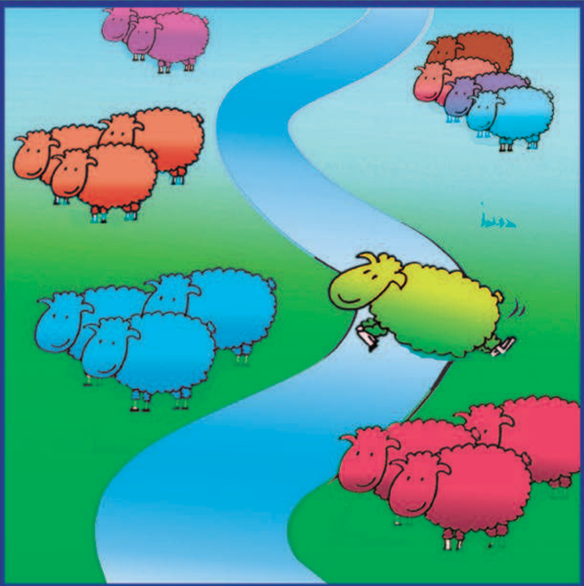

Achim Bühl

SPSS 18

Einführung in die moderne Datenanalyse

12., aktualisierte Auflage

ALWAYS LEARNING

PEARSON

$$\text{Lambda} = \frac{43{,}7\ \% - 28{,}1\ \%}{43{,}7\ \%} = 0{,}357$$

Erreicht der Fehler bei der zweiten Diagnose den Wert 0, wird Lambda = 1. Ist der Fehler bei der zweiten Prognose gleich dem der ersten Prognose, wird Lambda = 0. Hier liefert die zweite Variable keinerlei Unterstützung bei der Vorhersage der ersten (abhängigen) Variablen; beide Variablen sind völlig unabhängig voneinander.

Da der Computer nicht wissen kann, welche Variable als abhängig zu betrachten ist, gibt PASW automatisch in diesem Zusammenhang zwei Werte an, wobei jede der beiden Variablen jeweils einmal als abhängig betrachtet wird. Für den Fall, dass Sie keine der beiden Variablen als abhängig deklarieren können, wird unter der Bezeichnung »symmetrisch« das gewichtete Mittel dieser beiden Werte angegeben.

Goodman und Kruskals Tau

Dies ist eine Variante des Lambda-Koeffizienten, die von PASW stets zusammen mit diesem ausgegeben wird, wobei die Anzahl der richtig vorhergesagten Personen anders berechnet wird, und zwar werden die beobachteten Häufigkeiten mit ihren Prozenten gewichtet und aufaddiert. Dies ergibt zunächst für die alleinige Betrachtung der abhängigen Variablen (erste Prognose):

36 * 56,3 % + 28 * 43,8 % = 32,53

Folgt man diesem Ansatz, werden von den insgesamt 64 Personen 31,47 falsch vorhergesagt, das sind 49,17%.

Unter Berücksichtigung der zweiten Variablen erhält man nach dem entsprechenden Ansatz für die richtig vorhergesagten Personen (zweite Prognose):

13 * 59,1 % + 16 * 88,9 % + 7 * 29,2 % +9 * 40,9 % + 2 * 11,1 % +17 * 70,8 %

= 39,89

Bei der zweiten Prognose wurden also von den insgesamt 64 Personen 24,11, das sind 37,67%, falsch vorhergesagt. Dies ergibt in diesem Fall eine Fehlerreduktion von

$$\frac{49{,}17\ \% - 37{,}67\ \%}{49{,}17\ \%} = 0{,}235$$

Dieser Wert wird unter der Bezeichnung »Goodman-und-Kruskal-Tau« ausgegeben. Auch hier gibt PASW einen zweiten Wert an; dieser ergibt sich, wenn man die andere Variable als abhängig betrachtet.

Unsicherheitskoeffizient

Dies ist eine weitere Variante von Lambda, wobei anstelle von fehlerhaften Prognosen von »Unsicherheit« gesprochen wird und diese Unsicherheit nach recht komplizierten Formeln bestimmt wird, so dass auf eine nähere Erläuterung des Rechengangs verzichtet wird.

Der Unsicherheitskoeffizient kann ebenfalls Werte zwischen 0 und 1 annehmen. Der Wert 1 bedeutet, dass die eine Variable exakt aus der anderen vorhergesagt werden kann.

