Title

tetrachoric — Tetrachoric correlations for binary variables

Syntax

`tetrachoric` *varlist* [*if*] [*in*] [*weight*] [, *options*]

options	Description
Main	
`stats(`*statlist*`)`	list of statistics; select up to 4 statistics; default is `stats(rho)`
`edwards`	use the noniterative Edwards and Edwards estimator; default is the maximum likelihood estimator
`print(`#`)`	significance level for displaying coefficients
`star(`#`)`	significance level for displaying with a star
`bonferroni`	use Bonferroni-adjusted significance level
`sidak`	use Šidák-adjusted significance level
`pw`	calculate all the pairwise correlation coefficients by using all available data (pairwise deletion)
`zeroadjust`	adjust frequencies when one cell has a zero count
`matrix`	display output in matrix form
`notable`	suppress display of correlations
`posdef`	modify correlation matrix to be positive semidefinite

statlist	Description
`rho`	tetrachoric correlation coefficient
`se`	standard error of rho
`obs`	number of observations
`p`	exact two-sided significance level

`by` is allowed; see [D] **by**.

`fweight`s are allowed; see [U] **11.1.6 weight**.

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Tetrachoric correlations

Description

`tetrachoric` computes estimates of the tetrachoric correlation coefficients of the binary variables in *varlist*. All these variables should be 0, 1, or missing values.

Tetrachoric correlations assume a latent bivariate normal distribution (X_1,X_2) for each pair of variables (v_1,v_2), with a threshold model for the manifest variables, $v_i = 1$ if and only if $X_i > 0$. The means and variances of the latent variables are not identified, but the correlation, r, of X_1 and X_2 can be estimated from the joint distribution of v_1 and v_2 and is called the tetrachoric correlation coefficient.

`tetrachoric` computes pairwise estimates of the tetrachoric correlations by the (iterative) maximum likelihood estimator obtained from bivariate probit without explanatory variables (see [R] **biprobit**) by using the Edwards and Edwards (1984) noniterative estimator as the initial value.

The pairwise correlation matrix is returned as `r(Rho)` and can be used to perform a factor analysis or a principal component analysis of binary variables by using the `factormat` or `pcamat` commands; see [MV] **factor** and [MV] **pca**.

Options

Main

`stats(`*statlist*`)` specifies the statistics to be displayed in the matrix of output. `stats(rho)` is the default. Up to four statistics may be specified. `stats(rho se p obs)` would display the tetrachoric correlation coefficient, its standard error, the significance level, and the number of observations. If *varlist* contains only two variables, all statistics are shown in tabular form. `stats()`, `print()`, and `star()` have no effect unless the `matrix` option is also specified.

`edwards` specifies that the noniterative Edwards and Edwards estimator be used. The default is the maximum likelihood estimator. If you analyze many binary variables, you may want to use the fast noniterative estimator proposed by Edwards and Edwards (1984). However, if you have skewed variables, the approximation does not perform well.

`print(`*#*`)` specifies the maximum significance level of correlation coefficients to be printed. Correlation coefficients with larger significance levels are left blank in the matrix. Typing `tetrachoric ..., print(.10)` would list only those correlation coefficients that are significant at the 10% level or lower.

`star(`*#*`)` specifies the maximum significance level of correlation coefficients to be marked with a star. Typing `tetrachoric ..., star(.05)` would "star" all correlation coefficients significant at the 5% level or lower.

`bonferroni` makes the Bonferroni adjustment to calculated significance levels. This option affects printed significance levels and the `print()` and `star()` options. Thus `tetrachoric ..., print(.05) bonferroni` prints coefficients with Bonferroni-adjusted significance levels of 0.05 or less.

`sidak` makes the Šidák adjustment to calculated significance levels. This option affects printed significance levels and the `print()` and `star()` options. Thus `tetrachoric ..., print(.05) sidak` prints coefficients with Šidák-adjusted significance levels of 0.05 or less.

`pw` specifies that the tetrachoric correlation be calculated by using all available data. By default, `tetrachoric` uses casewise deletion, where observations are ignored if any of the specified variables in *varlist* are missing.

`zeroadjust` specifies that when one of the cells has a zero count, a frequency adjustment be applied in such a way as to increase the zero to one-half and maintain row and column totals.

`matrix` forces `tetrachoric` to display the statistics as a matrix, even if *varlist* contains only two variables. `matrix` is implied if more than two variables are specified.

`notable` suppresses the output.

