Behav Res (2012) 44:1191–1196
DOI 10.3758/s13428-012-0200-6

TETRA-COM: A comprehensive SPSS program for estimating the tetrachoric correlation

Urbano Lorenzo-Seva · Pere J. Ferrando

Published online: 5 April 2012

Abstract We provide an SPSS program that implements descriptive and inferential procedures for estimating tetrachoric correlations. These procedures have two main purposes: (1) bivariate estimation in contingency tables and (2) constructing a correlation matrix to be used as input for factor analysis (in particular, the SPSS FACTOR procedure). In both cases, the program computes accurate point estimates, as well as standard errors and confidence intervals that are correct for any population value. For purpose (1), the program computes the contingency table together with five other measures of association. For purpose (2), the program checks the positive definiteness of the matrix, and if it is found not to be Gramian, performs a nonlinear smoothing procedure at the user's request. The SPSS syntax, a short manual, and data files related to this article are available as supplemental materials from brm.psychonomic-journals.org/content/supplemental.

Keywords Tetrachoric correlation · Binary variables · Factor analysis · Difficulty factor

Variables measured on a dichotomous scale are quite common in social science measurement. For example, knowledge of some topic or ability in some domain can be measured by items scored as correct/incorrect, whereas an attitude or a personality trait can be measured by items scored as agree/disagree or endorse/reject. As a third example, a clinical psychologist can diagnose a patient as having/not having a high anxiety level. In the examples provided, the measurements are not inherently dichotomous but, rather, arise from a dichotomization of a continuum at a given threshold. In the first two examples, the continuum would be of knowledge, degree of agreement, and response strength, respectively. In the third, the continuum would be the level of anxiety.

The tetrachoric correlation (Pearson, 1900) is an old measure of association specifically intended for the type of variables illustrated above. It is first assumed that the two dichotomies whose association is to be assessed are actually obtained by dichotomizing truly continuous variables that are not observed. The tetrachoric correlation is, therefore, an estimate of the product–moment correlation that would have been obtained with the underlying continuous variables if its joint distribution were bivariate normal.

In the social science domain, the computation of tetrachoric correlations is of interest mainly in two types of applications. First, it is of interest "per se" as a measure of association in contingency tables to assess, for example, attitude change or rater reliability (e.g., Fleiss, 1981; Guilford & Fruchter, 1973). Second, it can be used as input for a data reduction analysis. The main application of the second type is in factor analysis (FA). However, viewed as a proximity measure, it can also be used in other techniques, such as multidimensional scaling.

Possibly, most of the discussion on the uses and properties of the tetrachoric correlation has been within the framework of FA. The theory of FA has been developed by assuming that (1) the observed variables have continuous, multivariate normal distributions and (2) their relations are linear. If so, their associations are fully summarized by a covariance or a product–moment correlation matrix. If the variables are dichotomous, however, assumption (1) is indeed false and assumption

Other materials: TETRA-COM.SPS; FACTOR.SPS; FLEISS.SAV; BINATET.SAV; R.SAV

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0200-6) contains supplementary material, which is available to authorized users.

U. Lorenzo-Seva · P. J. Ferrando
Universitat Rovira i Virgili,
Tarragona, Spain

U. Lorenzo-Seva (✉)
Centre de Reçerca en Avalució i Mesura de la Conducta,
Departament de Psicologia, Universitat Rovira i Virgili,
Carretera Valls s/n,
43007 Tarragona, Spain
e-mail: urbano.lorenzo@urv.cat

(2) may not be fulfilled, especially if the variables have different splits and are skewed in opposite directions. If this occurs, the product–moment correlation (termed the phi coefficient) is inappropriate, partly because it cannot attain its maximum unit value unless both variables have the same split. As Olsson (1979) pointed out, when variables with widely different splits (some of which are possibly nonlinearly related to the factors) are factor analyzed, two problems can occur: First, the loading estimates can be downwardly biased (attenuated), and second, additional factors with no conceptual interpretation may be needed to obtain an acceptable degree of model–data fit. These additional factors have traditionally been known as *difficulty factors*, although it would be more correct to refer to them as *skew factors* or *factors of curvilinearity* (Greer, Dunlap, & Beatty, 2003; McDonald & Ahlawat, 1974).