9.3.4 Assoziationsmaße für ordinalskalierte Variablen

Alle diese Maße basieren auf der Anzahl der Fehlordnungen (Inversionen: I), die sich ergeben, wenn man die Werte einer der beiden Variablen in aufsteigender Reihenfolge niederschreibt und die Werte der anderen Variablen entsprechend zuordnet. Zusammen mit der Anzahl der richtigen Ordnungen (Proversionen: P) gehen sie auf verschiedene Weise in die entsprechenden Formeln für die Assoziationsmaße ein, die Werte zwischen minus 1 und plus 1 annehmen können.

Gamma

Gamma berechnet sich aus der einfachen Formel:

$$G = \frac{P - I}{P + I}$$

Treten keine Inversionen auf (I = 0), wird G = 1 (totaler Zusammenhang). Treten keine Proversionen, also ausschließlich Inversionen auf (P = 0), handelt es sich um einen maximal gegenläufigen Zusammenhang (G = -1). Ist P = I, ist keinerlei Zusammenhang gegeben (G = 0).

Somers d

Berechnet werden asymmetrisches und symmetrisches Somers d. Die Formel für Gamma wird durch einen Korrekturterm erweitert, der im Falle von Bindungen (Auftreten gleicher Messwerte) bei der abhängigen Variablen relevant wird:

$$d = \frac{P - I}{P + I + T_y}$$

Ein entsprechendes asymmetrisches Somers d gibt es mit dem Korrekturterm T_x. Das symmetrische Somers d verwendet im Nenner den Mittelwert der beiden asymmetrischen Koeffizienten.

Kendalls Tau-b

Bei dieser Maßzahl werden die Bindungen sowohl der abhängigen als auch der unabhängigen Variablen berücksichtigt:

$$\tau_b = \frac{P - I}{\sqrt{(P + I + T_x) \cdot (P + I + T_y)}}$$

Die Werte -1 und +1 können nur bei einer quadratischen Kreuztabelle angenommen werden.

Kendalls Tau-c

Hier können die Werte -1 und +1 bei jeder Tabelle erreicht werden:

$$\tau_c = \frac{2 \cdot m \cdot (P - I)}{N^2 \cdot (m - 1)}$$

Dabei ist N die Gesamtsumme der Häufigkeiten; m ist die kleinere der Anzahlen der Zeilen und Spalten.

9.3.5 Weitere Assoziationsmaße

PASW bietet weitere spezielle Assoziationsmaße an, die im Folgenden vorgestellt werden sollen.

Eta

Dieser Koeffizient ist geeignet, wenn die abhängige Variable Intervall- und die unabhängige Variable Ordinal- oder Nominalskalenniveau hat. Das Quadrat von Eta ist der Anteil der Gesamtvarianz, der durch die unabhängige Variable erklärt wird.

Kappa-Koeffizient

Cohens Kappa-Koeffizient kann nur für quadratische Kreuztabellen berechnet werden, in denen dieselben numerischen Kodierungen für die Zeilen- und die Spaltenvariable verwendet wurden. Der typische Anwendungsfall ist der, dass Personen oder Objekte durch zwei Gutachter beurteilt werden. Kappa gibt dann den Grad der Übereinstimmung zwischen den beiden Beurteilern an.

Risiko

Unter dieser Option verbergen sich bei PASW drei verschiedene Koeffizienten, die für Vierfeldertafeln bestimmt werden können, wobei eine bestimmte Testsituation gegeben sein muss: Eine so genannte Risikovariable, die angibt, ob ein bestimmtes Ereignis eintrifft oder nicht, wird in Abhängigkeit von einer unabhängigen (ursächlichen) und ebenfalls dichotomen Variablen untersucht.

Dies soll an einem typischen Beispiel erläutert werden. Bei insgesamt 294 Probanden aus einer Studie über Depressionen ergab sich die folgende Häufigkeitsverteilung:

	Depression	
	Ja	Nein
Weiblich	a = 40	b = 143
Männlich	c = 10	d = 101

Die Depression mit den beiden Kategorien ja – nein ist die Risikovariable, das Geschlecht mit den beiden Kategorien weiblich – männlich die unabhängige (ursächliche) Variable.