`posdef` modifies the correlation matrix so that it is positive semidefinite, that is, a proper correlation matrix. The modified result is the correlation matrix associated with the least-squares approximation of the tetrachoric correlation matrix by a positive-semidefinite matrix. If the correlation matrix is modified, the standard errors and significance levels are not displayed and are returned in `r()`.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Association in 2-by-2 tables

Although a wide variety of measures of association in cross tabulations have been proposed, such measures are essentially equivalent (monotonically related) in the special case of 2×2 tables—there is only 1 degree of freedom for nonindependence. Still, some measures have more desirable properties than others. Here we compare two measures: the standard Pearson correlation coefficient and the tetrachoric correlation coefficient. Given asymmetric row or column margins, Pearson correlations are limited to a range smaller than -1 to 1, although tetrachoric correlations can still span the range from -1 to 1. To illustrate, consider the following set of tables for two binary variables, X and Z:

	Z = 0	Z = 1	
X = 0	$20 - a$	$10 + a$	30
X = 1	a	$10 - a$	10
	20	20	40

For a equal to 0, 1, 2, 5, 8, 9, and 10, the Pearson and tetrachoric correlations for the above table are

a	0	1	2	5	8	9	10
Pearson	0.577	0.462	0.346	0	−0.346	−0.462	−0.577
Tetrachoric	1.000	0.792	0.607	0	−0.607	−0.792	−1.000

The restricted range for the Pearson correlation is especially unfortunate when you try to analyze the association between binary variables by using models developed for continuous data, such as factor analysis and principal component analysis.

The tetrachoric correlation of two variables (Y_1, Y_2) can be thought of as the Pearson correlation of two latent bivariate normal distributed variables (Y_1^*, Y_2^*) with threshold measurement models $Y_i = (Y_i^* > c_i)$ for unknown cutpoints c_i. Or equivalently, $Y_i = (Y_i^{**} > 0)$ where the latent bivariate normal (Y_1^{**}, Y_2^{**}) are shifted versions of (Y_1^*, Y_2^*) so that the cutpoints are zero. Obviously, you must judge whether assuming underlying latent variables is meaningful for the data. If this assumption is justified, tetrachoric correlations have two advantages. First, you have an intuitive understanding of the size of correlations that are substantively interesting in your field of research, and this intuition is based on correlations that range from -1 to 1. Second, because the tetrachoric correlation for binary variables estimates the Pearson correlation of the latent continuous variables (assumed multivariate normal distributed), you can use the tetrachoric correlations to analyze multivariate relationships between the dichotomous variables. When doing so, remember that you must interpret the model in terms of the underlying continuous variables.

▷ Example 1

To illustrate tetrachoric correlations, we examine three binary variables from the `familyvalues` dataset (described in example 2).

```
. use http://www.stata-press.com/data/r13/familyvalues
(Attitudes on gender, relationships and family)

. tabulate RS075 RS076

   fam att: │      fam att: trad
   women in │    division of labor
 charge bad │         0          1 │     Total
────────────┼──────────────────────┼──────────
          0 │     1,564        979 │     2,543
          1 │       119        632 │       751
────────────┼──────────────────────┼──────────
      Total │     1,683      1,611 │     3,294

. correlate   RS074 RS075 RS076
(obs=3291)

             │    RS074    RS075    RS076
─────────────┼───────────────────────────
       RS074 │   1.0000
       RS075 │   0.0396   1.0000
       RS076 │   0.1595   0.3830   1.0000

. tetrachoric RS074 RS075 RS076
(obs=3291)

             │    RS074    RS075    RS076
─────────────┼───────────────────────────
       RS074 │   1.0000
       RS075 │   0.0689   1.0000
       RS076 │   0.2480   0.6427   1.0000
```

As usual, the tetrachoric correlation coefficients are larger (in absolute value) and more dispersed than the Pearson correlations.

◁

Factor analysis of dichotomous variables

▷ Example 2

Factor analysis is a popular model for measuring latent continuous traits. The standard estimators are appropriate only for continuous unimodal data. Because of the skewness implied by Bernoulli-distributed variables (especially when the probability is distributed unevenly), a factor analysis of a Pearson correlation matrix can be rather misleading when used in this context. A factor analysis of a matrix of tetrachoric correlations is more appropriate under these conditions (Uebersax 2000). We illustrate this with data on gender, relationship, and family attitudes of spouses using the Households in The Netherlands survey 1995 (Weesie et al. 1995). For attitude variables, it seems reasonable to assume that agreement or disagreement is just a coarse measurement of more nuanced underlying attitudes.