If their underlying assumptions are reasonable, the use of tetrachoric corrrelations as input for FA can substantially alleviate the problem: The factor loadings will be more correct and less biased, and the dimensionality of the data will be more accurately assessed (Knol & Berger, 1991; Parry & McArdle, 1991). In fact, the common FA of the tetrachoric correlation matrix is a simple approach—known as the *heuristic solution* (Bock & Lieberman, 1970)—for fitting the nonlinear two-parameter normal ogive model. At present, there are more sophisticated procedures for fitting the two-parameter model (e.g., Bock & Aitkin, 1981; Fraser & McDonald, 1988; Muthén, 1993). However, the simple heuristic solution based on tetrachorics is still an acceptable procedure for evaluating the dimensionality and structure of a set of binary items (Knol & Berger, 1991; Parry & McArdle, 1991). Actually, nowadays it is still widely used in applied research; for example, a search in *Google Scholar* reported 723 documents that used tetrachoric correlation in 2010 and 2011.

While the tetrachoric correlation can be useful for a variety of purposes, it is not free of problems and shortcomings. First, it can be reasonably used only when both variables can be conceived as arising from normally distributed underlying continuous variables. This assumption is not clear in many cases and is totally inappropriate in inherent dichotomies such as gender.

A second shortcoming concerns the reliability of the tetrachoric estimate. As expected, it is generally far less stable than the product–moment estimate obtained from continuous data and becomes more unstable (1) as the variables become more extreme and (2) if the sample is small (McNemar, 1969). Guilford and Fruchter (1973) and McNemar noted that, at the very least, the sample has to be twice the size of the corresponding product–moment for the same accuracy, and they advise using samples of at least 200 or 300 observations.

Factor estimates cannot be stable and accurate if the tetrachoric correlations that serve as the input are not also stable and accurate. In other words, factor-loading estimates computed from tetrachoric correlations could have large standard errors (and standard errors are not computed in a typical exploratory FA program) if the tetrachoric correlations on which they have been based have large standard errors. Muthén (1993) proposed that FA of tetrachorics should be considered as a two-step procedure. So, it is advisable (1) to examine standard errors of tetrachoric correlations (first step analysis) and (2) to estimate factor loadings only if the standard errors in step 1 are low.

Finally, a third shortcoming, which concerns FA applications, is that the tetrachoric correlation matrix is not necessarily positive definite (Gramian). If it is not, it is not computationally appropriate for standard FA methods such as ML and GLS (see Wothke, 1993).

In this article, we present an SPSS program that is based on a comprehensive approach and implements a variety of descriptive and inferential procedures for estimating tetrachoric correlations. As for the most basic procedures, (1) it uses a procedure for estimating the tetrachoric correlation that is both accurate and fast, (2) it computes standard error estimates that are essentially valid for any population value, and (3) it provides essentially correct confidence intervals. It also provides additional features for the two main uses discussed above. As for the first use, the program computes both the contingency table and several measures of agreement based on this table: the associated chi-square, the contingency coefficient, the phi coefficient, the tau-a coefficient, and the kappa index (see, e.g., Liebetrau, 1983). To be used as FA input, the program (1) constructs the correlation matrix, (2) checks its positive definiteness, and, (3) if requested, performs a smoothing procedure that makes the matrix amenable to any FA estimation procedure.

The three basic estimates that are implemented and that are mentioned above—point estimate, standard error, and confidence interval—are those proposed by Bonett and Price (2005; see also Bonett, 2006). The point estimate is a modified cosine-π approximation (Guilford & Fruchter, 1973). This approximation seems to be the most accurate tetrachoric approximation currently available within the class of simple (i.e., easily computed) approximations. The standard error and confidence interval are asymptotic approximations obtained from the distributions of the odds ratio (Agresti, 2002). As for the additional procedures, the smoothing procedure implemented is the one proposed by Devlin, Gnanadesikan, and Kettenring (1975) based on a nonlinear transformation of the elements of the matrix. This smoothing not only makes the matrix usable with all FA procedures, but also does not affect the results obtained with the FA procedures that do not require positive definiteness (Knol & Berger, 1991).