In Zusammenhang mit solchen Studien spricht man von Kohorten- bzw. Fall-Kontrollstudien. Dabei untersuchen Kohortenstudien eine festgelegte Fallgruppe, bei denen das untersuchte Ereignis noch nicht eingetreten ist, über einen gewissen Zeitraum hinweg und stellen fest, bei welchen Fällen dieses Ereignis eintritt und bei welchen nicht und ob sich das Risiko des Eintretens zwischen den Kategorien einer unabhängigen Variablen unterscheidet. Bei Fall-Kontrollstudien hingegen wird eine Fallgruppe, bei der das Ereignis bereits eingetreten ist, mit einer Kontrollgruppe verglichen.

Zwei der von PASW berechneten Koeffizienten werden üblicherweise bei Kohortenstudien, der dritte Koeffizient bei Fall-Kontrollstudien verwendet. In Kohortenstudien wird für die beiden Kategorien der unabhängigen Variablen (hier: das Geschlecht) die Inzidenzrate bestimmt. Bei den weiblichen Probanden ist die Inzidenzrate für das Auftreten einer Depression

$$\frac{40}{40 + 143} = 0{,}219$$

Bei den männlichen Probanden ist die Inzidenzrate

$$\frac{10}{10 + 101} = 0{,}09$$

Der Quotient aus den beiden Inzidenzraten

$$\frac{0{,}219}{0{,}090} = 2{,}426$$

wird als relatives Risiko bezeichnet. Das Risiko, depressiv zu werden, liegt bei den Frauen um das 2,426-Fache über dem bei den Männern. Da der Rechner nicht wissen kann, welche der beiden Kodierungen der Risikovariablen für das Vorhandensein der Depression steht, wird dieses relative Risiko für beide Ausprägungen der Risikovariablen berechnet.

Bei Fall-Kontrollstudien wird eine etwas andere Variante der Koeffizientenberechnung, die auch Odds Ratio genannt wird, verwendet. So sind die »Chancen« (odds) bei den Frauen, depressiv zu werden, 40/143, bei den Männern 10/101. Das Chancenverhältnis (Odds Ratio) ist demnach

$$\frac{40 * 101}{143 * 10} = 2{,}825$$

Bezeichnen wir die vier Häufigkeiten der Vierfeldertafel mit a, b, c und d (siehe oben), so werden demnach von PASW die drei folgenden Koeffizienten berechnet:

$$R0 = \frac{a * d}{b * c}$$

$$R1 = \frac{a \ * \ (c \ + \ d)}{c \ * \ (a \ + \ b)}$$

$$R2 = \frac{b \ * \ (c \ + \ d)}{d \ * \ (a \ + \ b)}$$

Wir wollen das gegebene Beispiel in PASW durchrechnen.

- Laden Sie die Datei depr.sav.

Die Datei enthält als Risikovariable die Variable depr mit den beiden Kodierungen 1 = ja und 2 = nein sowie die unabhängige (ursächliche) Variable sex mit den beiden Kodierungen 1 = weiblich und 2 = männlich. Ferner ist als weitere Variable die Häufigkeitsvariable n enthalten.

- Treffen Sie die Menüwahl

  Daten
  Fälle gewichten...

  und geben Sie n als Häufigkeitsvariable an.

- Definieren Sie anschließend in der Dialogbox *Kreuztabellen* die Variable sex als Zeilen- und die Variable depr als Spaltenvariable und aktivieren Sie über den Schalter *Statistiken...* die Option *Risiko*.

Es werden die folgenden Ergebnisse angezeigt.

Risikoschätzer

		95%-Konfidenzintervall	
	Wert	Untere	Obere
Quotenverhältnis für Geschlecht (Weiblich / Männlich)	2,825	1,350	5,911
Für Kohorten-Analyse Depression = Ja	2,426	1,265	4,655
Für Kohorten-Analyse Depression = Nein	,859	,780	,946
Anzahl der gültigen Fälle	294		

Es werden nacheinander Odds Ratio (R0) und die beiden relativen Risiken (R1 und R2) ausgegeben, wobei auch ein 95%-Konfidenzintervall berechnet wird.

Möchten Sie Odds Ratio und das relative Risiko korrekt berechnen, müssen Sie Folgendes beachten:

- Definieren Sie die ursächliche (unabhängige) Variable als Zeilenvariable und die Risikovariable als Spaltenvariable.
- In der ersten der beiden Zeilen der Vierfeldertafel muss die Gruppe mit dem größeren Risiko eingetragen sein.
- In der ersten der beiden Spalten der Vierfeldertafel muss die Kodierung für das Eintreffen des Ereignisses enthalten sein.

