How to create a fuzzy cross table with SPSS

(Version 1.0)

Ruben P. Konig

Radboud University Nijmegen

Please address all your comments, questions, and requests concerning this paper or the syntax discussed in it to Ruben P. Konig, Department of Communication, Faculty of Social Science, Radboud University Nijmegen, P.O.Box 9104, 6500 HE Nijmegen, The Netherlands; r.konig@ru.nl; +31-24-3615789

Please refer to this paper as: Konig, R. P. (2008). *How to create a fuzzy cross table with SPSS (version 1.0)* [unpublished paper]. Nijmegen, The Netherlands: Radboud University Nijmegen. Retrieved from http://oase.uci.kun.nl/~rkonig/downloads/fuzzyCrosstable.pdf, <fill in date of retrieval>.

Abstract

In this paper I discuss the need to create a crosstable of fuzzy coded variables and how this may be achieved with the help of SPSS's matrix procedure.

Introduction

Suppose you are interested in the relationship between occupational class and newspaper subscription, and suppose that you are aware of the fact that newspaper subscription is not an individual characteristic of your respondents, but a characteristic of their households. At the same time you are aware of the fact that some households consist of only one person, but others consist of more than one person and sometimes more than one of these persons have a job and consequently a score on occupational class. Also suppose that you want to describe the relationship between occupational class and newspaper subscription with a cross table (newspaper subscription and occupational class are nominal variables). How do you proceed?

To measure occupational class at household level, you cannot calculate the mean occupational class of the household members with a job, because occupational class is not an interval level variable. This is a situation where 'fuzzy coding' might bring relief (cf. Murtagh, 2005, pp. 77-92). You could count the respondent's household in part in every applicable category of occupational class. That is, if there are *n* people with a job in a respondent's household, you count 1/*n*-th of that household in the occupational class of every household member with a job.[1] As a consequence, the frequencies of the households' occupational class may not consist of whole numbers only, and the same is true of the cell counts in a cross table of the households' occupational class and newspaper subscription.

Table 1
Fictitious data to illustrate fuzzy coding of households' occupational class and newspaper subscription

occupational classes of household members	household's occupational class				household's newspaper subscription			
	working class	lower middle class	upper middle class	upper class	subscription to newspapers	no paper	popular paper	quality paper
working class	1				popular		1	
upper middle & lower middle class		.5	.5		popular		1	
upper middle class			1		quality			1
upper & upper middle class				.5	quality			1
working class & working class	1				no paper	1		
upper class				1	quality			1
upper middle & working class	.5		.5		quality & popular		.5	.5
lower middle, working & working class	.66	.33			popular		1	
lower middle & upper middle class		.5	.5		popular		1	
lower middle class		1			popular		1	
upper middle & upper class			.5	.5	quality & popular		.5	.5
lower middle & lower middle class		1			quality			1
working, lower middle & upper middle class	.33	.33	.33		quality & popular		.5	.5
upper middle & lower middle class		.5	.5		quality			1
lower middle class		1			no paper	1		
et cetera...								

Table 1 is an illustration of fuzzy coding with some fictitious data. In the first column you see what classes the household members belong to and in the second to fifth column you see how this can be coded fuzzily. Since households are not obliged to subscribe to a newspaper and are not not restricted to subscribe to only one newspaper, the same principle is applied to newspaper subscription as well in the other columns.

But how to make a cross table for such fuzzy coded data? In fact that is fairly simple, but your standard statistical package may not be fitted out for making it seem simple – I know that SPSS is not. The columns two to five and seven to nine in Table 1 constitute two so-called indicator matrices; respectively for the households' occupational class and for the households' newspaper subscription. If you transpose one of these indicator matrices and multiply it with the other, you end up with the desired cross table. And that is not all. With knowledge of some matrix algebra you can also compute statistics such as row and column

percentages adjusted standardized residuals, Pearson's chi-square, and the likelihood ratio chi square.

