Statistik II im Wintersemester 2006/2007

Themen am 7.11.

- Drittvariablenkontrolle in der Tabellenanalyse
- Korrelation und Kausalität
- Statistische Tests in dreidimensionalen Kreuztabellen

Lernziele:

1. Aufbau einer dreidimenionalen Kreuztabellen, Partialtabellen
2. Ursachen von Unterschieden zwischen bivariaten und konditionalen Zusammenhängen: Konfundierung, Suppression, Verzerrung, Scheinbare Nichtbeziehung, Scheinkausalität Mediatoreffekt, Interaktionseffekt (Moderatoreffekt)
3. Prüfung der Unabhängigkeit von drei Variablen und Unabhängigkeit einer Variablen von zwei weiteren Variablen

Wiederholung vom 31.10:

Test von Mittelwertdifferenzen in abhängigen Stichproben

Methode A: Berechung einer neuen Variable $D = X_1 - X_2$ und Test einer Hypothese über den Populationsmittelwert μ_D von D.

Methode B: Anwendung der generellen Teststatistik:

$$Z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \mu}{\hat{\sigma}\left(\bar{x}_1 - \bar{x}_2\right)}$$

mit: $$\hat{\sigma}\left(\bar{x}_1 - \bar{x}_2\right) = \sqrt{\frac{\hat{\sigma}^2\left(X_1\right) + \hat{\sigma}^2\left(X_2\right) - 2 \cdot \hat{\sigma}\left(X_1, X_2\right)}{n}} = \sqrt{\frac{s^2\left(X_1\right) + s^2\left(X_2\right) - 2 \cdot s\left(X_1, X_2\right)}{n-1}}$$

Tests von Anteilsdifferenzen:

$$Z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \mu}{\hat{\sigma}\left(\bar{x}_1 - \bar{x}_2\right)} = \frac{\left(p_1 - p_2\right) - \pi}{\hat{\sigma}\left(p_1 - p_2\right)}$$

Wiederholung vom 31.10:

1. Unabhängige Stichproben, postulierte Anteilsdifferenz $\pi \neq 0$:

$$Z = \frac{(p_1 - p_2) - \pi}{\sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}}}$$

2. Unabhängige Stichproben, postulierte Anteilsdifferenz $\pi = 0$:

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}_{pooled} \cdot (1-\hat{\pi}_{pooled}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{p_1 - p_2}{\sqrt{\frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2} \cdot \left(1 - \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. Abhängige Stichproben, postulierte Anteilsdifferenz $\pi = 0$:

	$X_1=1$	$\mathbf{X_1}$	$X_1=0$	
Schwangerschaftsabbruch	... bei finanzieller Notlage sollte			insgesamt
$\mathbf{X_2}$ ... wenn die Frau es will	... erlaubt sein		... verboten sein	
$X_2=1$ ... solte erlaubt sein	41.7% (1197) = a		5.3% (151) = b	46.9% (1348) p_2
$X_2=0$ ... sollte verboten sein	18.1% (519) = c		35.0% (1006) = d	53.1% (1525)
insgesamt	59.7% (1716) p_1		40.3% (1157)	100.0% (2875)

$$p_1 - p_2 = (a+c)/n - (a+b)/n = (c-b)/n$$

Wiederholung vom 31.10:

Schwangerschaftsabbruch	$X_1=1$ ... bei finanzieller Notlage sollte ... erlaubt sein	$X_1=0$ ... verboten sein	insgesamt	
X_2 ... wenn die Frau es will				
$X_2=1$ ... solte erlaubt sein	41.7% (1197) = a	5.3% (151) = b	46.9% (1348)	p_2
$X_2=0$ ... sollte verboten sein	18.1% (519) = c	35.0% (1006) = d	53.1% (1525)	
insgesamt	59.7% (1716) p_1	40.3% (1157)	100.0% (2875)	

$$p_1 - p_2 = (a+c)/n - (a+b)/n = (c-b)/n$$

$$Z = \frac{p_{D=1} - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n_D}}} = \frac{c/(b+c) - 0.5}{\sqrt{\frac{0.25}{b+c}}}$$

4. Abhängige Stichproben, postulierte Anteilsdifferenz $\pi \neq 0$:

$$Z = \frac{(p_1 - p_2) - \pi}{\hat{\sigma}(p_1 - p_2)}$$

$$\hat{\sigma}(p_1 - p_2) = \left(\frac{p_1 \cdot (1-p_1)}{n} + \frac{p_2 \cdot (1-p_2)}{n} - 2 \cdot \frac{p_{11} \cdot (1-p_1) \cdot (1-p_2) - p_{12} \cdot p_1 \cdot (1-p_2) - p_{21} \cdot (1-p_1) \cdot p_2 + p_{22} \cdot p_1 \cdot p_2}{n} \right)^{0.5}$$

Drittvariablenkontrolle in der Tabellenanalyse

Mit der Tabellenanalyse können Zusammenhänge zwischen zwei nominal- oder ordinal-skalierten Variablen mit nicht zu vielen Ausprägungen analysiert werden.
Bei einer kausalen Interpretation allerdings die Gefahr von Fehlinterpretationen, wie das folgende empirische Beipsiel zeigt.

Bewertung von Schwangerschaftsabbrüchen in Abhängigkeit von Telefonanschluss im Haushalt

Abtreibung, wenn die Frau es will, ...	Telefonanschluss im Haushalt? ja	nein
... sollte verboten sein	54.7%	33.0%
... sollte erlaubt sein	45.3%	67.0%
	(2331)	(782)

(Quelle: ALLBUS 1992)

In Haushalten, die 1992 über einen Telefonanschluss verfügten, war der Anteil der Abtreibungsgegner um 21.7 (54.7% – 33.0%) Prozentpunkte höher als in Haushalten, die über kein Telefon verfügen!

Es erscheint nciht sehr plausibel, dass das Verfügen über einen Telefonanschluss dazu führt, dass die Haltung zu Schwangerschaftsabbrüchen regider wird.
Tatsächlich handelt es sich bei der beobachteten empirischen Beziehung um ein Artefakt, dass dadurch zustande kommt, dass 1992 sowohl die Telefondichte in den alten und den neuen Bundesländern, als auch die Haltung zu Schwangerschaftsabbrüchen sehr unterschiedlich war.

Drittvariablenkontrolle in der Tabellenanalyse

Deutlich wird dies, wenn der Zusammenhang für die alten und neuen Bundesländer getrennt betrachtet wird.

Bewertung von Schwangerschaftsabbrüchen in Abhängigkeit von Telefonanschluss im Haushalt

	Alte Länder Telefonanschluss im Haushalt?		Neue Länder Telefonanschluss im Haushalt	
Abtreibung, wenn die Frau es will, ...	ja	nein	ja	nein
... sollte verboten sein	58.5%	62.8%	28.9%	29.7%
... sollte erlaubt sein	41.5%	37.2%	71.1%	70.3%
	(2026)	(78)	(305)	(704)

(Quelle: ALLBUS 1992)

Die gleichzeitige Betrachtung von drei Variablen (Telefonbesitz X, Haltung zu Schwangerschaftsabbrüchen (Y) und Region (Z) ermöglicht es, die Fehlinterpretation des bivariaten Zusammenhangs zu vermeiden.
Die sogenannte ***Drittvariablenkontrolle*** ist daher für statistische Zusammenhnagsanalysen äußerst wichtig.
In der Drittvariablenkontrolle, werden mehrdimensionale Kreuztabellen analysiert.

Aufbau einer trivariaten Kreuztabelle

Die ***trivariate Kreuztabelle*** zeigt die gemeinsame Verteilung von drei Variablen. Die Ausprägungen der dritten Variablen definieren ***Partialtabellen***, d.h. bivariate Tabellen bei vorgegebenem Wert einer dritten Variablen, dessen Ausprägung für alle Fälle der Partialtabelle ***konstant*** ist. Drittvariablenkontrolle wird daher oft mit ***Konstanthalten der Ausprägung einer Drittvariablen*** gleichgesetzt.