To demonstrate, we examine a few of the variables from the `familyvalues` dataset.

```
. use http://www.stata-press.com/data/r13/familyvalues
(Attitudes on gender, relationships and family)
. describe RS056-RS063

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
RS056           byte    %9.0g                 fam att: should be together
RS057           byte    %9.0g                 fam att: should fight for relat
RS058           byte    %9.0g                 fam att: should avoid conflict
RS059           byte    %9.0g                 fam att: woman better nurturer
RS060           byte    %9.0g                 fam att: both spouses money goo
RS061           byte    %9.0g                 fam att: woman techn school goo
RS062           byte    %9.0g                 fam att: man natural breadwinne
RS063           byte    %9.0g                 fam att: common leisure good
. summarize RS056-RS063

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       RS056 |       3298    .5630685    .4960816          0          1
       RS057 |       3296    .5400485    .4984692          0          1
       RS058 |       3283    .6387451    .4804374          0          1
       RS059 |       3308     .654474    .4756114          0          1
       RS060 |       3302    .3906723     .487975          0          1
-------------+---------------------------------------------------------
       RS061 |       3293    .7102946    .4536945          0          1
       RS062 |       3307    .5857272    .4926705          0          1
       RS063 |       3298    .5379018     .498637          0          1
. correlate RS056-RS063
(obs=3221)

             |    RS056    RS057    RS058    RS059    RS060    RS061    RS062
-------------+---------------------------------------------------------------
       RS056 |   1.0000
       RS057 |   0.1350   1.0000
       RS058 |   0.2377   0.0258   1.0000
       RS059 |   0.1816   0.0097   0.2550   1.0000
       RS060 |  -0.1020  -0.0538  -0.0424   0.0126   1.0000
       RS061 |  -0.1137   0.0610  -0.1375  -0.2076   0.0706   1.0000
       RS062 |   0.2014   0.0285   0.2273   0.4098  -0.0793  -0.2873   1.0000
       RS063 |   0.2057   0.1460   0.1049   0.0911   0.0179  -0.0233   0.0975

             |    RS063
-------------+---------
       RS063 |   1.0000
```

Skewness in these data is relatively modest. For comparison, here are the tetrachoric correlations:

```
. tetrachoric RS056-RS063
(obs=3221)

             |    RS056    RS057    RS058    RS059    RS060    RS061    RS062
-------------+---------------------------------------------------------------
       RS056 |   1.0000
       RS057 |   0.2114   1.0000
       RS058 |   0.3716   0.0416   1.0000
       RS059 |   0.2887   0.0158   0.4007   1.0000
       RS060 |  -0.1620  -0.0856  -0.0688   0.0208   1.0000
       RS061 |  -0.1905   0.1011  -0.2382  -0.3664   0.1200   1.0000
       RS062 |   0.3135   0.0452   0.3563   0.6109  -0.1267  -0.4845   1.0000
       RS063 |   0.3187   0.2278   0.1677   0.1467   0.0286  -0.0388   0.1538

             |    RS063
-------------+---------
       RS063 |   1.0000
```

Again we see that the tetrachoric correlations are generally larger in absolute value than the Pearson correlations. The bivariate probit and Edwards and Edwards estimators (the `edwards` option) implemented in `tetrachoric` may return a correlation matrix that is not positive semidefinite—a mathematical property of any real correlation matrix. Positive definiteness is required by commands for analyses of correlation matrices, such as `factormat` and `pcamat`; see [MV] **factor** and [MV] **pca**. The `posdef` option of `tetrachoric` tests for positive definiteness and projects the estimated correlation matrix to a positive-semidefinite matrix if needed.

```
. tetrachoric RS056-RS063, notable posdef
. matrix C = r(corr)
```

This time, we suppressed the display of the correlations with the `notable` option and requested that the correlation matrix be positive semidefinite with the `posdef` option. Had the correlation matrix not been positive definite, `tetrachoric` would have displayed a warning message and then adjusted the matrix to be positive semidefinite. We placed the resulting tetrachoric correlation matrix into a matrix, `C`, so that we can perform a factor analysis upon it.

