

## Bivariate Verteilungen

Tabellarische Darstellung:

Bivariate Tabellen entstehen durch Kreuztabulation zweier Variablen.

Beispiel:

	X	Y
<b>Student(in)</b>	<b>Herkunft</b>	<b>Fakultät</b>
0001	Europa	Jura
0002	Nicht-Europa	Medizin
0003	Nicht-Europa	Philosophie
0004	Europa	Sowi
.	.	.
.	.	.
1240	Europa	Wiwi

# Tabelle: Herkunft und Studium

Studierende an den Fakultäten der Universität zu Köln

Herkunft (x)

Fakultät  
(y)

	Europa	Nicht-Europa	$\Sigma$
Med	46	82	<b>128</b>
Jura	38	20	<b>58</b>
Phil	179	197	<b>376</b>
Sowi	238	189	<b>427</b>
Wiwi	97	154	<b>251</b>
$\Sigma$	<b>598</b>	<b>642</b>	<b>1240</b>

# Kontingenztabelle

	weiblich	männlich	$\Sigma$
Kein Unfall	300 (30%)	300 (30%)	600 (60%)
Unfall ohne Personen- Schaden	75 (7,5%)	150 (15%)	225 (22,5%)
Unfall mit Personen- Schaden	25 (2,5%)	150 (15%)	175 (17,5%)
$\Sigma$	400 (40%)	600 (60%)	1000 (100%)

Frauen: kein Unfall/weiblich  $\rightarrow \frac{300}{400} \Rightarrow 75\%$   
Unfall ohne/weiblich  $\rightarrow \frac{75}{400} \Rightarrow 18,75\%$   
Unfall mit/ weiblich  $\rightarrow \frac{25}{400} \Rightarrow 6,25\%$

Männer: kein Unfall  $\rightarrow \frac{300}{600} \Rightarrow 50\%$   
Unfall ohne/männlich  $\rightarrow \frac{150}{600} \Rightarrow 25\%$   
Unfall mit/männlich  $\rightarrow \frac{150}{600} \Rightarrow 25\%$

# Tabelle: Die generelle Struktur der bivariaten Tabelle

	$X_1$	$X_2$	$X_3$	
$y_{1\cdot}$	$J_{11}$	$J_{12}$	$J_{13}$	$n_{1\cdot}$
$y_{2\cdot}$	$J_{21}$	$J_{22}$	$J_{23}$	$n_{2\cdot}$
$y_{3\cdot}$	$J_{31}$	$J_{32}$	$J_{33}$	$n_{3\cdot}$
$y_{4\cdot}$	$J_{41}$	$J_{42}$	$J_{43}$	$n_{4\cdot}$
$y_{5\cdot}$	$J_{51}$	$J_{52}$	$J_{53}$	$n_{5\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$N$

## **Aufgabe:**

106 Studierende wurden vor der Klausur gebeten, ihre derzeitige psychische Verfassung einzuschätzen. Zur Beurteilung wurde ihnen eine vierstufige Skala von „äußerst labil“, „labil“, „stabil“ und „sehr stabil“ vorgelegt, und sie wurden gebeten, ihre psychische Lage anhand dieser Skala einzuordnen.

Die Untersuchung kam zu folgendem Ergebnis:

Von den 44 weiblichen und 62 männlichen Befragten beschrieben 16 Frauen und lediglich 3 Männer ihre Verfassung als „äußerst labil“. 18 Frauen und 22 Männer ordneten sich der Kategorie „labil“ ein. „Stabil“ fühlten sich dagegen 9 Frauen und 32 Männer, „sehr stabil“ lediglich 1 Frau und 5 Männer.

a) Stellen Sie das Ergebnis in einer Kontingenztafel dar.

- b) Wie viel Prozent der Frauen beurteilen ihre psychische Situation als sehr stabil?
  
- c) Wie viel Prozent der befragten Studierenden, die ihre psychische Situation als äußerst labil bezeichnen, sind Frauen?
  