	Z															
	z_1					z_2					....	z_K				
	X					X						X				
Y	x_1	x_2	...	x_J	Σ	x_1	x_2	...	x_J	Σ		x_1	x_2	...	x_J	Σ
y_1	n_{111}	n_{121}	...	n_{1I1}	$n_{1\bullet1}$	n_{112}	n_{122}	...	n_{1I2}	$n_{1\bullet2}$	...	n_{11K}	n_{12K}	...	n_{1IK}	$n_{1\bullet K}$
y_2	n_{211}	n_{221}	...	n_{2I1}	$n_{2\bullet1}$	n_{212}	n_{222}	...	n_{2I2}	$n_{2\bullet2}$	...	n_{21K}	n_{22K}	...	n_{2IK}	$n_{2\bullet K}$
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Y_I	n_{I11}	n_{I21}	...	n_{IJ1}	$n_{I\bullet1}$	n_{I12}	n_{I22}	...	n_{II2}	$n_{I\bullet2}$	...	n_{I1K}	n_{I2K}	...	n_{IIK}	$n_{I\bullet K}$
Σ	$n_{\bullet11}$	$n_{\bullet21}$	...	$n_{\bullet J1}$	$n_{\bullet\bullet1}$	$n_{\bullet12}$	$n_{\bullet22}$	...	$n_{\bullet I2}$	$n_{\bullet\bullet2}$	...	$n_{\bullet1K}$	$n_{\bullet2K}$	...	$n_{\bullet IK}$	$n_{\bullet\bullet K}$

Die formale Kennzeichnung der absoluten Häufigkeiten erfolgt in einer trivariaten Tabelle analog zu der in bivariaten Tabellen:

- „n“ steht für die absoluten Häufigkeiten;
- die ersten beiden Indizes von n stehen für die Ausprägungen der Zeilenvariable (X) und der Spaltenvariable (Y) in einer Partialtabelle;
  der zusätzliche dritte Index steht für die Ausprägung der ***Dritt-*** oder ***Kontrollvariable*** (Z).

Randtabellen

Bivariate Tabellen ergeben sich aus der trivariaten Kreuztabelle analog der Bildung univariater Randverteilung bei bivariaten Tabellen durch ***Aggregation*** (Aufsummieren) über die Werte einer der drei Variablen und werden daher als ***Randtabellen*** bezeichnet.

Als Beispiel wird der Zusammenhang zischen der Bewertung der eigenen und der allgemeinen Wirtschafslage bei Kontrolle der Erwerbstätigkeit betrachtet:

Trivariate Kreuztabelle:

	Erwerbstätigkeit (Z)			
	ja (z_1)		nein (z_2)	
	Allgemeine Wirtschaftslage (X)		Allgemeine Wirtschafslage (X)	
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	gut (x_1)	nicht gut (x_2)
gut (y_1)	$n_{111} = 170$	$n_{121} = 751$	$n_{112} = 168$	$n_{122} = 614$
nicht gut (y_2)	$n_{211} = 72$	$n_{221} = 928$	$n_{212} = 54$	$n_{222} = 732$

(Quelle: Allbus 1996)

Randtabelle Y nach X durch Aggregation über Kontrollvariable (Z):

	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Summe
gut (y_1)	$n_{11-} = 170 + 168 = 338$	$n_{12-} = 751 + 614 = 1365$	$n_{1++} = 1703$
nicht gut (y_2)	$n_{21-} = 72 + 54 = 126$	$n_{22-} = 928 + 732 = 1660$	$n_{2++} = 1786$
Summe	$n_{+1+} = 464$	$n_{+2+} = 3025$	$n_{+++} = 3489$

Randtabellen

Trivariate Kreuztabelle:

	Erwerbstätigkeit (Z)			
	ja (z_1)		nein (z_2)	
	Allgemeine Wirtschaftslage (X)		Allgemeine Wirtschafslage (X)	
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	gut (x_1)	nicht gut (x_2)
gut (y_1)	$n_{111} = 170$	$n_{121} = 751$	$n_{112} = 168$	$n_{122} = 614$
nicht gut (y_2)	$n_{211} = 72$	$n_{221} = 928$	$n_{212} = 54$	$n_{222} = 732$

(Quelle: Allbus 1996)

Randtabelle Y nach Z durch Aggregation über Spaltenvariable (X):

	Erwerbstätigkeit (Z)		
Eigene Wirtschaftslage (Y)	ja (z_1)	nein (z_2)	Summe
gut (y_1)	$n_{1\text{-}1} = 170 + 751 = 921$	$n_{1\text{-}2} = 168 + 614 = 782$	$n_{++1} = 1703$
nicht gut (y_2)	$n_{2\text{-}1} = 72 + 928 = 1000$	$n_{2\text{-}2} = 54 + 732 = 786$	$n_{++2} = 1786$
Summe	$n_{+1+} = 1921$	$n_{+2+} = 1568$	$n_{+++} = 3489$

Randtabelle X nach Z durch Aggregation über Zeilenvariable (Y):

	Erwerbstätigkeit (Z)		
Allgem. Wirtschaftslage (X)	ja (z_1)	nein (z_2)	Summe
gut (x_1)	$n_{+11} = 170 + 72 = 242$	$n_{+12} = 168 + 54 = 222$	$n_{+1+} = 464$
nicht gut (x_2)	$n_{+21} = 751 + 928 = 1679$	$n_{+22} = 614 + 732 = 1346$	$n_{+2+} = 3025$
Summe	$n_{++1} = 1921$	$n_{++2} = 1568$	$n_{+++} = 3489$

Anteile und Prozentuierungen

Für die Interpretation werden üblicherweise Anteilen bzw. Prozentwerte berechnet.
Eher selten werden allerdings Anteile bezogen auf die gesamte dreidimensionale Kreuztabelle berechnet.

$$p_{ijk} = \frac{n_{ijk}}{n_{...}};\ p_{ij.} = \frac{n_{ij.}}{n_{...}};\ p_{i.k} = \frac{n_{i.k}}{n_{...}};\ p_{.jk} = \frac{n_{.jk}}{n_{...}};\ p_{i..} = \frac{n_{i..}}{n_{...}};\ p_{.j.} = \frac{n_{.j.}}{n_{...}};\ p_{..k} = \frac{n_{..k}}{n_{...}}$$

Eher beziehen sich Anteile auf die Partialtabellen: $p_{ij(k)} = \frac{p_{ijk}}{p_{..k}} = \frac{n_{ijk}}{n_{..k}}$

nur Erwerbstätige (z_1)	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	$p_{11(1)}$ = 170/1921 = 0.088	$p_{12(1)}$ = 751/1921 = 0.391	$p_{1+(1)}$ = 921/1921 = 0.479
nicht gut (y_2)	$p_{21(1)}$ = 72/1921 = 0.037	$p_{22(1)}$ = 928/1921 = 0.483	$p_{2+(1)}$ = 1000/1921 = 0.521
Total	$p_{+1(1)}$ = 0.126	$p_{+2(1)}$ = 0.874	$p_{++(1)}$ = 1.000

nur Nichterwerbstätige (z_2)	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	$p_{11(2)}$ = 168/1568 = 0.107	$p_{12(2)}$ = 614/1568 = 0.392	$p_{1+(2)}$ = 782/1568 = 0.499
nicht gut (y_2)	$p_{21(2)}$ = 54/1568 = 0.034	$p_{22(2)}$ = 732/1568 = 0.467	$p_{2+(2)}$ = 768/1568 = 0.501
Total	$p_{+1(2)}$ = 0.142	$p_{+2(2)}$ = 0.858	$p_{++(2)}$ = 1.000

