AUSSDA
AUSTRIAN SOCIAL SCIENCE DATA ARCHIVE

BIVARIATE KREUZTABELLE MIT SPSS

Anwendungsbeispiel mit Daten von AUSSDA – The Austrian Social Science Data Archive

Erstellt von:

Otto Bodi-Fernandez

Universität Graz

Kontakt:

otto.bodi@uni-graz.at

DOI 10.5281/zenodo.3612749

Version: 1.0

This work is licensed under a Creative Commons Attribution 4.0 International License.

Bivariate Kreuztabelle

In diesem Anwendungsbespiel wird Schritt für Schritt eine bivariate Kreuztabellenanalyse mit der Statistiksoftware SPSS (Version 25) durchgeführt. Die Datenbasis ist der Soziale Survey Österreich (SSÖ) aus dem Jahr 2018. Der Datensatz ist bei AUSSDA – The Austrian Social Science Data Archive unter https://data.aussda.at für Forschung und Lehre frei zugänglich. Die Ergebnisse dieses Beispiels können durch das Ausführen der dokumentierten Syntaxbefehle repliziert werden.

Datensatz

Hadler, Markus; Höllinger, Franz; Muckenhuber, Johanna, 2019, "Social Survey Austria 2018 (SUF edition)", doi:10.11587/ERDG3O, AUSSDA Dataverse, V3

Der Datensatz repräsentiert eine Mehrthemenumfrage, die repräsentativ für die österreichische Wohnbevölkerung ab 18 Jahren ist. Mit der Frage C1 im Fragbogen soll erhoben werden, wie glücklich oder unglücklich sich die befragten Personen fühlen (Happiness).

C1. Wenn Sie einmal Ihr Leben heute betrachten, was würden Sie sagen, wie glücklich oder unglücklich sind Sie alles in allem? (Vorlesen)

→ Nur EINE Markierung möglich!

Sehr glücklich	☐	1
Ziemlich glücklich	☐	2
Nicht sehr glücklich	☐	3
Überhaupt nicht glücklich	☐	4
Kann ich nicht sagen	☐	8

Wir sind nun daran interessiert, ob sich verheiratete Personen im Vergleich zu geschiedenen, ledigen oder verwitweten Personen im Ausmaß ihres Glücklichseins (Happiness) unterscheiden. Anders ausgedrückt: Wir möchten untersuchten, ob der Familienstand einen Einfluss auf die Happiness hat. Familienstand als unabhängige Variable weist lediglich nominales Skalenniveau auf. Es können daher sinnvollerweise keine je-desto Hypothesen aufgestellt werden. Die abhänge Variable Happiness ist ordinalskaliert, weist mit den vier Antwortmöglichkeiten von ‚Sehr glücklich' bis ‚Überhaupt nicht glücklich' aber nur wenige Kategorien auf. Zur Untersuchung des Zusammenhangs zwischen Familienstand und Happiness bietet sich daher eine bivariate Kreuztabelle an.

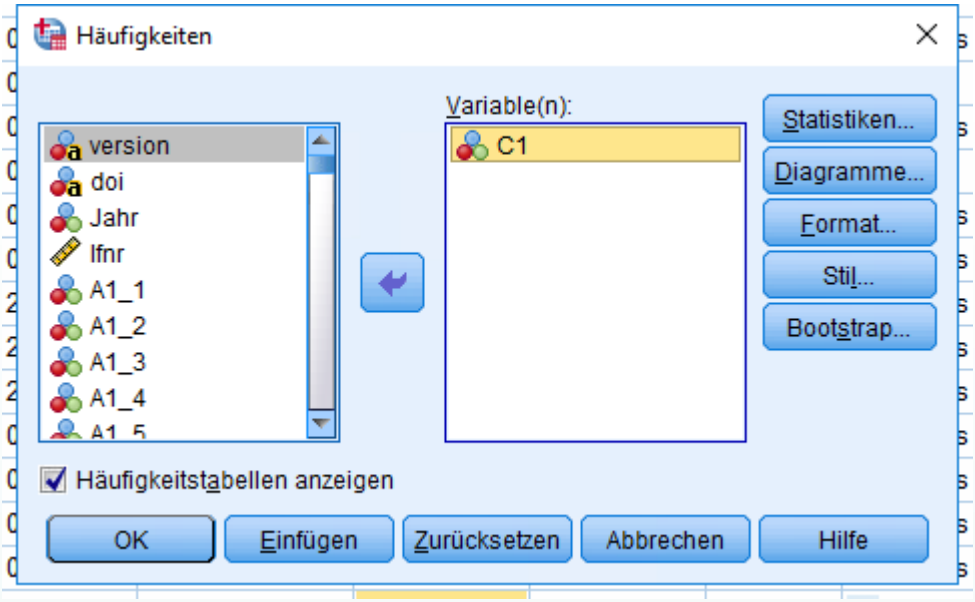
Vorbereitung

Zunächst aktivieren wir im Datensatz die Gewichtungsvariable (**Weight_Combined**). Wir verwenden dazu folgende Syntax.