TETRA-COM: A comprehensive SPSS program for tetrachoric correlations

We created an SPSS program to implement the approaches described above. The program runs automatically from the

SPSS (Norusis, 1988) syntax window, and the output can be configured in a variety of ways. We developed the program on the basis of the MATRIX command language (see, e.g., Einspruch, 2003, pp. 137–149). It should be noted, however, that users do not need to know how to program in this language in order to run TETRA-COM; they need only specify the values of some variables in order to configure the analysis of the data at hand. The aim of the program is to produce an *R.dat* file containing the correlation matrix that is ready to be analyzed with FACTOR or ALSCAL modules of SPSS. The Appendix shows the extract of the code in the file *tetra-com.sps*, which can be modified by the user to adapt the syntax. The following computation parameters can be configured: (1) the significance level to be used in the output, (2) whether the cross-tabulation matrices are to be printed in the output or not, (3) whether the smoothing procedure must be computed if the tetrachoric correlation matrix turns out to be non-Gramian, (4) whether the computed correlation matrix is to be analyzed with FACTOR or with ALSCAL modules of SPSS, and (5) the default path for saving the *R.dat* file that contains the computed correlation matrix. To run TETRA-COM, the user has to have an active SPSS data file containing only the variables to be included in the correlation matrix. Once the *R.dat* has been computed, it must be loaded and activated to compute further analysis. For example, to extract a factor using unweighted least squares (ULS), the following SPSS syntax can be used:

```
FACTOR MATRIX IN(COR=*)
  /PRINT  INITIAL KMO ROTATION
  /PLOT EIGEN
  /CRITERIA FACTORS(1) ITERATE(100)
  /EXTRACTION ULS
  /CRITERIA ITERATE(25).
```

Illustrative examples

The first use of the tetrachoric correlation is illustrated with an example taken from Fleiss (1981). The example was also analyzed by Bonnet and Price (2005), which serves to check the accuracy of our results. A total of 100 patients were classified into *neurosis* or *other* categories by two raters A and B. The aim of the study was to assess the degree of rater agreement. Figure 1 shows part of the output provided by TETRA-COM.

The tetrachoric correlation is clearly higher than the other measures of association and, particularly, the product–moment (phi) estimate. The point estimate suggests a notable degree of agreement between both raters. The 95% confidence interval is relatively wide, which is to be expected because (1) the variables have quite extreme splits and (2) the sample is rather small. Note that the upper end of the interval is almost one, which means that the hypothesis of perfect agreement can hardly be rejected. A larger sample would be needed to assess this point in more detail.

The second example is a simulation that illustrates the use and advantages of the tetrachoric correlation matrix as input in FA. First, we simulated the responses of 1,000 individuals on a 10-item test that behaved according to the one common factor model. The original responses were continuous, with a common population factor loading of .70. Next, the responses were dichotomized. The first 5 items were "easy," with a common split of 80 (correct)/20(incorrect), whereas the last 5 items were "difficult," with a reverse split of 20/80. TETRA-COM obtained the tetrachoric matrix from the binary responses and checked that it was positive definite. The correlation matrix was then used as input for the SPSS FACTOR program. The unidimensional solution was fitted by the ULS procedure and provided a good fit to the data. The corresponding loading estimates are shown in Table 1.

For the purpose of comparability, the common factor model with ULS estimation was also fitted to the ordinary product–moment (i.e., phi) correlation matrix by the standard SPSS FACTOR procedure. This time, an unrotated two-factor solution was requested. The outcome factor solution is shown in Table 2.

There are two particular points of interest. First, note that the loadings estimated from the tetrachoric correlations (see Table 1) are essentially unbiased and their values are reasonably close to the "true" population loading of .70. In contrast, the loadings on the first factor based on the phi correlation (see Table 2) are all downwardly biased (i.e., attenuated), and the amount of attenuation is substantial. Second, while the first factor in the bidimensional solution is clearly the content factor, the second factor is artifactual: It is a classical *difficulty* or *skew* factor in which the "easy" items load negatively and the "difficult" items load positively.