Anteile und Prozentuierungen

In der Regel werden bedingte relative Häufigkeiten in den Partialtabellen berechnet.
Zur Unterschiedung von bedingender Variable und Kontrollvariable wird zwischen erklärender Variablen und Kontrollvariablen ein Punkt gesetzt: $p_{Y(X.Z)}$ bezeichnet die bedingten Anteile von Y gegeben X in den durch die Ausprägungen von Z definierten Partialtabellen

$$p_{i(j.k)} = \frac{p_{ijk}}{p_{jk}} = \frac{n_{ijk}}{n_{jk}}$$

nur Erwerbstätige (z_1)	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	$p_{1(1.1)}$ = 170/242 = 0.702	$p_{1(2.1)}$ = 751/1679 = 0.447	$p_{1(+.1)}$ = 921/1921 = 0.479
nicht gut (y_2)	$p_{2(1.1)}$ = 72/242 = 0.298	$p_{2(2.1)}$ = 928/1679 = 0.553	$p_{2(+.1)}$ = 1000/1921 = 0.521
Total	$p_{+(1.1)}$ = 1.000 (242)	$p_{+(2.1)}$ = 1.000(1679)	$p_{+(+.1)}$ = 1.000 (1921)

nur Nichterwerbstätige (z_2)	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	$p_{1(1.2)}$ = 168/222 = 0.757	$p_{1(2.2)}$ = 614/1346 = 0.456	$p_{1(+.2)}$ = 782/1568 = 0.499
nicht gut (y_2)	$p_{2(1.2)}$ = 54/222 = 0.243	$p_{2(2.2)}$ = 732/1346 = 0.544	$p_{2(+.2)}$ = 786/1568 = 0.501
Total	$p_{+(1.2)}$ = 1.000 (222)	$p_{+(2.2)}$ = 1.000 (1346)	$p_{+(+.2)}$ = 1.000 (1568)

Konditionale Zusammenhangsmaße

Zusammenhangsmaße in Partialtabellen geben die bedingten Zusammenhänge bei gegebenem Wert der Kontrollvariable an und sind daher ***konditionale Zusammenhangsmaße***.

Nur Erwerbstätige (z_1)	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	70.2% (170)	44.7% (751)	47.9% (921)
nicht gut (y_2)	29.8% (72)	55.3% (928)	52.1% (1000)
	(242)	(1679)	(1921)

$d_{YX(Z=1)}\% = 25.5$ Punkte

$\Phi_{YX(Z=1)} = 0.169$

Nur Nichterwerbstätige (z_2)	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	75.7% (168)	45.6% (614)	49.9% (782)
nicht gut (y_2)	24.3% (54)	54.4% (732)	50.1% (786)
	(222)	(1346)	(1568)

$d_{YX(Z=2)}\% = 30.1$ Punkte

$\Phi_{YX(Z=2)} = 0.210$

Alle	Allgemeine Wirtschaftslage (X)		
Eigene Wirtschaftslage (Y)	gut (x_1)	nicht gut (x_2)	Total
gut (y_1)	72.8% (338)	45.1% (1365)	48.8% (1703)
nicht gut (y_2)	27.2% (126)	54.9% (1660)	51.2% (1786)
	(464)	(3025)	(3489)

$d_{YX}\% = 27.7$ Punkte

$\Phi_{YX} = 0.188$

Unterschiede zwischen konditionalen und bivariaten Effekten

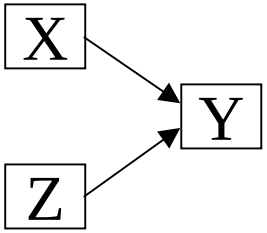
Der Vergleich von konditionalen Effekten bei Drittvariablenkontrolle und korrespondierenden bivariaten Effekten in Partialtabellen kann zu ganz unterschiedlichen Ergebnissen führen:

- Verglichen mit der bivariaten Beziehung kann ein konditionaler Zusammenhang geringer ausfallen als der korrespondierende bivariate Zusammenhang.
- Verglichen mit der bivariaten Beziehung kann ein konditionaler Zusammenhang ganz verschwinden.
- Das Vorzeichen der Beziehung zwischen abhängiger und unabhängier Variable kann sich bei konditionaler und bivariater Berechnung umdrehen.
- Die konditionalen Beziehungen in den Partialtabellen können sich unterscheiden.
- Verglichen mit der bivariaten Beziehung kann ein konditionaler Zusammenhang größer ausfallen als der korrespondierende bivariate Zusammenhang.
- Obwohl bivariat keine Beziehung besteht, zeigt sich in den Partialtabellen eine Beziehung zwischen abhängiger und erklärender Variable.

Ursache dieser Differenzen sind verschiedenen Möglichkeiten von Kausalbeziehungen zwischen drei Variablen.
Im folgenden werden solche Möglichkeiten jeweils an einem fiktiven Datenbeispiel demonstriert.

Additive Effekte bei unkorrelierten erklärenden Variablen

Sowohl X als auch Z wirken jeweils für sich auf Y.
Bei unkorrelierten, additiven Effekten sind bivariate und konditionale Prozentsatzdifferenzen gleich groß.



	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	62% (155)	46% (115)	54% (135)	38% (95)
y_2	38% (95)	54% (135)	46% (115)	62% (155)
Total	(250)	(250)	(250)	(250)

$d_{Y.X(Z=1)}\% = 16.0$ $d_{Y.X(Z=2)}\% = 16.0$
$d_{X.Y(Z=1)}\% = 16.1$ $d_{X.Y(Z=2)}\% = 16.1$
$\Phi_{XY(Z=1)} = 0.161$ $\Phi_{XY(Z=2)} = 0.161$

	X		
Y	x_1	x_2	Total
y_1	58% (290)	42% (210)	50% (500)
y_2	42% (210)	58% (290)	50% (500)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = 16.0$, $d_{X.Y}\% = 16.0$
$\Phi_{XY} = 0.160$

	Z		
Y	z_1	z_2	Total
y_1	54% (270)	46% (230)	50% (500)
y_2	46% (230)	54% (270)	50% (500)
Total	(500)	(500)	(1000)

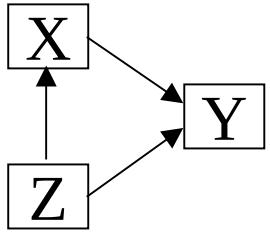
$d_{Y.Z}\% = 8.0$, $d_{Z.Y}\% = 8.0$
$\Phi_{YZ} = 0.080$

	Z		
X	z_1	z_2	Total
x_1	50% (250)	50% (250)	50% (500)
x_2	50% (250)	50% (250)	50% (500)
Total	(500)	(500)	(1000)

$d_{X.Z}\% = 0$, $d_{Z.X}\% = 0$
$\Phi_{XZ} = 0$

Additive Effekte bei korrelierten erklärenden Variablen: Konfundierung

Eine bivariater Effekt von X auf Y ist ***konfundiert***, wenn er aufgrund einer Beziehung zwischen Kontrollvariable und erklärender Variable andere (und sehr oft höhere) Werte aufweist als die korrespondierenden konditionalen Effekte.



	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	80% (320)	70% (70)	40% (40)	30% (120)
y_2	20% (80)	30% (30)	60% (60)	70% (280)
Total	(400)	(100)	(100)	(400)

$d_{Y.X(Z=1)}\% = 10.0$ $\quad$ $d_{Y.X(Z=2)}\% = 10.0$

$d_{X.Y(Z=1)}\% = 9.3$ $\quad$ $d_{X.Y(Z=2)}\% = 7.3$

$\Phi_{XY(Z=1)} = 0.097$ $\quad$ $\Phi_{XY(Z=2)} = 0.857$

	X		
Y	x_1	x_2	Total
y_1	72% (360)	38% (190)	55% (550)
y_2	28% (140)	62% (310)	45% (450)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = 34.0$, $d_{X.Y}\% = 34.3$

$\Phi_{XY} = 0.342$

	Z		
Y	z_1	z_2	Total
y_1	78% (390)	32% (160)	55% (550)
y_2	22% (110)	68% (340)	45% (450)
Total	(500)	(500)	(1000)

$d_{Y.Z}\% = 46.0$, $d_{Z.Y}\% = 46.5$

$\Phi_{YZ} = 0.462$

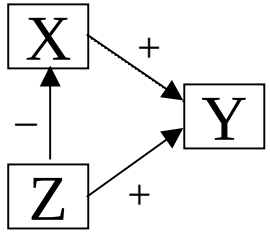
	Z		
X	z_1	z_2	Total
x_1	80% (400)	20% (100)	50% (500)
x_2	20% (100)	80% (400)	50% (500)
Total	(500)	(500)	(1000)

$d_{X.Z}\% = 60.0$, $d_{Z.X}\% = 60.0$

$\Phi_{XZ} = 0.600$

Additive Effekte bei korrelierten erklärenden Variablen: Scheinbare Nichtbeziehung

Zwischen X und Y besteht eine ***scheinbare Nichtbeziehung***, wenn erst durch die Drittvariablenkontrolle eine Beziehung sichtbar wird.