Syntax:

```
WEIGHT BY Weight_Combined.
```

Wir betrachten zunächst die abhängige Variable Happiness (**C1**). Mit der Befehlskette ***Analysieren → Deskriptive Statistik → Häufigkeiten*** kann folgendes Dialogfeld aufgerufen werden:



Wir wählen Variable C1 aus. Im Menüpunkt ‚Diagramme' wählen wir zusätzlich ‚Balkendiagramm' aus.

Die Häufigkeitstabelle sowie das Balkendiagramm können auch mit folgender Syntax aufgerufen werden:

Syntax:

```
FREQUENCIES VARIABLES=C1
  /BARCHART FREQ
  /ORDER=ANALYSIS.
```

Wie glücklich ("Happiness")

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Sehr glücklich	341	28,4	28,8	28,8
	Ziemlich glücklich	733	61,1	62,0	90,9
	Nicht sehr glücklich	102	8,5	8,6	99,5
	überhaupt nicht glücklich	6	,5	,5	100,0
	Gesamt	1182	98,5	100,0	
Fehlend	[Kann ich nicht sagen]	18	1,5		
Gesamt		1200	100,0		

Wir sehen, dass sich mit über 60% die Mehrheit der befragten Personen ‚ziemlich glücklich' fühlt. Weitere knapp 30% fühlen sich ‚sehr glücklich'. Die restlichen Befragten fallen in die Kategorien ‚nicht sehr glücklich' oder ‚überhaupt nicht glücklich', wobei der Anteil letzterer verschwindend gering ist.

Weiters betrachten wir nun die Verteilung der unabhängigen Variable Familienstand (D13):

Syntax:

```
FREQUENCIES VARIABLES=D13
  /ORDER=ANALYSIS.
```

Familienstand

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	verheiratet	681	56,8	56,9	56,9
	eingetragene Lebenspartnerschaft	31	2,6	2,6	59,5
	geschieden / getrennt	131	10,9	11,0	70,5
	verwitwet / Lebenspartner / in verstorben	87	7,2	7,2	77,7
	ledig	267	22,3	22,3	100,0
	Gesamt	1198	99,8	100,0	
Fehlend	[Keine Angabe]	2	,2		
Gesamt		1200	100,0		

Mit etwa 57% nehmen verheiratete Personen den größten Anteil aller Befragten ein. Lediglich 31 Personen der 1200 Befragten (2,6%) leben in einer eingetragenen Partnerschaft.

Aufgrund teils geringer Häufigkeiten in einzelnen Kategorien der abhängigen und unabhängigen Variablen und zur einfacheren Interpretierbarkeit wollen wir in beiden Variablen Zusammenfassungen von Kategorien vornehmen.

In der unabhängigen Variable Familienstand (**D13**) fassen wir die ersten beiden Kategorien (verheiratet und eingetragene Lebenspartnerschaft) zusammen und erhalten so die neue Variable **D13_kat**, welche aus nunmehr vier Kategorien besteht. Dies kann mit folgender Syntax vorgenommen werden:

Syntax:

```
RECODE D13 (3=2) (4=3) (5=4) (1 thru 2=1) INTO D13_kat.
VARIABLE LABELS  D13_kat 'Familienstand (kategorisiert)'.
EXECUTE.

value labels D13_kat
1 'verheiratet/eingetragene Partnerschaft'
2 'geschieden/getrennt'
3 'verwitwet/Partner verstorben'
4 'ledig'
```

