. χ^2-Verteilung; diese hat unterschiedliche Werte je nach Zahl der Freiheitsgrade (siehe dazu Schritt 3).

Anwendungsvoraussetzungen des χ^2-Tests:

- Der Test ist nur gültig, wenn gilt: $e_{ij} > 5$ für alle (oder: die meisten) e_{ij}.
- Außerdem soll evtl. bei $n < 60$ die sog. Kontinuitätskorrektur nach Yates verwendet werden (das ist aber unter Statistikern umstritten). Bei sehr kleinen Fallzahlen ($n < 30$) muss zu sog. exakten Testverfahren gegriffen werden (hier nicht behandelt).

Im vorliegenden Fall gibt es in beiden Hinsichten keine Probleme, wir können also den χ^2-Test bedenkenlos anwenden.

Der χ^2-Test: Schritt 3

Das Signifikanzniveau

Üblicherweise wählen wir ein Signifikanzniveau von $\alpha = 0{,}05$. Wäre ein „Irrtum“ (d. h., eine Ablehnung der Nullhypothese, obwohl sie zutrifft), besonders gravierend, wäre auch an $\alpha = 0{,}01$ oder gar $\alpha = 0{,}001$ zu denken. Das Risiko hierbei: Die Nullhypothese beizubehalten, obwohl sie falsch ist (Fehler 2. Art).

Trifft die H_0 zu, beträgt der Wert der Teststatistik 0. Der kritische Wert der χ^2-Quadratverteilung (hier: $1 - 0{,}05 = 0{,}95$) liegt bei einer Vierfeldertabelle (1 Freiheitsgrad) bei 3,841. Erreicht oder übertrifft die Teststatistik diesen Wert, wird die Nullhypothese verworfen. (Besonders niedrige Werte sprechen also nicht gegen die Nullhypothese.)

Es gibt hier also keine Unterscheidung zwischen ein- oder zweiseitigen Hypothesen – wir können nur prüfen: Zusammenhang ja oder nein.

Allgemein beträgt die Zahl der Freiheitsgrade (d. f.) für den χ^2-Test $(I-1)(J-1)$, wobei I die Zahl der Zeilen und J die Zahl der Spalten in der Tabelle ist (ohne Randverteilung!).

Der χ^2-Test: Schritt 4

Berechnung der Teststatistik und Entscheidung über H_0:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

In Worten:

1. Berechne für jede einzelne Zelle der Tabelle die Differenz zwischen beobachtetem und unter der H_0 erwarteten Wert, quadriere diese Differenz und dividiere das Ergebnis durch den erwarteten Wert.
2. Summiere diese Werte über alle Spalten (j) und alle Zeilen (i).

Der χ^2-Test: Schritt 4

Berechnung der Teststatistik und Entscheidung über H_0 im Beispiel:

$$\begin{aligned}\chi^2 &= \frac{(360-252)^2}{252} + \frac{(40-148)^2}{148} \\ &\quad + \frac{(270-378)^2}{378} + \frac{(330-222)^2}{222} \\ &= 46{,}29 + 78{,}81 + 30{,}86 + 52{,}54 \\ &= 208{,}5\end{aligned}$$

$208{,}5 > 3{,}841 \rightarrow H_0$ wird verworfen (mit $\alpha = 0{,}05$).

Beachten Sie: Bei einer Vier-Felder-Kreuztabelle (und nur hier!) ist der Absolutbetrag der Abweichung zwischen beobachteten und erwarteten Werten (und damit der Zähler der einzelnen Summanden) für alle vier Felder gleich.

Der χ^2-Test: Probleme

- Wie bei allen Signifikanztests gilt auch hier: Der Stichprobenumfang entscheidet. Bei $n = 1000$ werden schon relativ kleine Unterschiede signifikant (ab Prozentsatzdifferenz von etwa 6 bis 7).
- Daraus folgt: Signifikanz sagt nichts über die Stärke des Zusammenhanges (gilt grundsätzlich, kann aber nicht oft genug wiederholt werden).
- Der Chi-Quadrat-Test ist ein „Omnibus-Test“: Er sagt nur, dass *irgendwelche* empirischen Werte von den erwarteten Werten abweichen. Er sagt aber nichts über die Richtung der Abweichung.

Assoziationsmaße

Zusammenhänge in größeren Kreuztabellen lassen sich im Prinzip durch eine Vielzahl von Prozentsatzdifferenzen/Relativen Risiken/Odds Ratios ausdrücken. Diese sind dann aber kaum mehr zusammenhängend zu interpretieren. Assoziationsmaße (die es im übrigen auch für Vier-Felder-Tabellen gibt) sind Versuche, die Stärke des Zusammenhanges in einer einzigen Maßzahl auszudrücken.

Die Vielzahl entsprechender Maßzahlen verbietet eine ausführliche Diskussion. Die folgende Übersicht gibt nur einige Hinweise. Die meisten Maßzahlen sind aus statistischer Sicht auch wenig sinnvoll, sie werden jedoch immer wieder verwendet.