	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	78% (117)	62% (217)	38% (133)	22% (33)
y_2	22% (33)	38% (133)	62% (217)	78% (117)
Total	(150)	(350)	(350)	(150)

$d_{Y.X(Z=1)}\% = 16.0$ $\quad d_{Y.X(Z=2)}\% = 16.0$

$d_{X.Y(Z=1)}\% = 15.2$ $\quad d_{X.Y(Z=2)}\% = 15.2$

$\Phi_{XY(Z=1)} = 0.156$ $\quad \Phi_{XY(Z=2)} = 0.156$

	X		
Y	x_1	x_2	Total
y_1	50% (250)	50% (250)	50% (500)
y_2	50% (250)	50% (250)	50% (500)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = 0$, $d_{X.Y}\% = 0$

$\Phi_{XY} = 0$

	Z		
Y	z_1	z_2	Total
y_1	66.8% (334)	33.2% (166)	50% (500)
y_2	33.2% (166)	66.8% (334)	50% (500)
Total	(500)	(500)	(1000)

$d_{Y.Z}\% = 33.6$, $d_{Z.Y}\% = 33.6$

$\Phi_{YZ} = 0.336$

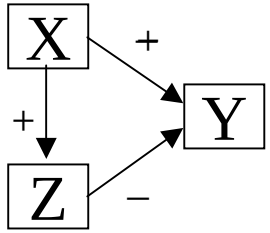
	Z		
X	z_1	z_2	Total
x_1	30% (150)	70% (350)	50% (500)
x_2	70% (350)	30% (150)	50% (500)
Total	(500)	(500)	(1000)

$d_{X.Z}\% = -40.0$, $d_{Z.X}\% = -40.0$

$\Phi_{XZ} = -0.400$

Additive Effekte bei korrelierten erklärenden Variablen: Verzerrung

Die bivariate Beziehung zwischen X und Y ist ***verzerrt***, wenn das Vorzeichen der bivariaten Beziehung umgekehrt zu den konditionalen Beziehungen ist.



	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	40% (160)	30% (45)	80% (80)	70% (245)
y_2	60% (240)	20% (105)	20% (20)	30% (105)
Total	(400)	(150)	(100)	(350)

$d_{Y.X(Z=1)}\% = 10.0$ $d_{Y.X(Z=2)}\% = 10.0$
$d_{X.Y(Z=1)}\% = 8.5$ $d_{X.Y(Z=2)}\% = 8.6$
$\Phi_{XY(Z=1)} = 0.092$ $\Phi_{XY(Z=2)} = 0.093$

	X		
Y	x_1	x_2	Total
y_1	48% (240)	58% (290)	53% (530)
y_2	52% (260)	42% (210)	47% (470)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = -10.0$, $d_{X.Y}\% = -10.0$
$\Phi_{XY} = -0.100$

	Z		
Y	z_1	z_2	Total
y_1	37.3% (205)	72.2% (325)	53% (530)
y_2	62.7% (345)	27.8% (125)	47% (470)
Total	(550)	(450)	(1000)

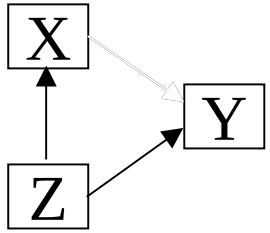
$d_{Y.Z}\% = -34.9$, $d_{Z.Y}\% = -34.7$
$\Phi_{YZ} = -0.348$

	X		
Z	x_1	x_2	Total
z_1	80% (400)	30% (150)	50% (550)
z_2	20% (100)	70% (350)	50% (450)
Total	(500)	(500)	(1000)

$d_{Z.X}\% = 50.0$, $d_{X.Z}\% = 50.5$
$\Phi_{XZ} = 0.503$

Scheinkausalität

Wenn eine bivariate Beziehung zwischen zwei Variablen X und Y dadurch hervorgerufen wird, dass eine Drittvariable auf X und Y wirkt, dann liegt eine ***Scheinkausalität*** vor.



	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	70% (280)	70% (70)	30% (30)	30% (120)
y_2	30% (120)	30% (30)	70% (70)	70% (280)
Total	(400)	(100)	(100)	(400)

$d_{Y.X(Z=1)}\% = 0$ $\quad$ $d_{Y.X(Z=2)}\% = 0$
$d_{X.Y(Z=1)}\% = 0$ $\quad$ $d_{X.Y(Z=2)}\% = 0$
$\Phi_{XY(Z=1)} = 0$ $\quad$ $\Phi_{XY(Z=2)} = 0$

	X		
Y	x_1	x_2	Total
y_1	62% (310)	38% (190)	50% (500)
y_2	38% (190)	62% (310)	50% (500)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = 24.0$, $d_{X.Y}\% = 24.0$
$\Phi_{XY} = 0.240$

	Z		
Y	z_1	z_2	Total
y_1	70% (350)	30% (150)	50% (500)
y_2	30% (150)	70% (350)	50% (500)
Total	(500)	(500)	(1000)

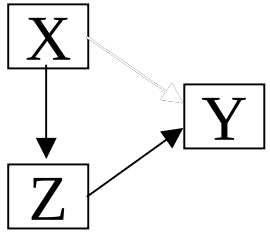
$d_{Y.Z}\% = 40.0$, $d_{Z.Y}\% = 40.0$
$\Phi_{YZ} = 0.400$

	Z		
X	z_1	z_2	Total
x_1	80% (400)	20% (100)	50% (500)
x_2	20% (100)	80% (400)	50% (500)
Total	(500)	(500)	(1000)

$d_{X.Z}\% = 60.0$, $d_{Z.X}\% = 60.0$
$\Phi_{XZ} = 0.600$

Mediation über eine intervenierende Variable

Wenn der Effekt einer erklärenden Variable über eine dritte ***intervenierende Variable*** (auch als ***Mediator*** bezeichnet) vermittelt wird, spricht man von Mediation.



	$Z = z_1$ X		$Z = z_2$ X	
Y	x_1	x_2	x_1	x_2
y_1	70% (280)	70% (105)	20% (20)	20% (70)
y_2	30% (120)	30% (45)	80% (80)	80% (280)
Total	(400)	(150)	(100)	(350)

$d_{Y.X(Z=1)}\% = 0$ $\quad$ $d_{Y.X(Z=2)}\% = 0$

$d_{X.Y(Z=1)}\% = 0$ $\quad$ $d_{X.Y(Z=2)}\% = 0$

$\Phi_{XY(Z=1)} = 0$ $\quad$ $\Phi_{XY(Z=2)} = 0$

	X		
Y	x_1	x_2	Total
y_1	60% (300)	35% (175)	47.5% (475)
y_2	40% (200)	65% (325)	52.5% (525)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = 25.0$, $d_{X.Y}\% = 25.1$

$\Phi_{XY} = 0.250$

	Z		
Y	z_1	z_2	Total
y_1	70% (385)	20% (90)	47.5% (475)
y_2	30% (165)	80% (360)	52.5% (525)
Total	(550)	(450)	(1000)

$d_{Y.Z}\% = 50.0$, $d_{Z.Y}\% = 49.6$

$\Phi_{YZ} = 0.498$

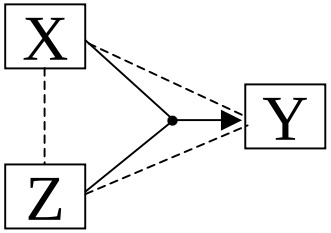
	X		
Z	x_1	x_2	Total
z_1	80.0% (400)	30.0% (150)	55% (550)
z_2	20.0% (100)	70.0% (350)	45% (450)
Total	(500)	(500)	(1000)

$d_{Z.X}\% = 50.0$, $d_{X.Z}\% = 50.5$

$\Phi_{XZ} = 0.503$

Interaktionseffekt

Ein Interaktionseffekt liegt vor, wenn sich die konditionalen Effekte bei verschiedenen Ausprägungen einer Drittvariable unterscheiden.