Chi-Quadrat-Test nach McNemar

Der Chi-Quadrat-Test nach McNemar wird bei zwei abhängigen dichotomen Variablen eingesetzt; er ist in Kapitel 13.2 beschrieben.

Cochran- und Mantel-Haenszel-Statistik

Diese Statistik knüpft an die Berechnung des Odds Ratio bei Vierfeldertafeln an, die mit der Option *Risiko* ermöglicht wird. Dabei kommt eine Schichtenvariable (Kovariate) ins Spiel, wobei getestet wird, ob sich das Quotenverhältnis (Odds Ratio) über die Kategorien dieser Variable hinweg signifikant von 1 (oder auch einem anderen Wert) unterscheidet. Ein Beispiel mag dies verdeutlichen.

- Laden Sie die Datei angst.sav.

Von insgesamt 1737 Personen sind drei Variablen gespeichert, nämlich das Geschlecht (1 = weiblich, 2 = männlich), eine Angabe darüber, ob eine Angststörung vorliegt (1 = ja, 2 = nein), und eine Angabe über das Körpergewicht (1 = nicht übergewichtig, 2 = übergewichtig). Wir wollen getrennt nach Übergewichtigen und Nichtübergewichtigen eine Kreuztabelle zwischen dem Geschlecht und dem Auftreten von Angststörungen erstellen und dabei jeweils das Odds Ratio berechnen.

- Treffen Sie die Menüwahl

  Daten

  Datei aufteilen...

- Aktivieren Sie die Option *Ausgabe nach Gruppen aufteilen* und geben Sie die Variable gewicht als Gruppierungsvariable an.
- Wählen Sie anschließend aus dem Menü

  Analysieren

  Deskriptive Statistiken

  Kreuztabellen...

- Tragen Sie die Variable sex als Zeilenvariable und die Variable angst als Spaltenvariable ein.
- Aktivieren Sie über den Schalter *Zellen...* die Ausgabe zeilenweiser Prozentwerte und über den Schalter *Statistiken...* die Option *Risiko*.

Die wesentliche Ausgabe ist im Folgenden wiedergegeben.

Übergewicht = Nein

Geschlecht * Angststörung Kreuztabelle[a]

			Angststörung		
			Ja	Nein	Gesamt
Geschlecht	Weiblich	Anzahl	154	592	746
		% innerhalb von Geschlecht	20,6%	79,4%	100,0%
	Männlich	Anzahl	79	715	794
		% innerhalb von Geschlecht	9,9%	90,1%	100,0%
Gesamt		Anzahl	233	1307	1540
		% innerhalb von Geschlecht	15,1%	84,9%	100,0%

a. Übergewicht = Nein

Risikoschätzer[a]

	Wert	95%-Konfidenzintervall	
		Untere	Obere
Quotenverhältnis für Geschlecht (Weiblich / Männlich)	2,354	1,758	3,154
Für Kohorten-Analyse Angststörung = Ja	2,075	1,612	2,670
Für Kohorten-Analyse Angststörung = Nein	,881	,844	,920
Anzahl der gültigen Fälle	1540		

a. Übergewicht = Nein

Übergewicht = Ja

Geschlecht * Angststörung Kreuztabelle[a]

			Angststörung		Gesamt
			Ja	Nein	
Geschlecht	Weiblich	Anzahl	22	62	84
		% innerhalb von Geschlecht	26,2%	73,8%	100,0%
	Männlich	Anzahl	9	104	113
		% innerhalb von Geschlecht	8,0%	92,0%	100,0%
Gesamt		Anzahl	31	166	197
		% innerhalb von Geschlecht	15,7%	84,3%	100,0%

a. Übergewicht = Ja

Risikoschätzer[a]

	Wert	95%-Konfidenzintervall	
		Untere	Obere
Quotenverhältnis für Geschlecht (Weiblich / Männlich)	4,100	1,776	9,468
Für Kohorten-Analyse Angststörung = Ja	3,288	1,597	6,771
Für Kohorten-Analyse Angststörung = Nein	,802	,698	,921
Anzahl der gültigen Fälle	197		

a. Übergewicht = Ja

Das Auftreten der Angststörung ist in beiden Fällen bei den Frauen deutlich erhöht. Das Odds Ratio beträgt bei den Nichtübergewichtigen 2,354 und bei den Übergewichtigen 4,100.