- d) Beurteilen Männer ihre psychische Situation häufiger als Frauen als labil?



# Tabelle: Verschiedene Grade der Beziehung in 2x2-Tabellen mit gleichen Randverteilungen

a) Keine Beziehung

25	25	50
25	25	50
50	50	100

c) starke Beziehung

40	10	50
10	40	50
50	50	100

b) Schwache Beziehung

28	22	50
22	28	50
50	50	100

d) perfekte Beziehung

50		50
	50	50
50	50	100

# Die Analyse der Beziehung zwischen nominalen Variablen

## Assoziationsmaße

1. Die Prozentsatzdifferenz (d%)
2. Chi-Quadrat
3. Phi-Koeffizient
4. Cramer's V
5. Kontingenzkoeffizient C

# Die Prozentsatzdifferenz

Beispiel:

Berufstätigkeit in Jahren und Vorgesetztenfunktion

	bis 25	über 25	
	1	2	Row Total
Nein	21	15	36
	70.0	50.0	60.0
Ja	9	15	24
	30.0	50.0	40.0
Column	30	30	60
Total	50.0	50.0	100.0

Von den bis 25 Jahre lang Berufstätigen haben 70%, von den über 25 Jahre lang berufstätigen nur 50% keine Vorgesetztenfunktion. Diese Differenz zwischen den Prozentsätzen ist ein Maß der Beziehung zwischen den Variablen.

Die Prozentsatzdifferenz ist wie folgt definiert:

$$d\% = \frac{100(ad - bc)}{(a + c)(b + d)}$$

# Lösung

$$d\% = \frac{100(ad - bc)}{(a + c)(b + d)}$$

$$d\% = \frac{100(21 * 15 - 15 * 9)}{(21 + 15)(9 + 15)}$$

$$d\% = 20,83$$

Die Prozentsatzdifferenz beträgt 0 bei vollständiger Unabhängigkeit (Indifferenz) und +/- 100 bei vollständiger Abhängigkeit.

Aus einem positiven Vorzeichen ist zu schließen, dass die Beziehung entlang der (ad)-Diagonalen verläuft, aus einem negativen Vorzeichen ist zu schließen, dass das Übergewicht auf der (bc)-Diagonalen liegt.

Dieses Assoziationsmaß eignet sich nur für 2 x 2 Tabellen.

Für größere Tabellen werden andere Assoziationsmaße benötigt.

# Chi-Quadrat

Hierbei vergleicht man die vorgefundenen Häufigkeiten in der *Kontingenztabelle* ( $f$ ) mit den Häufigkeiten aus der *Indifferenztabelle* ( $f_e$ ). In der Indifferenztabelle finden sich die Häufigkeiten, die man erwarten würde, wenn keine Beziehung zwischen den Variablen bestünde. Berechnung:

$$\chi^2 = \sum \frac{(f_b - f_e)^2}{f_e}$$

# Kontingenztafel (fb)

Berufswechsel/ Abitur	ja	nein	
ja	9	17	26
nein	24	10	34
	33	27	60

Die Häufigkeiten der Indifferenztabelle werden auf Basis der Randhäufigkeiten der Kontingenztabelle berechnet.

$$f_{ij} = \frac{n_{i.} \cdot n_{.j}}{N}$$

Randhäufigkeit der Zeile • Randhäufigkeit der Spalte  
N

# Indifferenztabelle (fe)

Berufswechsel/ Abitur	ja	nein	
ja	14,3	11,7	26
nein	18,7	15,3	34
	33	27	60

## Berechnung von Chi-Quadrat

$$\chi^2 = \sum \frac{(f_b - f_e)^2}{f_e}$$

Um die Differenz zwischen den beobachteten und den erwarteten Häufigkeiten festzustellen, muss Chi-Quadrat berechnet werden.