	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	70% (175)	50% (125)	50% (125)	70% (175)
y_2	30% (75)	50% (125)	50% (125)	30% (75)
Total	(250)	(250)	(250)	(250)

$d_{Y.X(Z=1)}\% = 20.0$ $\quad d_{Y.X(Z=2)}\% = -20.0$

$d_{X.Y(Z=1)}\% = 20.8$ $\quad d_{X.Y(Z=2)}\% = -20.8$

$\Phi_{XY(Z=1)} = 0.204$ $\quad \Phi_{XY(Z=2)} = 0.204$

	X		
Y	x_1	x_2	Total
y_1	60% (300)	60% (300)	60% (600)
y_2	40% (200)	40% (200)	40% (400)
Total	(500)	(500)	(1000)

$d_{Y.X}\% = 0$, $d_{X.Y}\% = 0$

$\Phi_{XY} = 0$

	Z		
Y	z_1	z_2	Total
y_1	60% (300)	60% (300)	60% (600)
y_2	40% (200)	40% (200)	40% (400)
Total	(500)	(500)	(1000)

$d_{Y.Z}\% = 0$, $d_{Z.Y}\% = 0$

$\Phi_{YZ} = 0$

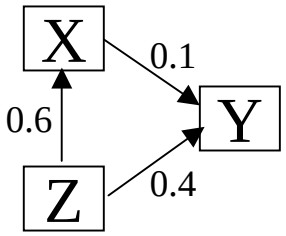
	Z		
X	z_1	z_2	Total
x_1	50% (250)	50% (250)	50% (500)
x_2	50% (250)	50% (250)	50% (500)
Total	(500)	(500)	(1000)

$d_{X.Z}\% = 0$, $d_{Z.X}\% = 0$

$\Phi_{YZ} = 0$

Direkte, indirekte, korrelierte und totale Effekte

Die unterschiedlichen Beziehungsmuster lassen sich leichter nachvollziehen, wenn zwischen direkten und indirekten Beziehungen unterschieden wird.



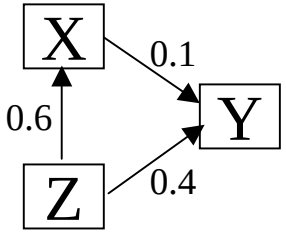
$d_{Y.X(Z=1)} = 0.100$ $d_{Y.Z(X=1)} = 0.400$ $d_{X.Z} = 0.600$ $d_{Y.Z} = 0.460$
$d_{Y.X(Z=2)} = 0.100$ $d_{Y.Z(X=2)} = 0.400$ $d_{Y.X} = 0.340$

Im Beispiel wirkt X auf Y, wobei die Anteilsdifferenzen $d_{Y.X(Z)}$ bei Kontrolle von Z jeweils 0.1 betragen.
Außerdem wirkt auch Z auf Y, wobei die Anteilsdifferenzen $d_{Y.Z(X)}$ bei Kontrolle von X hier jeweils 0.4 betragen.
Schließlich wirkt Z auch auf X, wobei hier die Anteilsdifferenz $d_{X.Z}$ 0.6 beträgt.

Da außer Z keine weitere Variable auf X wirkt, kann hier der bivariate Effekt betrachtet werden.

Alle drei Effekte sind ***direkte Effekte***, da die Erklärungsvariable ohne "Umweg" auf die jeweilige abhängige Variable wirkt.
Im Beispiel hat Z zusätzlich noch einen ***indirekten Effekt*** über X auf Y: Wenn der Wert von Z sich ändert, ändert sich der Wert von X, was wiederum eine Änderung von Y auslöst.
Im additiven Modell von Anteilsveränderungen ist der indirekte Effekt gleich dem Produkt aller direkten Effekte auf dem Pfad von Z über X auf Y, hier also $0.6 \times 0.1 = 0.06$

Direkte, indirekte, korrelierte und totale Effekte



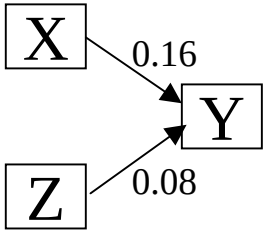
$d_{Y.X(Z=1)} = 0.100$ $d_{Y.Z(X=1)} = 0.400$ $d_{X.Z} = 0.600$ $d_{Y.Z} = 0.460$
$d_{Y.X(Z=2)} = 0.100$ $d_{Y.Z(X=2)} = 0.400$ $d_{Y.X} = 0.340$

Der ***totale Effekt*** von Z auf Y ist die Summe des direkten Effekts und aller indirekten Effekte, im Beispiel also 0.4 + 0.06 = 0.46.
Dieser Wert ist hier gerade gleich dem bivariaten Effekt $d_{Y.Z}$ von Z auf Y.

Da Z sowohl X als auch Y beeinflusst, haben X und Y (zum Teil) eine gemeinsame Ursache, was zu einem Zusammenhang zwischen X und Y führt, der unabhängig von dem Effekt von X auf Y ist. Dieser durch die gemeinsame erklärende Variable Z hervorgerufene Zusammenhang wird ***korrelierter Effekt*** genannt.
Im Beispiel ist der korrelierte Effekt das Produkt des Effekts von Z auf X und von Z auf Y, also 0.6×0.4 = 0.24

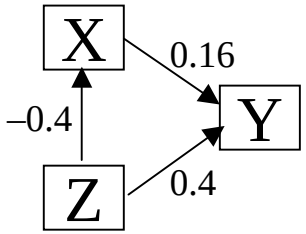
Der bivariate Effekt von X auf Y ist Folge des direkten Effekts von X auf Z plus dem korrelierten Effekt von Z auf X und von Z auf Y, hier also 0.1 + 0.24 = 0.34.
Konfundierung tritt bei additiven Beziehungen immer dann auf, wenn es neben direkten auch indirekte oder korrelierte Effekte gibt.

Direkte, indirekte, korrelierte und totale Effekte

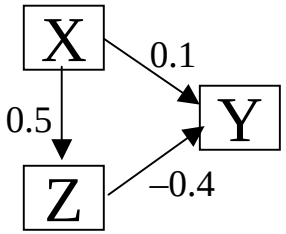


$d_{Y.X(Z=1)} = 0.160$ $d_{Y.Z(X=1)} = 0.080$ $d_{X.Z} = 0$ $d_{Y.Z} = 0.080$
$d_{Y.X(Z=2)} = 0.160$ $d_{Y.Z(X=2)} = 0.080$ $d_{Y.X} = 0.160$

Wenn es weder indirekte noch korrelierte Effekte gibt, sind die bivariaten und die konditinalen Effekte identisch.
Dies ist im Beispiel der additiven Effekte bei unkorrelierten erklärenden Variablen der Fall.



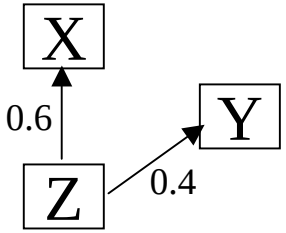
$d_{Y.X(Z=1)} = 0.160$ $d_{Y.Z(X=1)} = 0.400$ $d_{X.Z} = –0.400$ $d_{Y.Z} = 0.336$
$d_{Y.X(Z=2)} = 0.160$ $d_{Y.Z(X=2)} = 0.400$ $d_{Y.X} = 0$



$d_{Y.X(Z=1)} = 0.100$ $d_{Y.Z(X=1)} = –0.400$ $d_{Z.X} = 0.500$ $d_{Y.Z} = –0.349$
$d_{Y.X(Z=2)} = 0.100$ $d_{Y.Z(X=2)} = –0.400$ $d_{Y.X} = –0.100$

Wenn direkte und indirekte Effekte in entgegengesetzte Richtung wirken, sind die bivariaten Effekte geringer als die konditionalen Effekte. Man bezeichnet dies auch als ***Suppression*** des Effekts durch eine *Suppressorvariable*. Als Folge von Suppression kann es bivariate zu einer ***scheinbaren Nichtbeziehung*** oder zu einer ***Verzerrung*** der Beziehungsrichtung kommen.

