Bivariate Analyse:

Tabellarische Darstellung:

Gemeinsame (bivariate) Häufigkeitstabelle. Sie wird auch **Kontingenz-, Assoziations-** oder **Korrelationstabelle (f_b)**genannt.

Beispiel:

Häufigkeitsverteilung der Variablen						
Student/in	1	2	3	4	5	6
Mathematiknote	3	2	4	1	2	2
Physiknote	3	4	1	3	1	4

⇨ Merkmalsausprägungen der Variable X

⇨ Merkmalsausprägungen der Variable Y

Kontingenztabelle (f_b)

	1	2	3	4	⇨ Variable X
1		1		1	2
2					0
3	1		1		2
4		2			2
⇩ Variable Y	1	3	1	1	N = 6

Randhäufigkeiten der Zeilenvariablen

Randhäufigkeiten der Spaltenvariablen

Korrelationskoeffizienten für nominale Variablen:

1) Prozentsatzdifferenz (d%)

- Mit Bezug auf die Nomenklatur der 2 x 2-Tabelle ist die Prozentsatzdifferenz (d%) wie folgt definiert:

$$d\% = \frac{100 \cdot (ad - bc)}{(a+c)(b+d)} \qquad \text{oder} \qquad d\% = 100 \cdot \left(\frac{a}{a+c} - \frac{b}{b+d} \right)$$

Beispiel:

Tabellarische Darstellung:

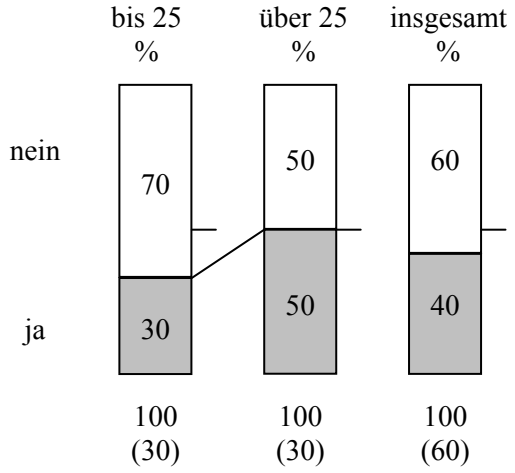
(bc)-Diagonale

Vorgesetztenfunktion (Y-Variable)		Berufstätigkeit (X-Variable) bis 25 Jahren	über 25 Jahre	Σ
	nein	21 **a** (70%)	15 **b** (50%)	36 (60%)
	ja	9 **c** (30%)	15 **d** (50%)	24 (40%)
	Σ	30 (50%)	30 (50%)	60 (100%)

Spaltenprozentwerte

(ad)-Diagonale

Graphische Darstellung:



$$d\% = \frac{100 \cdot (21 \cdot 15 - 15 \cdot 9)}{(21+9)(15+15)} = \frac{100 \cdot (315 - 135)}{900} = 20 \text{ oder } d\% = 100 \cdot \left(\frac{21}{30} - \frac{15}{30} \right) = 100 \cdot (0{,}70 - 0{,}50) = 20$$

Interpretation:

- Je länger jemand berufstätig ist, desto eher übt er/sie auch Vorgesetztenfunktionen aus

Allgemein:

- Die Prozentsatzdifferenz beträgt bei vollständiger Unabhängigkeit (Indifferenz) **Null**, bei vollständiger Abhängigkeit bzw. Assoziation **± 100**. Insofern nimmt der Koeffizient unter den Assoziationsmaßen eine Sonderstellung ein.
- Die Prozentsatzdifferenz vermittelt als einfaches, leicht errechnetes Maß einen plastischen Ausdruck von der Beziehung zwischen den Variablen.
- Die Richtung wird durch das Vorzeichen ausgedrückt.
- Ein **positives Vorzeichen** gibt zu erkennen, dass die **Beziehung entlang der (ad)-Diagonalen verläuft**, während ein **negatives Vorzeichen** das Übergewicht der **(bc)-Diagonalen** anzeigt.
- Größere als 2 x 2-Tabellen weisen mehr als eine Prozentsatzdifferenz auf. Dies kann allerdings eher zur Verwirrung bei der Interpretation führen.
- **Dieses Assoziationsmaß eignet sich also nur für 2 x 2-Tabellen. Für größere Tabellen werden andere Assoziationsmaße benötigt.**