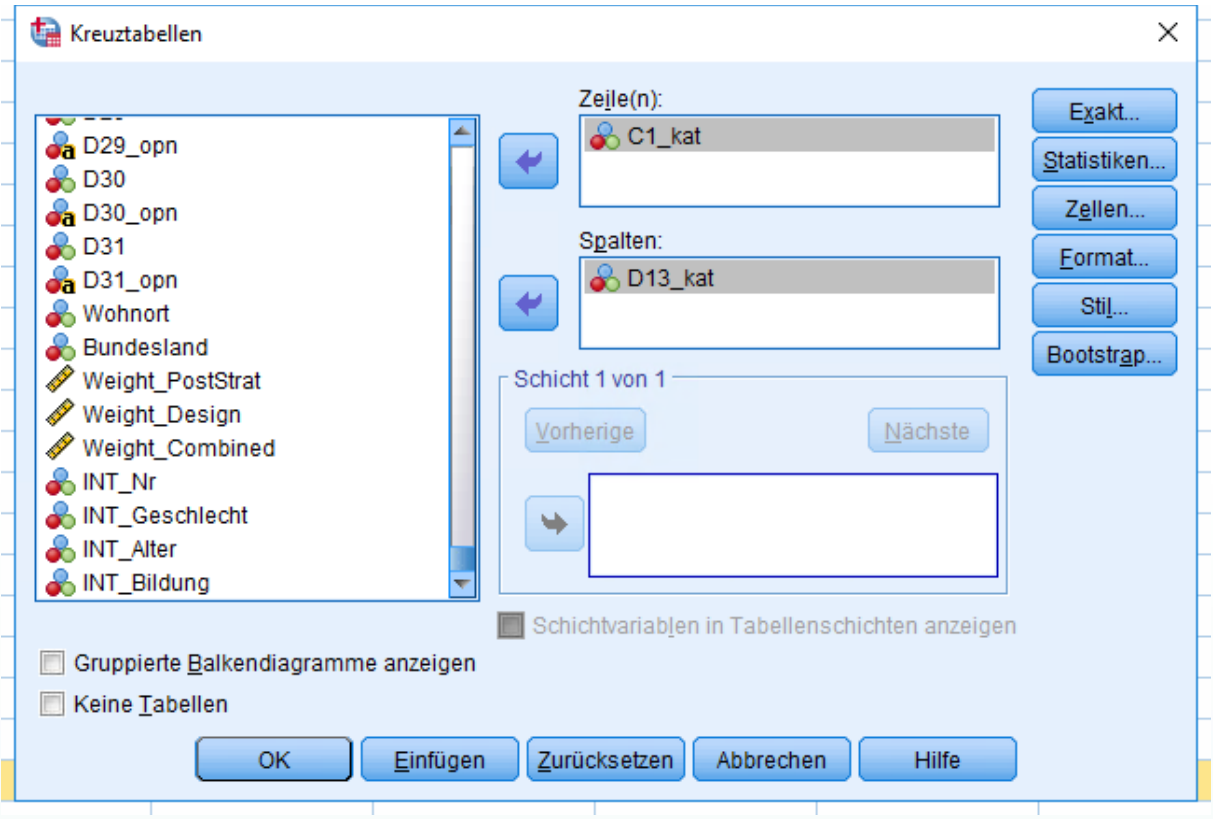
In der abhängigen Variable Happiness (**C1**) fassen wir die ersten beiden sowie die letzten beiden Kategorien zusammen und erhalten so die neue Variable **C1_kat** mit den Ausprägungen 1=(sehr) glücklich und 2=(überhaupt) nicht glücklich. Dies kann mit folgender Syntax vorgenommen werden:

Syntax:

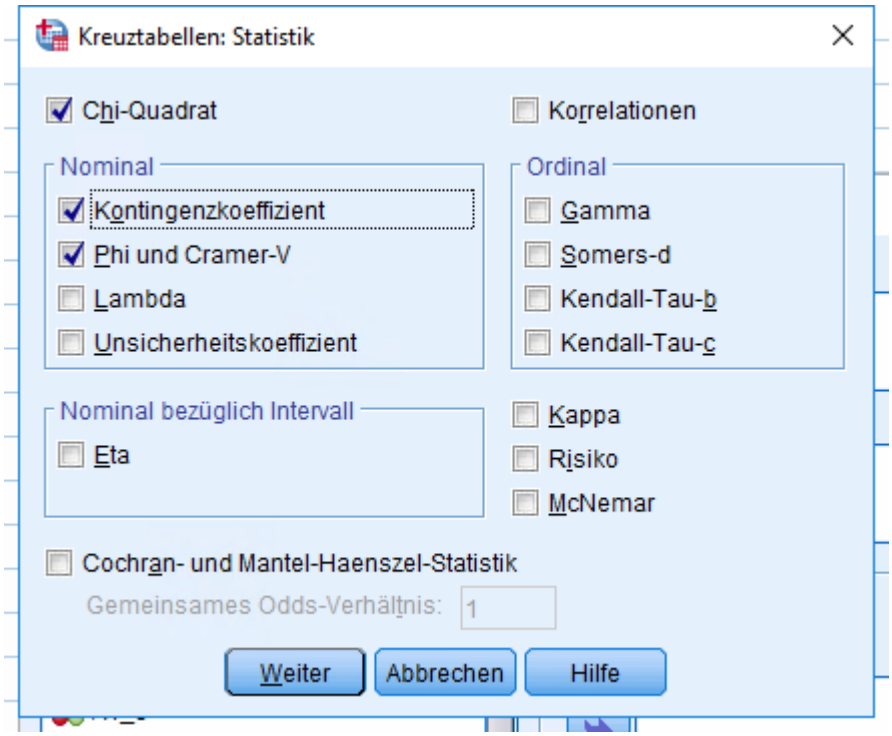
```
RECODE C1 (1 thru 2=1) (3 thru 4=2) INTO C1_kat.
VARIABLE LABELS  C1_kat 'Happiness (kategorisiert)'.
EXECUTE.

value labels C1_kat
1 '(sehr) glücklich'
2 '(überhaupt) nicht glücklich'
```

Durchführung

Nun soll die Kreuztabelle mit den beiden neu berechneten Variablen erstellt werden. Wir verwenden dafür die Befehlskette ***Analysieren → Deskriptive Statistik → Kreuztabelle*** und erhalten folgendes Dialogfeld:

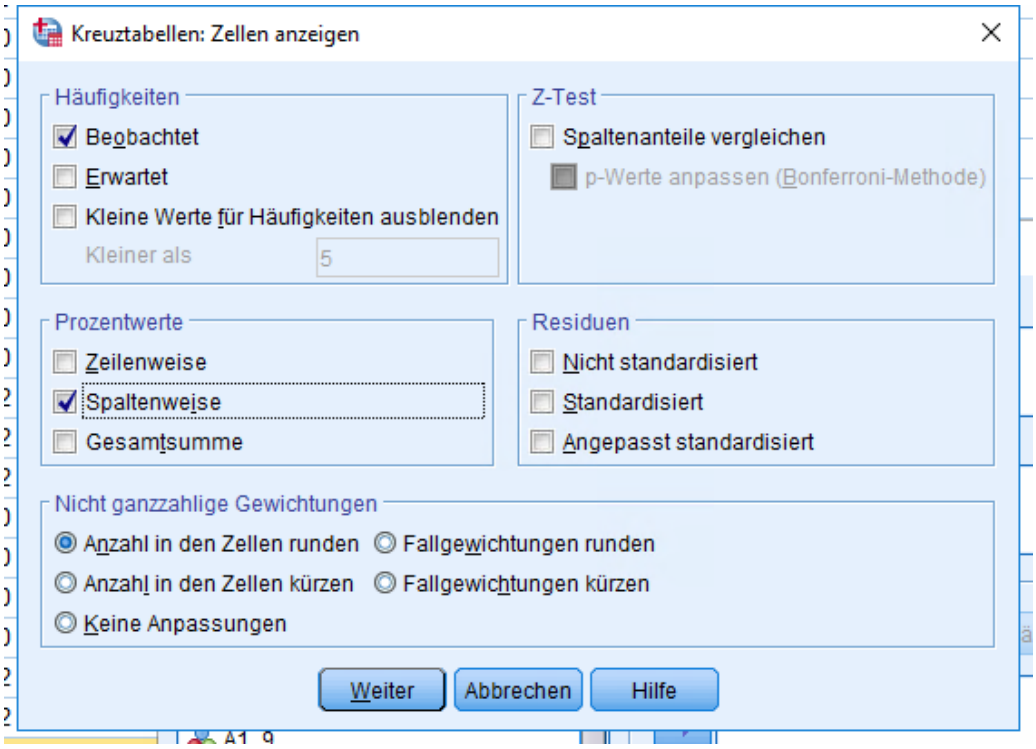


Nun müssen Zeilen- und Spaltenvariablen ausgewählt werden. Wir entscheiden uns dafür, die unabhängige Variable Familienstand (**D13_kat**) als Spaltenvariable zu analysieren und die abhängige Variable Happiness (**C1_kat**) als Zeilenvariable. Unter dem Menüpunkt ‚Statistiken' wollen wir noch weitere Optionen aufrufen.