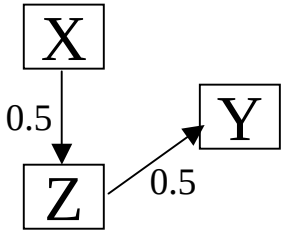
Direkte, indirekte, korrelierte und totale Effekte



$d_{Y.X(Z=1)} = 0$ $d_{Y.Z(X=1)} = 0.400$ $d_{X.Z} = 0.400$ $d_{Y.Z} = 0.400$

$d_{Y.X(Z=2)} = 0$ $d_{Y.Z(X=2)} = 0.400$ $d_{Y.X} = 0.240$

Wenn es weder direkte noch indirekte Effekte zwischen zwei Variablen giebt, sondern nur korrelierte Effekte, dann darf die bivariate Beziehung nicht als kausale Beziehung missverstanden werden. Man spricht daher von ***Scheinkausalität***.

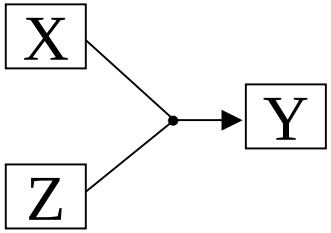


$d_{Y.X(Z=1)} = 0$ $d_{Y.Z(X=1)} = 0.500$ $d_{Z.X} = 0.500$ $d_{Y.Z} = 0.500$

$d_{Y.X(Z=2)} = 0$ $d_{Y.Z(X=2)} = 0.500$ $d_{Y.X} = 0.250$

Wenn es nur indirekte Effekte über eine ***intervenierende Variable*** gibt, liegt ein Mediatoreffekt vor, bei dem der bivariate Effekt durch die Kausalkette der intervenierenden oder mediierenden Variablen interpretiert wird.
Bei Scheinkausalität wie Mediation sind die konditionalen Effekte null, wenn die gemeinsame Ursache bei Scheinkausalität bzw. die intervenierende Variable bei Mediation als Kontrollvariable verwendet wird.

Interaktionseffekte



$d_{Y.X(Z=1)} = +0.200$ $d_{Y.Z(X=1)} = +0.400$ $d_{X.Z} = 0$ $d_{Y.Z} = 0$

$d_{Y.X(Z=2)} = -0.200$ $d_{Y.Z(X=2)} = -0.400$ $d_{Y.X} = 0$

Bei einem Interaktionseffekt wirken zwei (oder mehr) erklärende Variablen gemeinsam auf eine abhängige Variable. Die Effekte sind dann **nicht** additiv. Bei Interaktionseffekten macht es keinen Sinn, vom Effekt einer Variable auf eine andere zu sprechen.
Sinnvoll ist nur die Betrachtung der konditionalen Effekte, im Beispiel von X auf Y bei gegebenen Z bzw. von Z auf Y bei gegebenen Z.
Bei Interaktionseffekten sind die koditionalen Effekte stets verschieden.
Da sich im Beispiel die konditionalen Effekte gegenseitig aufheben, sind die bivariaten Effekt auf Y null. Eine scheinbare Nichtbeziehung kann daher auch Folge von Interaktionseffekten sein.
Da im Beipsiel auch die beiden erklärenden Variablen X und Z unabhängig voneinander sind, zeigt das Beispiel, dass es möglich ist, dass Variablen wechselseitig voneinander unabhängig sind und es gleichwohl eine (nicht additive) Beziehung zwischen ihnen geben kann.

Korrelation und Kausalität

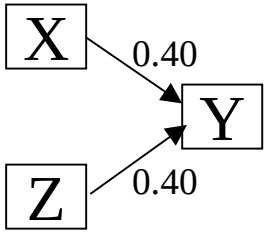
Durch die statistische Analyse von Zusammenhängen einschließlich Drittvariablenkontrolle ist es möglich, die Stärke kausaler Effekte zu untersuchen.
Darüber hianus ist es auch möglich, Konfundierung, Mediation, Scheinkausalitäten, scheinbare Nichtbeziehungen, Verzerrungen und Interaktionseffekte zu analysieren.
Es ist allerdings nicht möglich, ohne zusätzliche Informationen die Kausalrichtung festzustellen. Ob etwa eine Mediation vorliegt oder eine Scheinkausalität, lasst sich an den Daten allein nicht sehen.
Erst wenn durch ein experimentelles oder quasiexperimentelles Design sichergestellt ist, dass eine Kausalrichtung nicht umgekehrt ist, können über statistische Analysen mit einiger Sicherheit auch Effektstärken korrekt erfasst werden.
Es ist allerdings möglich, postulierte kausale Beziehungen auszuschließen, wenn die Datenanalyse zeigt, dass bei Modellierung der postulierten Kausalstruktur, die Effekte nicht in erwarteter Höhe und/oder Richtung auftreten.

Fehlspezifikation

Das eine statistische Analyse bei einer ***Fehlspezifikation*** der Kausalrichtung in die Irre leiten kann, zeigen das folgende Beispiel.
Die Daten sind so generiert, dass zwei statisch unabhängige Variablen X und Z eine abhängige Variable Y beeinflussen.
Wenn fälschlicherweise die Rolle von Kontrollvariable Z und abhängiger Variable Y vertauscht wird, ergeben sich vollkommen falsche Effekte:

	$Z = z_1$		$Z = z_2$	
	X		X	
Y	x_1	x_2	x_1	x_2
y_1	90% (216)	50% (180)	50% (80)	10% (24)
y_2	10% (24)	50% (180)	50% (80)	90% (216)
Total	(240)	(360)	(160)	(240)

$d_{Y.X(Z=1)}\% = 40.0$ $d_{Y.X(Z=2)}\% = 40.0$



tatsächliche Kausalstruktur

	$Y = y_1$		$Y = y_2$	
	X		X	
Z	x_1	x_2	x_1	x_2
z_1	73.0%(216)	88.2%(180)	23.1%(24)	45.5%(180)
z_2	27.0%(80)	11.8%(24)	76.9%(80)	54.5%(216)
Total	(296)	(204)	(104)	(396)

$d_{Z.X(Y=1)}\% = -15.3$ $d_{Z.X(Y=2)}\% = -22.4$

$d_{Z.Y(X=1)}\% = 49.9$
$d_{Z.Y(X=2)}\% = 42.8$

$d_{X.Y}\% = 38.4$
$d_{Y.X}\% = 40.0$
$d_{Z.X}\% = 0$
$d_{X.Z}\% = 0$

Hypothesenprüfung in trivariaten Kreuztabellen

Die Logik des Hypothesentestens kann von Hypothesentests in bivariaten Tabellen auf trivariate Tabellen verallgemeinert werden. Dabei kann die generelle Logik von Pearsons Chiquadrattest bzw. des LR-Tests genutzt werden.

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

Bei statistischer Unabhängigkeit muss gelten:

$$\pi_{ijk} = \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k} \text{ für alle i=1,2,...,I, j=1,2,...,J und k=1,2,...,K}$$

Daraus ergibt sich folgendes Hypothesenpaar:

$$H_0: \pi_{ijk} = \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k} \text{ versus } H_1: \pi_{ijk} \neq \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k}$$

In einfachen Zufallsauswahlen sind die Randanteile $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$ und $p_{\bullet\bullet k}$ konsistente und erwartungstreue Schätzer der entsprechenden Populationsanteile. Die bei statistischer Unabhängigkeit erwarteten Häufigkeiten berechnen sich dann nach:

$$e_{ijk} = n \cdot p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet k} = n \cdot \frac{n_{i\bullet\bullet}}{n} \cdot \frac{n_{\bullet j\bullet}}{n} \cdot \frac{n_{\bullet\bullet k}}{n} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}$$

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

Entsprechend der generellen Formeln für Pearsons Chiquadrattest oder den LR-Test lassen sich dann die Teststatistiken berechnen:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(n_{ijk} - e_{ijk}\right)^2}{e_{ijk}}$$

$$L^2 = 2\cdot\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\cdot\ln\left(\frac{n_{ijk}}{e_{ijk}}\right)$$

Wenn die Nullhypothese richtig ist, sind die beiden Teststatistiken asymptotisch chiquadratverteilt. Die Zahl der Freiheitsgrade ist gleich der Zahl der Tabellenzellen minus der Zahl der geschätzten Modellparameter.
Die Tabelle hat insgesamt I·J·K Tabellenzellen. Für die Berechnung der erwarteten Häufigkeiten werden (I–1)+(J–1)+(K–1) univariate Randwahrscheinlichkeiten plus die Stichprobenfallzahl als empirische Information benötigt.
Dann ergibt sich die Zahl der Freiheitsgrade als:

$$df = I\cdot J\cdot K - (I-1) - (J-1) - (K-1) - 1 = I\cdot J\cdot K - I - J - K + 2$$

Wenn die Nullhypothese falsch ist, sind die beiden Teststatistiken wiederum nichtzentral chiquadratverteilt. Die Nullhypothese wird daher bei einer Irrtumswahrscheinlichkeit von α abgelehnt, wenn die Teststatistik größer ist als das $(1-\alpha)$-Quantil der Chiqdratverteilung mit df = I·J·K–I–J–K+2 Freiheitsgraden.