`tetrachoric` with the `posdef` option asserted that `C` was positive definite because no warning message was displayed. We can verify this by using a familiar characterization of symmetric positive-definite matrices: all eigenvalues are real and positive.

```
. matrix symeigen eigenvectors eigenvalues = C
. matrix list eigenvalues
eigenvalues[1,8]
            e1          e2          e3          e4          e5          e6          e7
r1   2.5974789   1.3544664   1.0532476   .77980391   .73462018   .57984565   .54754512
            e8
r1   .35299228
```

We can proceed with a factor analysis on the matrix `C`. We use `factormat` and select iterated principal factors as the estimation method; see [MV] **factor**.

```
. factormat C, n(3221) ipf factor(2)
(obs=3221)
Factor analysis/correlation                        Number of obs    =     3221
    Method: iterated principal factors             Retained factors =        2
    Rotation: (unrotated)                          Number of params =       15

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.06855      1.40178            0.7562       0.7562
        Factor2  |      0.66677      0.47180            0.2438       1.0000
        Factor3  |      0.19497      0.06432            0.0713       1.0713
        Factor4  |      0.13065      0.10967            0.0478       1.1191
        Factor5  |      0.02098      0.10085            0.0077       1.1267
        Factor6  |     -0.07987      0.01037           -0.0292       1.0975
        Factor7  |     -0.09024      0.08626           -0.0330       1.0645
        Factor8  |     -0.17650            .           -0.0645       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) = 4620.01 Prob>chi2 = 0.0000
Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
           RS056 |   0.5528    0.4120 |      0.5247
           RS057 |   0.1124    0.4214 |      0.8098
           RS058 |   0.5333    0.0718 |      0.7105
           RS059 |   0.6961   -0.1704 |      0.4865
           RS060 |  -0.1339   -0.0596 |      0.9785
           RS061 |  -0.5126    0.2851 |      0.6560
           RS062 |   0.7855   -0.2165 |      0.3361
           RS063 |   0.2895    0.3919 |      0.7626
    -------------------------------------------------
```

◁

▷ Example 3

We noted in example 2 that the matrix of estimates of the tetrachoric correlation coefficients need not be positive definite. Here is an example:

```
. use http://www.stata-press.com/data/r13/familyvalues
(Attitudes on gender, relationships and family)
. tetrachoric RS056-RS063 in 1/20, posdef
(obs=18)
matrix with tetrachoric correlations is not positive semidefinite;
  it has 2 negative eigenvalues
  maxdiff(corr,adj-corr) =  0.2346
  (adj-corr: tetrachoric correlations adjusted to be positive semidefinite)
    adj-corr |    RS056    RS057    RS058    RS059    RS060    RS061    RS062
-------------+---------------------------------------------------------------
       RS056 |   1.0000
       RS057 |   0.5284   1.0000
       RS058 |   0.3012   0.2548   1.0000
       RS059 |   0.3251   0.2791   0.0550   1.0000
       RS060 |  -0.5197  -0.4222  -0.7163   0.0552   1.0000
       RS061 |   0.3448   0.4815  -0.0958  -0.1857  -0.0980   1.0000
       RS062 |   0.1066  -0.0375   0.0072   0.3909  -0.2333  -0.7654   1.0000
       RS063 |   0.3830   0.4939   0.4336   0.0075  -0.8937  -0.0337   0.4934
    adj-corr |    RS063
-------------+---------
       RS063 |   1.0000
```

```
. mata:
------------------------------------------------- mata (type end to exit) ------
: C2 = st_matrix("r(corr)")
: eigenvecs = .
: eigenvals = .
: symeigensystem(C2, eigenvecs, eigenvals)
: eigenvals
                  1              2              3              4
    +-------------------------------------------------------------+
  1 |   3.156592567    2.065279398    1.324911199    .7554904485  |
    +-------------------------------------------------------------+
                  5              6              7              8
    +-------------------------------------------------------------+
  1 |   .4845368741    .2131895139   -8.80914e-19   -1.90196e-16  |
    +-------------------------------------------------------------+

: end
--------------------------------------------------------------------------------
```

The estimated tetrachoric correlation matrix is rank-2 deficient. With this `C2` matrix, we can only use models of correlation that allow for singular cases.

◁

Tetrachoric correlations with simulated data

▷ Example 4

We use `drawnorm` (see [D] **drawnorm**) to generate a sample of 1,000 observations from a bivariate normal distribution with means −1 and 1, unit variances, and correlation 0.4.

```
. clear
. set seed 11000
. matrix m = (1, -1)
. matrix V = (1, 0.4 \ 0.4, 1)
. drawnorm c1 c2, n(1000) means(m) cov(V)
(obs 1000)
```