Ob ein Zusammenhang zwischen Zeilen- und Spaltenvariable auch in der Grundgesamtheit angenommen werden darf, kann mittels Ch^2-Test geprüft werden. Dieser ist im Menüpunkt ‚Statistiken' anzuwählen. Weiters können je nach vorliegendem Skalenniveau verschiedene Zusammenhangsmaße aufgerufen werden. Da unsere unabhängige Variable Familienstand lediglich nominal skaliert ist, kommen nur entsprechende Zusammenhangsmaße für nominale Daten in Frage. In unserem Beispiel wählen wir den ‚Kontingenzkoeffizienten C' sowie ‚Phi und Cramer-V'.

Weiters benötigen wir für die Interpretation der Kreuztabelle neben den absoluten Häufigkeiten der beobachteten Werte auch die relativen Häufigkeiten (Prozentwerte). Diese können im Menüpunkt ‚Zellen' angewählt werden.



Zur Ausgabe von relativen Häufigkeiten bietet SPSS drei Optionen an: Prozentwerte können *zeilenweise, spaltenweise* oder als *Gesamtsumme* ausgegeben werden. Welche Option Sinn macht, hängt von der Fragestellung ab. Soll wie in unserem Fall ein Zusammenhang zwischen einer unabhängigen und einer abhängigen Variable dargestellt werden, so sollte die Werte **in Richtung der unabhängigen Variable** prozentuiert werden. Da wir die unabhängige Variable als Spaltenvariable definiert haben, müssen wir auch **spaltenweise** prozentuieren. Dadurch können die Werte zeilenweise über die verschiedenen Kategorien der unabhängigen Variable verglichen werden. Mit dem ‚Weiter'-Button kommen wir wieder ins Hauptdialogfeld der Kreuztabelle. Die entsprechenden Syntaxbefehle können mit der Option ‚Einfügen' automatisch erstellt werden.

Syntax:

```
CROSSTABS
 /TABLES=C1_kat BY D13_kat
 /FORMAT=AVALUE TABLES
 /STATISTICS=CHISQ CC PHI
 /CELLS=COUNT COLUMN
 /COUNT ROUND CELL.
```

Ergebnisse

Im Ergebnisoutput erhalten wir die nachfolgende Kreuztabelle. Entsprechend unserer Auswahl enthält diese sowohl die absoluten Häufigkeiten der beobachteten Werte als auch die Spaltenprozentwerte. Dementsprechend summieren sich die Prozentwerte pro Spalte in der letzten Zeile jeweils auf 100% auf. Die einzelnen Spalten geben nun die jeweilige Verteilung innerhalb einer Kategorie des Familienstandes wieder. Die letzte Spalte ‚Gesamt' enthält die Verteilung der abhängigen Variable über die gesamte Datenbasis. Diese wird auch als Randverteilung bezeichnet.

Happiness (kategorisiert) * Familienstand (kategorisiert) Kreuztabelle

			Familienstand (kategorisiert)				
			verheiratet/eingetragene Partnerschaft	geschieden/getrennt	verwitwet/Partner verstorben	ledig	Gesamt
Happiness (kategorisiert)	(sehr) glücklich	Anzahl	653	105	73	241	1072
		% innerhalb von Familienstand (kategorisiert)	93,0%	81,4%	88,0%	90,6%	90,8%
	(überhaupt) nicht glücklich	Anzahl	49	24	10	25	108
		% innerhalb von Familienstand (kategorisiert)	7,0%	18,6%	12,0%	9,4%	9,2%
Gesamt		Anzahl	702	129	83	266	1180
		% innerhalb von Familienstand (kategorisiert)	100,0%	100,0%	100,0%	100,0%	100,0%

AUSSDA AUSTRIAN SOCIAL SCIENCE DATA ARCHIVE

Da in Richtung der unabhängigen Variable spaltenweise prozentuiert wurde, können die relativen Häufigkeiten in den Zellen zur Interpretation des Zusammenhangs zeilenweise verglichen werden (Merksatz: Prozentuiere spaltenweise, interpretiere zeilenweise!). Die einzelnen Werte können zeilenweise sowohl untereinander als auch mit der jeweiligen Randverteilung (welche den Gesamtdurchschnitt darstellt) verglichen werden.