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

Als Beispiel soll die Unabhängigkeit der allgemeinen Wirtschaftslage, der eigenen Wirtschaftslage und der Erwerbstätigkeit aus Tabelle 15.2 mit einer Irrtumswahrscheinlichkeit von 5% getestet werden.

	Erwerbstätigkeit (Z)			
	ja (z_1)		nein (z_2)	
Eigene Wirtschaftslage (Y)	Allgemeine Wirtschaftsl.(X) gut (x_1)	nicht gut (x_2)	Allgemeine Wirtschafsl. (X) gut (x_1)	nicht gut (x_2)
gut (y_1)	70.2% (170)	44.7% (751)	75.7% (168)	45.6% (614)
nicht gut (y_2)	29.8% (72)	55.3% (928)	24.3% (54)	54.4% (732)

(Quelle: Allbus 1996)

Randverteilungen		
Y	X	Z
1703	464	1921
1786	3025	1568
3489	3489	3489

Zur Berechnung müssen zunächst die erwarteten Häufigkeiten aus den drei Randverteilungen berechnet werden.

$$e_{ijk} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}$$

$$e_{111} = \frac{1703 \cdot 464 \cdot 1921}{3489^2} = 124.70,\ e_{211} = \frac{1786 \cdot 464 \cdot 1921}{3489^2} = 130.78,\ \ldots$$

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	170	124.70	16.46	105.37
ja	gut	nicht gut	72	130.78	26.42	-85.94
ja	nicht gut	gut	751	812.95	4.72	-119.06
ja	nicht gut	nicht gut	928	852.57	6.67	157.34
nein	gut	gut	168	101.78	43.08	168.38
nein	gut	nicht gut	54	106.74	26.06	-73.60
nein	nicht gut	gut	614	663.57	3.70	-95.33
nein	nicht gut	nicht gut	732	695.91	1.87	74.03
Summe			3489	3489	128.98	131.18

Aus den erwarteten und den beobachteten Häufigkeiten können dann die Chiquadratanteile in den einzelnen Zellen berechnet werden. Die Aufsummierung ergibt den Wert der Teststatistik.

$$\chi^2_{111} = \frac{\left(n_{111} - e_{111}\right)^2}{e_{111}} = \frac{\left(170 - 124.70\right)^2}{124.70} = 16.46$$

$$L^2_{111} = 2 \cdot n_{111} \cdot \ln\left(\frac{n_{111}}{e_{111}}\right) = 2 \cdot 170 \cdot \ln\left(\frac{170}{124.70}\right) = 105.37$$

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	170	124.70	16.46	105.37
ja	gut	nicht gut	72	130.78	26.42	-85.94
ja	nicht gut	gut	751	812.95	4.72	-119.06
ja	nicht gut	nicht gut	928	852.57	6.67	157.34
nein	gut	gut	168	101.78	43.08	168.38
nein	gut	nicht gut	54	106.74	26.06	-73.60
nein	nicht gut	gut	614	663.57	3.70	-95.33
nein	nicht gut	nicht gut	732	695.91	1.87	74.03
Summe			3489	3489	128.98	131.18

Die Teststatistiken sind die Summen der Zellenanteile und betragen 128.98 bzw. 131.18.
Die Zahl der Freiheitsgrade ist df = 8–2–2–2+2 = 4.
Bei einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese abgelehnt, wenn die Teststatistik größer ist als das 95%-Quantil der Chiquadratverteilug mit 4 Freiheitsgraden, d.h. größer ist als 9.488.
Da dies der Fall ist, ist die Nullhypothese abzulehnen.Vermutlich sind die drei Variablen nicht statistisch unabhängig voneinander.

Alle erwarteten Häufigkeiten sind größer 5. Daher ist diese Anwendungsvoraussetzung des Chiquadrattests erfüllt.

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Im Beispiel besteht eine Abhängigkeit zwischen den drei Varablen.
Möglicherweise ist aber die Erwerbstätigkeit von den beiden anderen Variablen EWL und AWL unabhängig.
Dies kann wiederum mit einem Chiquadrattest geprüft werden.
Wenn in einer trivariaten Häufigkeitstabelle die Zeilen- und Spaltenvariable von der dritten Variablen statistisch unabhängig sind, muss gelten:

$$\pi_{ijk} = \pi_{ij\bullet} \cdot \pi_{\bullet\bullet k} \text{ für alle } i=1,2,...,I,\ j=1,2,...,J \text{ und } k=1,2,...,K$$

Daraus ergibt sich folgendes Hypothesenpaar:

$$H_0: \pi_{ijk} = \pi_{ij\bullet} \cdot \pi_{\bullet\bullet k} \text{ versus } H_1: \pi_{ijk} \neq \pi_{ij} \cdot \pi_{\bullet\bullet k}$$

Die bei zutreffender Nullhypothese erwarteten Häufigkeiten berechnen sich dann nach:

$$e_{ijk} = n \cdot p_{ij\bullet} \cdot p_{\bullet\bullet k} = n \cdot \frac{n_{ij\bullet}}{n} \cdot \frac{n_{\bullet\bullet k}}{n} = \frac{n_{ij\bullet} \cdot n_{\bullet\bullet k}}{n}$$

Formal entspricht dies der Unabhängigkeit in einer bivariaten Kreuztabelle, bei der die Zeilenvariable aus der Kombination aller Ausprägungen der ersten beiden Variablen der trivariaten Tabelle, im Beispiel aus AWL und EWL gebildet wird. Die Nullhypothese wird daher bei einer Irrtumswahrscheinlichkeit von α abgelehnt, wenn die Teststatistik größer ist als das $(1-\alpha)$-Quantil der Chiqdratverteilung mit df = $(I \cdot J-1) \cdot (K-1)$ Freiheitsgraden.

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

	Erwerbstätigkeit (Z)			
	ja (z_1)		nein (z_2)	
Eigene Wirtschaftslage (Y)	Allgemeine Wirtschaftsl.(X) gut (x_1)	nicht gut (x_2)	Allgemeine Wirtschafsl. (X) gut (x_1)	nicht gut (x_2)
gut (y_1)	70.2% (170)	44.7% (751)	75.7% (168)	45.6% (614)
nicht gut (y_2)	29.8% (72)	55.3% (928)	24.3% (54)	54.4% (732)

(Quelle: Allbus 1996)

Randverteilungen		
Y	X	Z
1703	464	1921
1786	3025	1568
3489	3489	3489

Zur Berechnung der Teststatistik müssen zunächst wieder die erwarteten Häufigkeiten berechnet werden:

Bei Unabhängigkeit erwartete Häufigkeiten (n=3489)				
AWL	gut	gut	nicht gut	nicht gut
EWL	gut	nicht gut	gut	nicht gut
Erwerbstätigkeit	(n=338)	(n=126)	(n=1365)	(n=1660)
ja (n=1921)	186.10	69.37	751.55	913.98
nein (n=1568)	151.90	56.63	613.45	746.02

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Bei Unabhängigkeit erwartete Häufigkeiten (n=3489)				
AWL	gut	gut	nicht gut	nicht gut
EWL	gut	nicht gut	gut	nicht gut
Erwerbstätigkeit	(n=338)	(n=126)	(n=1365)	(n=1660)
ja (n=1921)	186.10	69.37	751.55	913.98
nein (n=1568)	151.90	56.63	613.45	746.02

Aus den erwarteten und den beobachteten Häufigkeiten werden sodann die Chiquadratanteile in den einzelnen Zellen berechnet. Die Aufsummierung ergibt wieder den Wert der Teststatistik.