Hierzu wird erstens die Differenz  $(f_b - f_e)$  zwischen der beobachteten Häufigkeit und der erwarteten Häufigkeit einer jeden Zelle berechnet, zweitens jede Differenz quadriert  $(f_b - f_e)^2$ , drittens jede quadrierte Differenz durch die erwartete Häufigkeit dividiert  $(f_b - f_e)^2 / f_e$  und schließlich über alle Zellen summiert  $\sum (f_b - f_e)^2 / f_e = \chi^2$ .

Der Nachteil dieses Assoziationsmaßes liegt darin, dass eine Verdoppelung der Zellenhäufigkeiten bei identischen Verteilungen zu einer Verdoppelung des Chi-Quadrat-Wertes führt. Chi-Quadrat variiert also mit N.



- Wenn die Kontingenztabelle gleich der Indifferenztabelle ist, dann besteht zwischen  $x$  und  $y$  kein Zusammenhang
- Wenn die Kontingenztabelle ungleich der Indifferenztabelle ist, dann besteht ein Zusammenhang zwischen  $x$  und  $y$ .

## Direkte Berechnung von Chi-Quadrat

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

# Arbeitstabelle zur Berechnung von chi-quadrat

Zeile i	Spalte j	fb	fe	(fb-fe)	(fb-fe) <sup>2</sup>	(fb-fe) <sup>2</sup> / fe
1	1	9	14,3	-5,3	28,09	1,96
1	2	17	11,7	5,3	28,09	2,40
2	1	24	18,7	-5,3	28,09	1,50
2	2	10	15,3	5,3	28,09	1,84

$$\chi^2 = \sum \frac{(f_b - f_e)^2}{f_e}$$

$$\chi^2 = 7,7$$

Zur Berechnung des Chi-quadrat-basierten Assoziationsmaßes werden:

1. Im ersten Schritt die Häufigkeiten der Indifferenztafel ermittelt;
2. Die Häufigkeiten der Kontingenztabelle mit den erwarteten/theoretischen Häufigkeiten aus der Indifferenztafel verglichen.
3. Die Differenz zwischen den Häufigkeiten zur Berechnung des Assoziationsmaßes heranziehen.

Es gilt, je größer die Differenz zwischen den Häufigkeiten der beiden Tabellen, desto größer ist die Abweichung von der statistischen Unabhängigkeit.

# **Phi-Koeffizient**

Der **Phi-Koeffizient** ist ein Maß, das die Anzahl der Fälle (N) berücksichtigt.

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Die Formel zur direkten Berechnung des Phi Koeffizienten lautet:

$$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Ein nach der ersten Formel berechneter Wert ist vorzeichenlos, ein nach der zweiten Formel berechneter Wert kann zwischen +1 und -1 variieren. Bei ordinalen und metrischen Daten könnte dieses Vorzeichen also auch inhaltlich interpretiert werden.

Für 2 x 2 Tabellen ist  $\Phi$  ein sensibles Assoziationsmaß. Es nimmt den Wert 0 an, wenn die beobachteten mit den erwarteten Häufigkeiten übereinstimmen. Phi erreicht den Wert 1, wenn Chi-Quadrat seinen maximalen Wert erreicht.

Für größere als 2 x 2 Tabellen wird  $\Phi > 1$ .

# Cramer's V

**Cramer's V** ist definiert als

$$\text{Cramers } V = \sqrt{\frac{\chi^2}{N \min(r-1, c-1)}}$$

Dabei steht  $r$  für die Anzahl der Zeilen und  $c$  für die Anzahl der Spalten. „Min“ steht für Minimum und besagt, dass zunächst zu prüfen ist, ob die Anzahl der Zeilen oder die der Spalten kleiner ist. Der kleinere Wert geht dann in die Berechnung des Koeffizienten ein.

# Kontingenzkoeffizient C

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Er hat den Vorteil, dass er für beliebig große Tabellen angewendet werden kann. Sein Nachteil liegt darin, dass seine praktische Grenze unter 1 liegt. Er nähert sich der 1, wenn die Spalten- und Zeilenanzahl zunimmt.