Wir wollen nun die Cochran- und Mantel-Haenszel-Statistik berechnen.

- Um zunächst die Aufteilung nach Gruppen wieder rückgängig zu machen, aktivieren Sie nach der Menüwahl

  Daten
  Datei aufteilen...

  die Option *Alle Fälle analysieren, keine Gruppen bilden*.

- In der Dialogbox *Kreuztabellen* geben Sie zusätzlich die Variable gewicht als Schichtenvariable an, deaktivieren über den Schalter *Statistiken...* die Option *Risiko* und aktivieren die Option *Cochran- und Mantel-Haenszel-Statistik*.

- Unter der Option *Gemeinsames Quoten-Verhältnis* belassen Sie es bei der Voreinstellung von 1.

Wir beschränken uns darauf, die Ausgabe der Cochran- und Mantel-Haenszel-Statistik wiederzugeben.

Tests auf Homogenität des Quotenverhältnisses

	Chi-Quadrat	df	Asymptotische Signifikanz (zweiseitig)
Breslow-Day	1,522	1	,217
Tarone	1,522	1	,217

Tests auf bedingte Unabhängigkeit

	Chi-Quadrat	df	Asymptotische Signifikanz (zweiseitig)
Cochran	44,665	1	,000
Mantel-Haenszel	43,724	1	,000

Unter Annahme der bedingten Unabhängigkeit ist die Cochran-Statistik nur dann als Chi-Quadrat-Verteilung mit 1 Freiheitsgrad asymptotisch verteilt, wenn die Anzahl der Schichten festgelegt ist. Die Mantel-Haenszel-Statistik ist unter dieser Annahme jedoch immer als Chi-Quadrat-Verteilung mit 1 Freiheitsgrad asymptotisch verteilt. Beachten Sie, daß die Kontinuitätskorrektur aus der Mantel-Haenszel-Statistik entfernt wird, wenn die Summe der Differenzen zwischen den beobachteten und erwarteten Größen 0 ist.

Schätzung des gemeinsamen Quotenverhältnisses nach Mantel-Haenszel

Schätzung			2,503
ln(Schätzung)			,918
Standardfehler von ln(Schätzung)			,141
Asymptotische Signifikanz (zweiseitig)			,000
Asymptotisches 95% Konfidenzintervall	Gemeinsames Quotenverhältnis	Untergrenze	1,901
		Obergrenze	3,297
	ln(gemeinsames Quotenverhältnis)	Untergrenze	,642
		Obergrenze	1,193

Die Schätzung des gemeinsamen Quotenverhältnisses nach Mantel-Haenszel ist unter der Annahme des gemeinsamen Quotenverhältnisses von 1,000 asymptotisch normalverteilt. Dasselbe gilt für den natürlichen Logarithmus der Schätzung.

Die Ergebnisse nach Cochran und nach Mantel-Haenszel sind einander sehr ähnlich; in beiden Fällen ergibt sich über beide Gewichtsgruppen hinweg ein höchst signifikanter Unterschied ($p < 0{,}001$) des Quotenverhältnisses (Odds Ratio) zum Wert 1. Sowohl nach Breslow-Day als auch nach Tarone kann die Annahme der Homogenität des Quotenverhältnisses über die beiden Gewichtsgruppen beibehalten werden ($p = 0{,}217$).

Die Schätzung des gemeinsamen Quotenverhältnisses ergibt die gleichen Werte, die sich über die Option Risiko ergeben würden, wenn Sie keine Aufspaltung nach der Schichtenvariablen vorgenommen hätten.

KAPITEL 10

Analyse von Mehrfachantworten

In diesem Kapitel wollen wir die Kodierung und Analyse von Mehrfachantworten behandeln. Fragen mit Mehrfachantwortmöglichkeit spielen bei den meisten Fragebogenauswertungen eine Rolle. Um solche Mehrfachantworten kodieren und auswerten zu können, bietet PASW zwei verschiedene Methoden an: die dichotome und die kategoriale Methode. Diese beiden Methoden werden in zwei getrennten Kapiteln jeweils anhand eines Beispiels dargestellt.