Berechnung der Teststatistiken χ^2 und L^2						
erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	170	186.10	1.46	-30.76
ja	gut	nicht gut	72	69.37	0.10	5.35
ja	nicht gut	gut	751	751.55	0.00	-1.10
ja	nicht gut	nicht gut	928	913.98	0.22	28.26
nein	gut	gut	168	151.90	1.71	33.85
nein	gut	nicht gut	54	56.63	0.12	-5.13
nein	nicht gut	gut	614	613.45	0.00	1.10
nein	nicht gut	nicht gut	732	746.02	0.26	27.78
Summe			3489	3489	3.80	3.79

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Berechnung der Teststatistiken χ^2 und L^2

erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	170	186.10	1.46	-30.76
ja	gut	nicht gut	72	69.37	0.10	5.35
ja	nicht gut	gut	751	751.55	0.00	-1.10
ja	nicht gut	nicht gut	928	913.98	0.22	28.26
nein	gut	gut	168	151.90	1.71	33.85
nein	gut	nicht gut	54	56.63	0.12	-5.13
nein	nicht gut	gut	614	613.45	0.00	1.10
nein	nicht gut	nicht gut	732	746.02	0.26	27.78
Summe			3489	3489	3.80	3.79

Bei einer Irrtumswahrscheinlichkeit von 5% ist die Nullhypothese abzulehnen, wenn die Teststatistik größer ist als das 95%-Quantil der Chiquadratverteilung mit df = (4–1)(2–1) = 3 Freiheitsgraden. Der kritische Wert beträgt 7.82.

Da die Teststatistiken Werte von 3.80 bzw. 3.79 haben, kann die Nullhypothese nicht verworfen werden. Bei einer Irrtumswahrscheinlichkeit von 5% kann nicht ausgeschlossen werden, dass die bivariate Verteilung der Einschätzung der allgemeinen und der eigenen Wirtschaftslage von der Erwerbstätigkeit unabhängig ist.

Dann kann es aber auch keinen Interaktionseffekt der Erwerbstätigkeit und der allgemeinen Wirtschaftslage mit der eigenen Wirtschafslage geben. Die beoabchteten Unterschiede der Prozentsatzdifferenzen in den beiden Partialtabellen sind somit nicht signifikant.

Prüfung der Beziehung in den Partialtabellen

Innerhalb jeder Partialtabelle lassen sich die Tests für bivariate Kreuztabellen anwenden.
Die Testergebnisse gelten dann für die Subpopulation, die durch die Ausprägung der Kontrollvariablen definiert ist.
Da die einzelnen Partialtabellen unabhängig voneinander sind, können die Chiquadratstatistiken und Freiheitsgrade in den Partialtabellen auch aufsummiert werden und gelten dann für die Gesamtheit der Partialtabellen.
Soll z.B. geprüft werden, ob statistische Unabhängigkeit in den beiden durch die Erwerbstätigkeit definierten Partialtabellen von AWL und EWL gelten, ist folgendes Hypothesenpaar zu testen:

$$H_0: \pi_{ij(k)} = \pi_{i\bullet(k)} \cdot \pi_{\bullet j(k)} \text{ versus } H_1: \pi_{ij(k)} \neq \pi_{i\bullet(k)} \cdot \pi_{\bullet j(k)} \text{ für alle k}$$

Die Teststatistiken werden getrennt für jede Partialtabelle berechnet und anschließend aufsummiert.

	erwerbstätig			nicht erwerbstätig		
	AWL			AWL		
EWL	gut	nicht gut	∑	gut	nicht gut	∑
gut	170	751	921	168	614	782
nicht gut	72	928	1000	54	732	786
∑	242	1679	1921	222	1346	1568

Prüfung der Beziehung in den Partialtabellen

	erwerbstätig			nicht erwerbstätig		
	AWL			AWL		
EWL	gut	nicht gut	∑	gut	nicht gut	∑
gut	170	751	921	168	614	782
nicht gut	72	928	1000	54	732	786
∑	242	1679	1921	222	1346	1568

$$\chi^2 = 1921 \cdot \frac{(170 \cdot 928 - 751 \cdot 72)^2}{921 \cdot 1000 \cdot 242 \cdot 1679} + 1568 \cdot \frac{(168 \cdot 732 - 54 \cdot 614)^2}{782 \cdot 786 \cdot 222 \cdot 1346} = 123.07$$

Bei einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese abgelehnt, wenn die Teststatistik größer ist als das 95%-Quantil der Chiquadratverteilung mit 2 Freiheitsgraden, d.h. größer ist als 5.99.
Da dies der Fall ist, ist die Nullhypothese abzulehnen. Auch bei Kontrolle der Erwerbstätigkeit besteht zwischen der allgemeinen und der eigenen Beurteilung der Wirtschaftslage ein Zusammenhang.
Da die kleinste erwartete Häufigkeit 782·222 / 1568 = 110.7 > 5 sind die Anwendungsvoraussetzungen für den Chiquadrattest gegeben.

Aufgabe:

Aufgabe 10.19

Die sozialpsychologische Theorie des Wahlverhaltens der Michigan-Schule geht davon aus, dass die Wahlentscheidung durch die Parteineigung und die Kandidatenpräferenz bestimmt wird.

In der Bundesrepublik Deutschland wird die Parteineigung üblicherweise durch die Frage erfasst: „In Deutschland neigen viele Leute längere Zeit einer bestimmten politischen Partei zu, obwohl sie auch ab und zu eine andere Partei wählen. Wie ist das bei Ihnen: Neigen Sie — ganz allgemein gesprochen — einer bestimmten Partei zu? Wenn ja, welcher?“

In Wahlumfragen zur Bundestagswahl 1998 wurde des weiteren nach dem bevorzugten Kanzler gefragt: „Wenn Helmut Kohl und Gerhard Schröder kandidieren, wer wäre Ihnen dann als Bundeskanzler lieber?“

Als dritte Frage ist hier die Wahlabsicht (Sonntagsfrage, siehe Kapitel 2) interessant.

Wenn bei der Parteineigung und der Wahlabsicht nur die Anhänger bzw. potentiellen Wähler von CDU/CSU und SPD berücksichtigt werden, dann kann aus den Daten der Umfragen des ZDF-Politbarometers zur Bundestagswahl 1998 folgende trivariate Häufigkeitsverteilung erstellt werden:

Aufgabe:

	Parteineigung			
	CDU/CSU		SPD	
bevorzugter Kandidat Wahlabsicht	Kohl	Schröder	Kohl	Schröder
CDU/CSU	2016	287	21	34
SPD	47	194	76	2896

a) Es wird vermutet, dass der (langfristige) Effekt der Parteineigung nur vermittelt (indirekt) über den (kurzfristigen) Effekt der Kandidatenpräferenz auf die Wahlabsicht wirkt.
Welches ist bei dieser Fragestellung der konditionale Effekt?

b) Welche Konsequenz erwarten Sie bei Gültigkeit dieser Hypothese für den konditionalen Effekt?

c) Wie können Sie die Vermutung mit einer Irrtumswahrscheinlichkeit von 1 % überprüfen? Beschreiben Sie die Vorgehensweise.

d) Führen Sie die zur Prüfung der Hypothese unter Teilaufgabe a) notwendigen Berechnungen durch.

e) Eine andere Vermutung besagt, dass ein additiver Effekt von Parteineigung und Kandidatenpräferenz auf die Wahlabsicht besteht.
Welches ist bei dieser Fragestellung der bedingte Effekt? Welche Konsequenzen erwarten Sie bei Gültigkeit dieser Hypothese für den bedingten Effekt?

Aufgabe:

f) Wie können Sie die Vermutung mit einer Irrtumswahrscheinlichkeit von 1 % überprüfen? Beschreiben Sie die Vorgehensweise.
g) Führen Sie die zur Prüfung der Hypothese unter Teilaufgabe d) notwendigen Berechnungen durch.
h) Eine dritte Vermutung lautet, dass es einen Interaktionseffekt zwischen der Parteineigung und der Kandidatenpräferenz gibt.
   Wie können Sie diese Hypothese prüfen? Beschreiben Sie die Vorgehensweise.

(Krebs u.a. 2003, S. 296ff.)