SPSS syntax

If you do not want to do all this by hand, the following SPSS syntax may be helpful. It uses SPSS' matrix procedure to compute all of the above and display it. It is meticulously annotated with comments to make it possible for everyone who wants to to understand what it does and maybe alter it to make some extra computations. The syntax presupposes that two variables (e.g., A and B) are represented in your data as a set of variables (e.g., catA1 to catA4 and catB1 to catB3, respectively) that each represent a category, like the columns in Table 1. At the start of the matrix procedure these variables are read as the two indicator matrices, that will be processed to create the cross table and compute the statistics. To run this syntax just cut and paste it into your own syntax. Then alter the syntax to fit your specific needs (at least change the bold printed lines) and run it.

```
comment matrix procedure to create a "fuzzy cross table".
matrix.
get A                                                          /*indicator matrix A*/
 /file=*
 /variables=catA1 catA2 catA3 catA4.               /* use any number of categories */
get B                                                          /*indicator matrix A*/
 /file=*
 /variables=catB1 catB2 catB3.                      /* use any number of categories */
compute observed=t(A)*B.                                  /* cross table of A and B */
compute rowsum=rsum(observed).                    /* column vector with row totals */
compute colsum=csum(observed).                    /* row vector with column totals */
compute N=rsum(colsum).                            /* total number of observations */
compute expected=(rowsum*colsum)&/N.                        /* expected frequencies */
compute chisquar=rsum(csum(((observed-expected)&**2)&/expected)).
                                                            /* Pearson's chi-square */
compute df=(ncol(observed)-1)*(nrow(observed)-1).             /* degrees of freedom */
compute as=make(nrow(observed),ncol(observed),0).
        /* preparation for computation of adjusted standardized residual (asresid) */
compute rowperc=make(nrow(observed),ncol(observed),0).
                                  /* preparation for computation of row percentages*/
compute colperc=make(nrow(observed),ncol(observed),0).
                               /* preparation for computation of column percentages*/
compute tosmall=0.   /* preparation for counting cells with expected frequency < 5 */
compute G2=0.        /* preparation for computation of likelihood ratio chi-square */
loop i=1 to nrow(expected) by 1.                    /* process the rows one by one */
+ loop j=1 to ncol(expected) by 1.   /* within row, process the columns one by one */
+ compute as(i,j)=((rowsum(i)*colsum(j)*(1-(rowsum(i)/N))*(1-(colsum(j)/N)))/N)&**.5.
+                         /* part of computation of adjusted standardized residual */
+ do if expected(i,j)<5.
+ compute tosmall=tosmall+1.          /* count cells with expected frequencies < 5 */
```

```
+ end if.
+ do if observed(i,j) <> 0.   /* if cell empty, ln(observed/expected) not defined, */
+               /* but lim n->0 n*ln(n) = 0, so the contribution of this cell is 0 */
+ compute G2=G2+(2*observed(i,j)*ln(observed(i,j)/expected(i,j))).
+                              /* cell contribution to likelihood ratio chi-square */
+ compute rowperc(i,j)=100*observed(i,j)/rowsum(i).              /* row percentage */
+ compute colperc(i,j)=100*observed(i,j)/colsum(j).           /* column percentage */
+ end if.
+ end loop.
end loop.
compute asresid=(observed-expected)&/as.         /* adjusted standardized residual */
print {observed,rowsum;colsum,N}                             /* display the results */
 /format="f10.3"
 /title="crosstable of (fuzzy coded) variables A and B"
 /clabels="catB1" "catB2" "catB3" "total"                /* choose category labels */
 /rlabel="catA1" "catA2" "catA3" "catA4" "total".         /* for columns and  rows */
print {chisquar,df,(1-chicdf(chisquar,df));G2,df,(1-chicdf(G2,df))}
 /format="f10.3"
 /title="chi-square-test of statistical independance"
 /clabels="chi2" "df" "p"
 /rlabels="Pearson" "likelihood ratio".
print {tosmall,(tosmall*100)/(nrow(expected)*ncol(expected))}
 /format="f10.3"
 /title="number of cells with expected frequencies < 5"
 /clabels="count" "%".
print {cmin(rmin(expected))}
 /format="f10.3"
 /title="smallest expected frequency".
print {rowperc,make(nrow(observed),1,100)}
 /format="f10.3"
 /title="row percentages"
 /clabels="catB1" "catB2" "catB3" "total"                /* choose category labels */
 /rlabel="catA1" "catA2" "catA3" "catA4".                 /* for columns and  rows */
print {colperc;make(1,ncol(observed),100)}
 /format="f10.3"
 /title="row percentages"
 /clabels="catB1" "catB2" "catB3"                        /* choose category labels */
 /rlabel="catA1" "catA2" "catA3" "catA4" "total".         /* for columns and  rows */
print asresid
 /format="f10.3"
 /title="adjusted standardized residuals"
 /clabels="catB1" "catB2" "catB3"                        /* choose category labels */
 /rlabel="catA1" "catA2" "catA3" "catA4".                 /* for columns and  rows */
end matrix.
```

Footnotes

[1] Here I am assuming that the occupational class of every household member is equally important in determining the occupational class of the household as a whole. That is not necessarily a good assumption, but for the sake of a simple argument in explaining fuzzy coding it suffices. Of course you can weigh the class of the different household members in any particular way that suits your needs.

References

Murtagh, F. (2005). Correspondence analysis and data coding with Java and R. Boca Raton, FL: Chapman & Hall/CRC.