Die wichtigste Unterscheidung betrifft das Messniveau (Skalenniveau): Maße für nominalskalierte Variablen (oder eine nominal- und eine ordinalskalierte Variable) sind von Maßen für zwei ordinalskalierte Variable zu unterscheiden.

Bei den erstgenannten unterscheiden wir zwischen χ^2-basierten und sonstigen Maßen.

χ^2-basierte Assoziationsmaße

Die χ^2-Teststatistik wird bei gegebenen Fallzahlen umso größer, je stärker der Zusammenhang ist. Sie wird jedoch grundsätzlich ganz wesentlich von der Fallzahl beeinflusst.

χ^2-basierte Assoziationsmaße korrigieren daher den χ^2-Wert um die Fallzahl.

Sie nehmen idealerweise Werte zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang) an. Dies gilt jedoch nicht für C. Für den Koeffizienten ϕ gibt es auch Berechnungsformeln, nach denen er Werte zwischen -1 (perfekter negativer) und $+1$ (perfekter positiver Zusammenhang annehmen kann.

χ^2-basierte Assoziationsmaße sind nicht sinnvoll, wenn beide Merkmale ordinalskaliert sind, weil es dann besser geeignete Maßzahlen gibt.

χ^2-basierte Assoziationsmaße

Name	Berechnung	Anmerkungen
ϕ (phi)	$\sqrt{\frac{\chi^2}{n}}$	Nur für 2x2-Tabellen geeignet; kann bei alternativer Berechnung auch Werte bis -1 annehmen.
V	$\sqrt{\frac{\chi^2}{n \cdot (\min(I,J)-1)}}$	I: Zahl der Werte von X; J: Zahl der Werte von Y Voller Name: Cramérs V
C	$\sqrt{\frac{\chi^2}{\chi^2+n}}$	Voller Name: Kontingenzkoeffizient. Achtung: Max(C) < 1!

Weitere Assoziationsmaße I: Yules Q

Die Maßzahl Q (ebenfalls für Vier-Felder-Tabellen geeignet) wird u. a. in den Lehrbüchern von Kühnel & Krebs oder Andreas Diekmann (Empirische Sozialforschung) empfohlen.

Sie ist jedoch mit größter Vorsicht zu genießen: Sie nimmt den Wert 1 (oder -1) an, wenn eine einzige Zelle die Häufigkeit Null aufweist. Daher kann Q nicht zwischen den beiden folgenden Tabellen (angegeben sind absolute Zahlen) unterscheiden:

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

Weitere Assoziationsmaße II: λ

Die Maßzahl λ (lambda) wird ebenfalls von Kühnel & Krebs empfohlen. Der Nachteil: Sie nimmt den Wert 0 an, wenn die Modalwerte pro Spalte alle in der gleichen Zeile auftreten. So beträgt der Wert von Lambda in beiden folgenden Tabellen 0:

	Arbeiter	Angestellte	Arbeiter	Angestellte
Haupt-/Realschule	210	310	400	310
Gymnasium	190	290	0	290

Alternatives Maß mit ähnlicher Logik: Goodmans und Kruskals Tau (nicht mit den nachfolgenden diskutierten Tau-a, -b und -c zu verwechseln).

Fazit zu Assoziationsmaßen für nominalskalierte Merkmale

Ein rundherum befriedigendes Assoziationsmaß für nominalskalierte Merkmale existiert nicht. (Der Unsicherheitskoeffizient – siehe wiederum Kühnel & Krebs – ist eine brauchbare, aber nicht sehr eingängige Alternative.)

In der Praxis werden oft ϕ (für 2x2-Tabellen) oder Cramérs V (für größere Tabellen) verwendet, aber nicht alle halten diese für die besten Maßzahlen (u. A. wegen fehlender inhaltlicher Interpretation und fehlender Differenzierung zwischen unabhängiger und abhängiger Variablen).

In aller Regel ist die Angabe von Prozentsatzdifferenzen, Odds Ratios oder (unter bestimmten Umständen) relativen Risiken vorzuziehen.

Assoziationsmaße für ordinalskalierte Merkmale I

Bei ordinalskalierten Merkmalen kann man von der Richtung eines Zusammenhanges sprechen. Beispiel: Salz im Mensaessen und Geschmacksbewertung (fiktive Daten; absolute Zahlen).

„Je mehr X, desto mehr Y“ → positiver Zusammenhang:

	kein	wenig	mittel	viel
lecker	10	8	6	4
sehr lecker	8	10	6	4
sensationell	4	6	8	10

„Je mehr X, desto weniger Y“ → negativer Zusammenhang:

	kein	wenig	mittel	viel
lecker	4	6	8	10
sehr lecker	8	10	6	4
sensationell	10	8	6	4

Assoziationsmaße für ordinalskalierte Merkmale II

Weil man bei zwei ordinalskalierten Merkmalen die Richtung des Zusammenhangs angeben kann, sollten Assoziationsmaße verwendet werden, die die Richtung durch das Vorzeichen ausdrücken.