Weitere Korrelationskoeffizienten:

- Statt, wie bei der Prozentsatzdifferenz, die konditionalen Verteilungen einer Vierfelder-Tabelle miteinander zu vergleichen, kann man die vorgefundene Besetzung der Zellen (auch größer als 2 x 2-Tabellen) mit einer Besetzung vergleichen, die man erwarten würde, wenn keine Beziehung zwischen den Variablen bestünde.
- Auf diesem **Vergleich der Häufigkeiten (beobachteten Häufigkeiten) der sogenannten Kontingenztabelle (f_b) mit den Häufigkeiten (erwarteten Häufigkeiten) der sogenannten Indifferenztabelle (f_e)** beruhen **Chi-Quadrat** als auch die traditionellen **chi-quadrat-basierenden Maßzahlen**.

2) Chi-Quadrat (χ^2):

- Dabei wird die Maßzahl Chi-Quadrat nach folgender Formel berechnet:

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

- Direkte Berechnung von Chi-Quadrat:

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

- **Wenn die Kontingenztabelle (f_b) und die Indifferenztabelle (f_e) identisch sind, dann nimmt χ^2 den Wert Null an.**
- **Chi-Quadrat kann maximal den Wert N annehmen**, und zwar im Falle einer 2 x 2-Tabelle mit zwei unbesetzten Diagonalzellen.

Indifferenztabelle (f_e):

- ist die Tabelle der statistischen Unabhängigkeit (fiktive Tabelle)
- Die Indifferenztabelle ist insofern eine imaginäre Tabelle, als sie die gemeinsamen Häufigkeiten bei gegebenen Randverteilungen in einer Weise darstellt, wie man sie anträfe bzw. zu erwarten hätte, wenn keine Beziehung zwischen den Variablen bestünde, d. h. wenn die Variablen voneinander unabhängig wären.

- Wenn die Kontingenztabelle (f_b) **gleich** der Indifferenztabelle (f_e) ist, dann besteht zwischen X und Y **kein** Zusammenhang.
- Wenn die Kontingenztabelle (f_b) **ungleich** der Indifferenztabelle (f_e) ist, dann besteht zwischen X und Y **ein** Zusammenhang.
- **Je größer die Differenz** zwischen den Häufigkeiten der beiden Tabellen ist, **desto größer ist die Abweichung von der statistischen Unabhängigkeit und der Grad der Assoziation** zwischen den Variablen.

Beispiel: $\mathbf{f_b = f_e}$

Kontingenztabelle (f_b)			
	1	2	⇨ Variable X
1	5	5	10
2	5	5	10
⇩ Variable Y	10	10	N = 20

Berechnung der Häufigkeiten der Indifferenztabelle (f_e):

Allgemeine Formel:

$$f_e - Wert = \frac{Randhäufigkeit\ der\ Zeile \cdot Randhäufigkeit\ der\ Spalte}{N}$$

Die Häufigkeiten der Indifferenztabelle muss für jede Zelle berechnet werden.

Ermittlung der Indifferenztabelle am Beispiel der gegebenen Kontingenztabelle:

$$f_e - Wert = \frac{10 \cdot 10}{20} = 5$$

- Der Wert 5 gilt für jede Zelle der Indifferenztabelle, da die Häufigkeiten in der Kontingenztabelle gleich groß sind.

Indifferenztabelle (f_e)			
	1	2	⇨ Variable X
1	5	5	10
2	5	5	10
⇩ Variable Y	10	10	N = 20

Interpretation:

- Es besteht keine statistische Beziehung zwischen den Variablen X und Y, da die Kontingenztabelle gleich der Indifferenztabelle ist. Chi-Quadrat nimmt demnach den Wert Null an.

Berechnung von Chi-Quadrat:

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

Zelle	f_b	f_e	$(f_b - f_e)$	$(f_b - f_e)^2$	$(f_b - f_e)^2 / f_e$
a	5	5	0	0	0
b	5	5	0	0	0
c	5	5	0	0	0
d	5	5	0	0	0
∑	20				**$\chi^2 = 0$**

Interpretation:

- Zwischen den beiden Variablen X und Y besteht kein Zusammenhang.