Discussion

One of the reasons we decided to develop TETRA-COM was that many fellow applied researchers who were SPSS users frequently needed to factor analyze binary items. They often complained that SPSS did not compute bivariate tetrachoric correlations or correlation matrices to be used in FA.

The procedures implemented in TETRA-COM can be useful not only for factor-analytic researchers, but also for social scientists in general who, sooner or later, need to obtain measures of agreement based on a 2 × 2 contingency table. And as Bonnet and Price (2005) have already noted, the program can also be used for pedagogical purposes. Most classroom presentations and practical sessions use the SPSS package, and the procedures implemented in TETRA-COM can be very useful for teaching the topic of tetrachoric correlation and demonstrating how the extremity of the splits and sample size affect the accuracy and stability of the point estimates.

Fig. 1 TETRA-COM output for the first example

```
Pearson correlation matrix
       x      y
x  1.000   .535
y   .535  1.000

Tetrachoric correlation matrix
       x      y
x  1.000   .835
y   .835  1.000

Alpha value
  .050

Point estimate, Se, and confidence interval for correlation coefficients
 Var vs    Var      r     Se  Lower  Upper
   1.00   2.00    .84    .11    .49    .96

Cross-tabulation count
       0      1
0     89      1
1      6      4

Symmetric Measures
    Chi-Squ    C Coeff        Phi      Tau-a      Kappa   Tetracho
     28.655       .472       .535       .071       .500       .835
```

It should be pointed out that Bonnet and Price's (2005) tetrachoric approximation is a reasonably accurate and easy-to-compute algorithm. A more accurate algorithm is the one proposed by Hamdan (1970), which was subsequently improved by Brown (1977). Implementations of Brown's algorithm are available in Ubersax (http://www.john-uebersax.com/stat/tetra.htm) and Fleming (2005). However, neither of these implementations is available in SPSS. Another alternative is to compute the tetrachoric correlations in SAS.

As was noted in the introduction section, the tetrachoric correlation also has important shortcomings that must be taken into account if it is to be correctly used in applied research. First, the variables must reasonably meet the starting assumptions of artificial dichotomization and underlying normality. As McNemar (1969) noted, the second assumption can be relaxed if we regard the tetrachoric correlation as the estimated correlation when the two measures are normalized. However, at the very least, the assumption that the variables have a linear relation must be plausible. Second, whether it is used in a bivariate application or as input for FA, it is important to ensure that the point estimate is accurate and stable (i.e., that the standard error is small and the confidence interval is narrow). Attaining an acceptable degree of accuracy and stability generally requires certain conditions: (1) Rather large samples (for example, sample sizes larger than 300) should be used, and (2) the splits should not be larger than 80/20 if this procedure is to be used with FA. However, if this degree is not attained, the results derived from the use of tetrachorics can be quite misleading. In addition, the chi-square goodness-of-fit statistic that can be obtained using the maximum likelihood (ML) factor extraction option in SPSS should be interpreted only as a descriptive index when tetrachoric correlations are analyzed, because its inferential interpretation is based on the assumption that the correlations are product–moment and the data are a sample from a multivariate normal distribution. ULS factor extraction seems a more advisable option in this case. We cannot control whether users use the program correctly. However, TETRA-COM provides tools for users to check the accuracy and stability of the estimates before they interpret the results or embark on FA.

Table 1 Unidimensional solution based on the tetrachoric correlation matrix for the second example

Item	Factor 1
1	.786
2	.726
3	.690
4	.766
5	.706
6	.668
7	.750
8	.737
9	.714
10	.802

Table 2 Bidimensional solution based on the phi correlation matrix for the second example

Item	Factor 1	Factor 2
1	.487	−.246
2	.482	−.263
3	.473	−.229
4	.508	−.274
5	.463	−.215
6	.452	.274
7	.473	.209
8	.455	.228
9	.473	.280
10	.484	.279

Program availability

The Appendix shows only a small portion of the SPSS code (i.e., the code lines that can be modified by the user). The SPSS syntax, a short manual, and data files related to this article are available as supplemental materials from brm.psychonomic-journals.org/content/supplemental. Alternatively, these materials can be obtained free of charge by e-mail (urbano.lorenzo@urv.cat).