Now consider the measurement model assumed by the tetrachoric correlations. We observe only whether `c1` and `c2` are greater than zero,

```
. generate d1 = (c1 > 0)
. generate d2 = (c2 > 0)
. tabulate d1 d2
```

d1	d2: 0	d2: 1	Total
0	176	6	182
1	656	162	818
Total	832	168	1,000

We want to estimate the correlation of `c1` and `c2` from the binary variables `d1` and `d2`. Pearson's correlation of the binary variables `d1` and `d2` is 0.170—a seriously biased estimate of the underlying correlation $\rho = 0.4$.

```
. correlate d1 d2
(obs=1000)

             |       d1       d2
-------------+------------------
          d1 |   1.0000
          d2 |   0.1704   1.0000
```

The tetrachoric correlation coefficient of `d1` and `d2` estimates the Pearson correlation of the latent continuous variables, `c1` and `c2`.

```
. tetrachoric d1 d2

   Number of obs =     1000
 Tetrachoric rho =       0.4790
       Std error =       0.0700

Test of Ho: d1 and d2 are independent
 2-sided exact P =       0.0000
```

The estimate of the tetrachoric correlation of `d1` and `d2`, 0.4790, is much closer to the underlying correlation, 0.4, between `c1` and `c2`.

◁

Stored results

`tetrachoric` stores the following in `r()`:

Scalars

`r(rho)`	tetrachoric correlation coefficient between variables 1 and 2
`r(N)`	number of observations
`r(nneg)`	number of negative eigenvalues (`posdef` only)
`r(se_rho)`	standard error of `r(rho)`
`r(p)`	exact two-sided significance level

Macros

`r(method)`	estimator used

Matrices

`r(Rho)`	tetrachoric correlation matrix
`r(Se_Rho)`	standard errors of `r(Rho)`
`r(Nobs)`	number of observations used in computing correlation
`r(P)`	exact two-sided significance level matrix

Methods and formulas

`tetrachoric` provides two estimators for the tetrachoric correlation ρ of two binary variables with the frequencies n_{ij}, $i, j = 0, 1$. `tetrachoric` defaults to the slower (iterative) maximum likelihood estimator obtained from bivariate probit without explanatory variables (see [R] **biprobit**) by using the Edwards and Edwards noniterative estimator as the initial value. A fast (noniterative) estimator is also available by specifying the `edwards` option (Edwards and Edwards 1984; Digby 1983)

$$\widehat{\rho} = \frac{\alpha - 1}{\alpha + 1}$$

where

$$\alpha = \left(\frac{n_{00}n_{11}}{n_{01}n_{10}}\right)^{\pi/4} \quad (\pi = 3.14\ldots)$$

if all $n_{ij} > 0$. If $n_{00} = 0$ or $n_{11} = 0$, $\widehat{\rho} = -1$; if $n_{01} = 0$ or $n_{10} = 0$, $\widehat{\rho} = 1$.

The asymptotic variance of the Edwards and Edwards estimator of the tetrachoric correlation is easily obtained by the delta method,

$$\text{avar}(\widehat{\rho}) = \left(\frac{\pi\alpha}{2(1+\alpha)^2}\right)^2 \left(\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}\right)$$

provided all $n_{ij} > 0$, otherwise it is left undefined (missing). The Edwards and Edwards estimator is fast, but may be inaccurate if the margins are very skewed.

`tetrachoric` reports exact p-values for statistical independence, computed by the `exact` option of [R] **tabulate twoway**.

References

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

Brown, M. B., and J. K. Benedetti. 1977. On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika* 42: 347–355.

Digby, P. G. N. 1983. Approximating the tetrachoric correlation coefficient. *Biometrics* 39: 753–757.

Edwards, J. H., and A. W. F. Edwards. 1984. Approximating the tetrachoric correlation coefficient. *Biometrics* 40: 563.

Golub, G. H., and C. F. Van Loan. 1996. *Matrix Computations*. 3rd ed. Baltimore: Johns Hopkins University Press.

Uebersax, J. S. 2000. Estimating a latent trait model by factor analysis of tetrachoric correlations. http://ourworld.compuserve.com/homepages/jsuebersax/irt.htm.

Weesie, J., M. Kalmijn, W. Bernasco, and D. Giesen. 1995. *Households in The Netherlands 1995*. Utrecht, Netherlands: Datafile, ISCORE, University of Utrecht.

Also see

[R] **biprobit** — Bivariate probit regression

[R] **correlate** — Correlations (covariances) of variables or coefficients

[R] **spearman** — Spearman's and Kendall's correlations

[R] **tabulate twoway** — Two-way table of frequencies

[MV] **factor** — Factor analysis

[MV] **pca** — Principal component analysis