- Obwohl der Chi-Quadrat-Wert das Ausmaß der Abweichung der Kontingenztabelle von der Indifferenztabelle, d.h. den Grad der Abweichung der beobachteten bivariaten Verteilung von der statistischen Unabhängigkeit reflektiert, kann er in dieser Form nicht als sinnvoller Kennwert der Beziehung zwischen den Variablen fungieren, da **Chi-Quadrat direkt mit N variiert**. D.h. eine Verdoppelung der Zellenhäufigkeiten bei identischen konditionalen Verteilungen bzw. bei denselben Proportionen der Tabellen führt zur Verdoppelung des Chi-Quadrat-Wertes.
- Da man nicht an einer Maßzahl interessiert ist, die bei identischen Graden der Beziehung in Abhängigkeit von der Anzahl der Fälle unterschiedliche Werte annimmt, muss ein auf Chi-Quadrat basierendes Assoziationsmaß die Anzahl der Fälle (N) berücksichtigen.

Maßzahlen auf Basis von Chi-Quadrat

- d.h. Chi-Quadrat steht im Zähler der Berechnungsformeln der jeweiligen Korrelationskoeffizienten

3) Phi (ϕ):

- Phi wird nach folgender Formel berechnet:

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

⇨ Der Wert des Koeffizienten ist vorzeichenlos (Wertebereich: 0 bis +1)

- Direkte Berechnung von Phi:

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

⇨ Der Wert des Koeffizienten kann zwischen -1 und +1 variieren

- **Für größere als 2 x 2-Tabellen kann Phi größer 1 werden, das ihn als Vergleichsgröße untauglich machen lässt.**

4) Cramer's V:

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

- wobei r die Anzahl der Zeilen und c die Anzahl der Spalten symbolisiert. Der Ausdruck „min" steht für Minimum und besagt, dass zunächst zu prüfen ist, ob die Anzahl der Zeilen und die Anzahl der Spalten kleiner ist; der kleinere Wert geht in die Berechnung des Koeffizienten ein.
- Bei 2 x 2-Tabellen ist V identisch mit ϕ.

5) Pearson's Kontingenzkoeffizient C:

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

- **Der Kontingenzkoeffizient C hat den Vorteil, für Tabellen beliebiger Größe (rechteckig oder quadratisch) berechnet werden zu können.**
- Der **Hauptnachteil** des Koeffizienten liegt allerdings darin, dass er praktisch eine unterhalb 1 liegende Obergrenze hat, obwohl sich die Obergrenze dem Wert 1 nähert, wenn die Anzahl der Zeilen und Spalten zunimmt. **Der Maximalwert hängt also von der Größe der zugrunde liegenden Tabelle ab.**

- Der Höchstwert von C ist wie folgt lediglich für quadratische Tabellen genau bestimmbar:

$C_{\max} = \sqrt{\frac{r-1}{r}}$, wobei r die Anzahl der Zeilen der quadratischen Tabelle bezeichnet

- **Hieraus folgt, dass sich C-Werte nur vergleichen lassen, wenn sie für Tabellen gleicher Größe berechnet wurden.**
- Sollen C-Werte unterschiedlich großer Tabellen miteinander verglichen werden, sind sie nach der folgenden Formel zu korrigieren:

$$C_{korr} = \frac{C}{C_{\max}}$$

Allgemein:

- Da Chi-Quadrat im Zähler der Koeffizientenformeln steht, hat dies zur Folge, dass wenn Chi-Quadrat gleich Null ist, alle Koeffizienten, die auf Chi-Quadrat basieren, ebenfalls den Wert Null ergeben.
- Wenn $f_b = f_e$, dann ist Chi-Quadrat und alle Korrelationskoeffizienten, die auf Chi-Quadrat basieren, also Phi, T, V, C, ebenfalls gleich Null.
- Allerdings ist die große Schwäche der chi-quadrat-basierten Maßzahlen, dass ihre Zahlenwerte kaum miteinander verglichen werden können.