10.1 Dichotome Methode

Wir wollen die dichotome Methode am Beispiel einer Jugendbefragung erläutern. Eine Frage des Fragebogens lautete: »Bei welchen der folgenden Sachen machst du aktiv mit?« Es waren die folgenden Antwortmöglichkeiten vorgegeben:

- Freiwillige Schulveranstaltung
- Klassensprecher/SMV
- Natur-, Umwelt-, Tierschutz
- Menschenrechte
- Kirchengruppe
- Freiwillige Organisation
- Etwas selbständig organisieren
- Sicherheit im Verkehr
- Gewerkschaft
- Parteien
- Stadtjugendring/Jugendverbände
- Politisch aktive Gruppe in der Stadt

Bei der Methode multipler Dichotomien wird für jede der Antwortmöglichkeiten eine eigene Variable definiert. Im gegebenen Beispiel werden dazu also zwölf Variablen benötigt. Kreuzt ein Jugendlicher die Antwort »Freiwillige Schulveranstaltung« an, wird der betreffenden Variablen der Code »1« zugeordnet, andernfalls eine »0«; kreuzt ein Jugendlicher die Antwort »Klassensprecher/SMV« an, wird der betreffenden Variablen der Code »1« zugeordnet, andernfalls eine »0« usw. So entstehen zwölf Variablen, jeweils mit 1 und 0 kodiert. Dabei ist die Wahl der Codezahlen natürlich beliebig; sie muss aber für alle Antwortmöglichkeiten gleich sein und dem Computer an der entsprechenden Stelle mitgeteilt werden.

Die Ergebnisse dieser Mehrfachfrage sind in der Datei mitmach.sav enthalten. Wir wollen zunächst eine Häufigkeitsauszählung der Frage »Bei welchen der folgenden Sachen machst du aktiv mit?« durchführen und dann eine Kreuztabelle dieser Frage mit dem Geschlecht erstellen.

10.1.1 Definition von Sets

Die Antworten auf unsere Frage sind in der beschriebenen Weise in den Variablen v1_1 bis v1_12 kodiert. In einem ersten Schritt müssen wir dem Computer mitteilen, dass diese zwölf Variablen zu einem »Variablenset« gehören.

- Laden Sie die Datei mitmach.sav.
- Treffen Sie die Menüwahl

Analysieren
  Mehrfachantworten
    Variablen-Sets definieren...

Es öffnet sich die Dialogbox *Mehrfachantworten-Sets*.

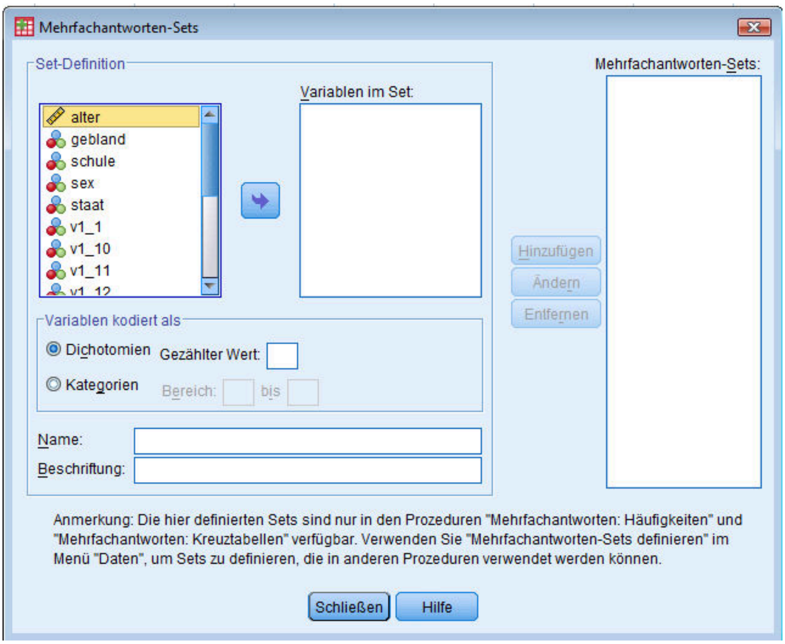

Bild 10.1: Dialogbox Mehrfachantworten-Sets

- Markieren Sie die Variablen v1_1 bis v1_12 in der Quellvariablenliste und übertragen Sie diese in die Zielvariablenliste *Variablen im Set*.
- Kennzeichnen Sie die Art der Kodierung als *Dichotomien* (dies ist die Voreinstellung). Legen Sie als zu zählenden Wert die »1« fest.