Acknowledgments The research was partially supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2009SGR1549) and by a grant from the Spanish Ministry of Education and Science (PSI2011-22683), with the collaboration of the European Fund for the Development of Regions.

Appendix

Extract of tetra-com.sps that shows the code lines configurable by the user

```
********************** ENTER HERE THE NAME OF THE FOLDER WHERE THE OUTPUT FILE "R.SAV" MUST BE STORED.
*                              HERE YOU HAVE TWO EXAMPLES.
*
*                              EXAMPLE 1. THE PATH C:\Data IS SPECIFIED:
*                              DEFINE !Path ()"C:\Data\" !enddefine.
*
*                              EXAMPLE 2. THE PATH C:\Users\uls\Desktop\tetra-com IS SPECIFIED:
*                              define !Path ()"C:\Users\uls\Desktop\tetra-com\" !enddefine.
DEFINE !Path ()"C:\Dades\Users\uls\Desktop\tetra-com\" !enddefine.

matrix.
********************** ENTER HERE THE SIGNIFICANTION LEVEL: 0.05, 0.01, or 0.001.
*                           compute alpha = 0.05.
*                           compute alpha = 0.01.
*                           compute alpha = 0.001.
compute alpha = 0.05.

********************** ENTER HERE IF YOU WANT TO INSPECT THE VARIABLES CROSSTABULATION TABLES.
*                          compute crosstab = 1.    The crosstabulation tables are printed.
*                          compute crosstab = 0.    The crosstabulation tables are not printed.
compute crosstab = 1.

********************** ENTER HERE IF THE OUTPUT CORRELATION MATRIX IS TO BE USED WITH "FACTOR" OR WITH
"ALSCAL" MODULES OF SPSS  .
*                          compute output = 1.    To be analyzed with FACTOR molude of SPSS.
*                          compute output = 2.    To be analyzed with ALSCAL module of SPSS.
compute output = 1.

********************** ENTER HERE IF YOU WANT YOU CORRELATION MATRIX BE CONVERTED TO POSITIVE DEFINITE.
********************** To smooth the correlation matrix if it is non positive definite.
*                          compute smooth = 1.      The matrix will be smoothed if necessary.
*                          compute smooth = 0.      The matrix will be smoothed.
compute smooth = 0.
```

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Bonett, D. G. (2006). C479: A new algorithm for the tetrachoric correlation coefficient. *Journal of Statistical Computation and Simulation, 76,* 737–739.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30,* 213–225.

Brown, M. B. (1977). Algorithm AS 116: The tetrachoric correlation and its asymptotic standard error. *Applied Statistics, 26,* 343–351.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika, 62,* 531–545.

Einspruch, E. L. (2003). *Next steps with SPSS*. London: Sage.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Fleming, J. S. (2005). TETCORR: A computer program to compute smoothed tetrachoric correlation matrices. *Behavior Research Methods, 37,* 59–64.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267–269.

Greer, T., Dunlap, W. P., & Beatty, G. O. (2003). A Monte Carlo evaluation of the tetrachoric correlation coefficient. *Educational and Psychological Measurement, 63,* 931–950. doi:10.1177/0013164403251318

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Hamdan, M. A. (1970). The equivalence of tetrachoric and maximum likelihood estimates of *r* in 2 ´ 2 tables. *Biometrika, 57*, 212–215.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparisons between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457–477.

Liebetrau, A. M. (1983). *Measures of association*. London: Sage.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82–99.

McNemar, Q. (1969). *Psychological statistics*. New York: Wiley.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Norusis, N. J. (1988). *The SPSS guide to data analysis for SPSS/PC+*. Chicago, IL: SPSS.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14,* 485–500.

Parry, C. D. H., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15,* 35–46.

Pearson, K. (1900). Mathematical contributions to the theory of evolution in the inheritance of characteristics not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society A, 195,* 79–150.

Wothke, W. (1993). Nonpositive definite matrices in structural models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.