So zeigt uns der Blick auf die Randverteilung, dass 90,8% aller Befragten angeben, (sehr) glücklich zu sein. Verheiratete Personen sind mit 93% leicht überdurchschnittlich unter den (sehr) glücklichen, während verwitwete Personen mit 80% leicht unterdurchschnittlich in dieser Kategorie sind. Geschiedene/getrennte Personen hingegen sind mit 81,4% deutlich unterproportional (sehr) glücklich. Fast jede/r fünfte dieser Gruppe gibt an, (überhaupt) nicht glücklich zu sein – deutlich öfter, als in allen anderen Gruppen. Ledige Personen liegen mit 90,6% (sehr) glücklichen nahezu genau im Durchschnitt. Die Prozentsatzdifferenz zwischen verheirateten und geschiedenen Personen beträgt 11,6% und fällt somit deutlich aus. Die Annahme, dass der Familienstand einen Einfluss auf die Happiness hat, wird somit gestützt.

Mit dem Chi²-Test ist nun zu prüfen, ob der beobachtete Zusammenhang in der Stichprobe auch für die Grundgesamtheit angenommen werden kann. Der Chi²-Test stellt die beobachteten Häufigkeiten den erwarteten Häufigkeiten gegenüber, die bei statistischer Unabhängigkeit von Zeilen- und Spaltenvariable auftreten würden. Die Tabelle dieser erwarteten Werte wird Indifferenztabelle [1] genannt und stellt die Nullhypothese des Chi²-Tests dar. Bei völliger Unabhängigkeit von Zeilen- und Spaltenvariable ergibt die Chi²-Prüfgröße Null. Je stärker die Abweichung der beobachteten von den erwarteten Werten, desto stärker ist der Zusammenhang zwischen Zeilen- und Spaltenvariable und desto höher ist die Chi²-Prüfgröße.

In unserem Beispiel erhalten wir einen Chi²-Wert von 18,702, wie aus der zweiten Tabelle im Ergebnisoutput abzulesen ist. Dieser Wert ist bei drei Freiheitsgraden (df), die sich aus einer 8-Felder-Tabelle ergeben, signifikant. Die Irrtumswahrscheinlichkeit p ist in der letzten Spalte abzulesen. Mit einem Wert von 0,000 ist diese kleiner als 1%. Somit kann die Nullhypothese, dass Zeilen- und Spaltenvariable in der Grundgesamtheit unabhängig sind, verworfen werden.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	18,702[a]	3	,000
Likelihood-Quotient	16,074	3	,001
Zusammenhang linear-mit-linear	2,345	1	,126
Anzahl der gültigen Fälle	1180		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 7,60.

[1] Die Indifferenztabelle kann auch in SPSS ausgegeben werden, indem im Menüpunkt ‚Zellen' die Option ‚erwartete Werte' angewählt wird.

Schließlich erhalten wir als Ergebnis den Kontingenzkoeffizienten C sowie Phi und Cramer-V als Zusammenhangsmaße. Diese errechnen sich aus dem Chi²-Wert und sollen als Maß der Zusammenhangsstärke im Wertebereich von 0 und +1 dienen. Der Phi-Koeffizient ist ein Maß für den Zusammenhang zwischen zwei dichotomen Variablen und ist und somit nur bei 4-Felder-Tabellen anzuwenden. Da unsere Tabelle größer ist, kommen nur der Kontingenzkoeffizient C und Cramers V in Frage. Beide Maße normieren im Prinzip den Chi²-Wert auf den Wertebereich 0 bis +1. Zu beachten ist aber, dass der maximal erreichbare Wert von C von der Tabellengröße abhängig ist und somit Tabellen unterschiedlicher Größe nicht miteinander verglichen werden können. Aus diesem Grund kommt häufiger Cramer-V zur Anwendung, da er immer den vollen Wertebereich von 0 bis +1 ausschöpft und bei Tabellen beliebiger Größe zur Anwendung kommen kann. In unserem Beispiel befinden sich aber alle Zusammenhangsmaße im selben Bereich bei etwa ,126. Dies deutet auf eine eher schwache (aber signifikante) Zusammenhangsstärke hin.

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,126	,000
	Cramer-V	,126	,000
	Kontingenzkoeffizient	,125	,000
Anzahl der gültigen Fälle		1180	

Inhaltlich können die Daten also dahingehend interpretiert werden, dass der Familienstand für die Happiness in der österreichischen Bevölkerung durchaus eine Rolle spielt, wenngleich keine besonders starke. Die größte Differenz ergibt sich zwischen verheirateten und geschiedenen/getrennten Personen. Verheiratete sind vergleichsweise häufiger glücklich, geschiedene/getrennte hingegen seltener. Ledige Personen liegen im Durchschnitt.