Für die folgenden Maße gilt, dass sie zwischen minimal -1 (perfekt negativer Zusammenhang) und maximal $+1$ (perfekt positiver Zusammenhang) liegen können. Der Wert 0 signalisiert einen fehlenden (gerichteten) Zusammenhang.

- Kendalls τ (tau) in den Varianten tau-a, tau-b und tau-c (am häufigsten verwendet: τ-b).
- Goodmans und Kruskals Gamma (wird gerne von Angebern verwendet, weil es große Werte annimmt; leidet an ähnlichem Problem wie Yules Q).
- Somers' d (unterscheidet zwischen unabhängiger und abhängiger Variable, wird aber besonders selten verwendet).

Für die Tabelle auf der vorherigen Folie beträgt Gamma 0,32, tau-b 0,23.

Assoziationsmaße für ordinalskalierte Merkmale III

Leider ist die Erläuterung dieser Assoziationsmaße recht aufwändig. Sie steht in keiner Relation zu ihrer statistischen Bedeutung (sie werden zwar immer wieder eingesetzt, sind jedoch veraltet).

Daher verzichte ich auf diese Erläuterung (wer will, kann bei Kühnel & Krebs nachlesen) und bespreche lieber das Problem der statistischen Absicherung von Zusammenhängen in Kreuztabellen mit zwei ordinalskalierten Merkmalen.

Hat man es mit ordinalskalierten Merkmalen zu tun und nimmt man einen gerichteten Zusammenhang an (je ... desto [weniger]), ist der Signifikanztest für das verwendete Assoziationsmaß dem χ^2-Test vorzuziehen, da letzterer auf Abweichungen von den erwarteten Werten in *beliebiger* Richtung reagiert.

Es kann daher vorkommen, dass der χ^2-Test einen signifikanten Zusammenhang andeutet, obwohl der *theoretisch angenommene* Zusammenhang keineswegs signifikant ist, und umgekehrt.

Signifikanztest bei ordinalskalierten Merkmalen I

Für die oben genannten Assoziationsmaße lassen sich Standardfehler und auf deren Grundlage standardnormalverteilte Teststatistiken berechnen. Es kann *asymptotisch* von einem signifikanten Zusammenhang (bei Signifikanzniveau $\alpha = 0{,}05$) ausgegangen werden, wenn gilt:

$$\left|\frac{\text{Koeff.}}{\text{S.E.}}\right| > 1{,}96 \text{ für } \mathrm{H_A}: \quad \text{Koeff.} \neq 0$$

$$\frac{\text{Koeff.}}{\text{S.E.}} > 1{,}645 \text{ für } \mathrm{H_A}: \quad \text{Koeff.} > 0$$

$$\frac{\text{Koeff.}}{\text{S.E.}} < -1{,}645 \text{ für } \mathrm{H_A}: \quad \text{Koeff.} < 0$$

Koeff. = Koeffizient (jeweiliges Assoziationsmaß); S. E. = Standardfehler. „Asymptotisch“ heißt: Die Gültigkeit der Aussagen steigt mit zunehmender Stichprobengröße.

Signifikanztest bei ordinalskalierten Merkmalen II

Beispiel (fiktiv):

Der Zusammenhang zwischen Salz im Mensaessen und Geschmack ist nach dem χ^2-Test nicht signifikant ($\chi^2 = 7{,}945$, krit. Wert 12,59 [bei $\alpha = 0{,}05$]) – aber die Assoziationsmaße für ordinalskalierte Merkmale sind sämtlich positiv und bei $\alpha = 0{,}05$ statistisch signifikant von 0 verschieden (für H_A: Koeff. $\neq 0$ bzw. H_A: Koeff. > 0).

	kein	wenig	mittel	viel
lecker	10	8	6	4
sehr lecker	8	10	6	4
sensationell	4	6	8	10

Gamma: 0,32, S. E. = 0,125 $\rightarrow$ Koeff./S. E. > 2

tau-b: 0,23, S. E. = 0,092 $\rightarrow$ Koeff./S. E. > 2

Signifikanztest bei ordinalskalierten Merkmalen III

Alternatives Beispiel (fiktiv):

Der Zusammenhang zwischen Salz im Mensaessen und Geschmack ist nach dem χ^2-Test klar signifikant ($\chi^2 = 218$, krit. Wert 12,59 [$\alpha = 0{,}05$]) – aber die Assoziationsmaße für ordinalskalierte Merkmale betragen 0 (und unterscheiden sich somit a fortiori nicht signifikant von 0).

	kein	wenig	mittel	viel
lecker	80	10	10	80
sehr lecker	10	10	10	10
sensationell	10	80	80	10

Die Tabelle zeigt absolute Häufigkeiten.