CROSSTABS

The notation and statistics refer to bivariate subtables defined by a row variable X and a column variable Y, unless specified otherwise. By default, CROSSTABS deletes cases with missing values on a table-by-table basis.

Notation

The following notation is used throughout this chapter unless otherwise stated:

X_i	Distinct values of row variable arranged in ascending order: $X_1 < X_2 < \cdots < X_R$
Y_j	Distinct values of column variable arranged in ascending order: $Y_1 < Y_2 < \cdots < Y_C$
f_{ij}	Sum of cell weights for cases in cell (i, j)
c_j	$\sum_{i=1}^{R} f_{ij}$, the jth column subtotal
r_i	$\sum_{j=1}^{C} f_{ij}$, the ith row subtotal
W	$\sum_{j=1}^{C} c_j = \sum_{i=1}^{R} r_i$, the grand total

Marginal and Cell Statistics

Count

$$\text{count} = f_{ij}$$

Expected Count

$$E_{ij} = \frac{r_i c_j}{W}$$

Row Percent

$$\text{row percent} = 100 \times \left(f_{ij} / r_i \right)$$

Column Percent

$$\text{column percent} = 100 \times \left(f_{ij} / c_j \right)$$

Total Percent

$$\text{total percent} = 100 \times \left(f_{ij} / W \right)$$

Residual

$$R_{ij} = f_{ij} - E_{ij}$$

Standardized Residual

$$SR_{ij} = \frac{R_{ij}}{\sqrt{E_{ij}}}$$

Adjusted Residual

$$AR_{ij} = \frac{R_{ij}}{\sqrt{E_{ij}\left(1-\frac{r_i}{W}\right)\left(1-\frac{c_j}{W}\right)}}$$

Chi-Square Statistics

Pearson's Chi-Square

$$\chi_p^2 = \sum_{ij} \frac{\left(f_{ij}-E_{ij}\right)^2}{E_{ij}}$$

The degrees of freedom are $(R-1)(C-1)$.

Likelihood Ratio

$$\chi_{LR}^2 = -2\sum_{ij} f_{ij} \ln\left(E_{ij}/f_{ij}\right)$$

The degrees of freedom are $(R-1)(C-1)$.

Fisher's Exact Test

If the table is a 2×2 table, not resulting from a larger table with missing cells, with at least one expected cell count less than 5, then the Fisher exact test is calculated. See Appendix 5 for details.

Yates Continuity Corrected for 2 x 2 Tables

$$\chi_c^2 = \begin{cases} \dfrac{W\left(\left|f_{11}f_{22} - f_{12}f_{21}\right| - 0.5W\right)^2}{r_1 r_2 c_1 c_2} & \text{if } \left|f_{11}f_{22} - f_{12}f_{21}\right| > 0.5\text{W} \\ 0 & \text{otherwise} \end{cases}$$

The degrees of freedom are 1.

Mantel-Haenszel Test of Linear Association

$$\chi_{MH}^2 = (W-1)r^2$$

where r is the Pearson correlation coefficient to be defined later. The degrees of freedom are 1.

Other Measures of Association

Phi Coefficient

For a table not 2×2

$$\varphi = \sqrt{\frac{\chi_p^2}{W}}$$

For a 2×2 table only, φ is equal to the Pearson correlation coefficient so that the sign of φ matches that of the correlation coefficients.

Coefficient of Contingency

$$CC = \left(\frac{\chi_p^2}{\chi_p^2 + W} \right)^{1/2}$$

Cramér's V

$$V = \left(\frac{\chi_p^2}{W(q-1)} \right)^{1/2}$$

where $q = \min\{R, C\}$.

Measures of Proportional Reduction in Predictive Error

Lambda

Let f_{im} and f_{mj} be the largest cell count in row *i* and column *j*, respectively. Also, let r_m be the largest row subtotal and c_m the largest column subtotal. Define $\lambda_{Y|X}$ as the proportion of relative error in predicting an individual's *Y* category that can be eliminated by knowledge of the *X* category. $\lambda_{Y|X}$ is computed as

$$\lambda_{Y|X} = \frac{\sum_{i=1}^{R} f_{im} - c_m}{W - c_m}$$

The standard errors are

$$ASE_0 = \frac{\sqrt{\left. \sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij}\left(\delta_{ij} - \delta_j\right)^2 - \left(\sum_{i=1}^{R} f_{im} - c_m\right)^2 \middle/ W \right.}}{W - c_m}$$

$$ASE_1 = \frac{\sqrt{\sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij}\left(\delta_{ij} - \delta_j + \lambda\delta_j\right)^2 - W\lambda_{Y|X}}}{W - c_m}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } j \text{ is column index for } f_{im} \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_j = \begin{cases} 1 & \text{if } j \text{ is index for } c_m \\ 0 & \text{otherwise} \end{cases}$$

Lambda for predicting *X* from *Y*, $\lambda_{Y|X}$, is obtained by permuting the indices in the above formulae.

The two asymmetric lambdas are averaged to obtain the symmetric lambda.

$$\lambda = \frac{\sum_{i=1}^{R} f_{im} + \sum_{j=1}^{C} f_{mj} - c_m - r_m}{2W - r_m - c_m}$$

The standard errors are

$$ASE_0 = \frac{\sqrt{\sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij}\left(\delta_{ij}^{r} + \delta_{ij}^{c} - \delta_{i}^{r} - \delta_{j}^{c}\right)^2 - \left[\left(\sum_{i=1}^{R} f_{im} + \sum_{j=1}^{C} f_{mj} - c_m - r_m\right)^2 \Big/ W\right]}}{2W - r_m - c_m}$$

$$ASE_1 = \frac{\sqrt{\sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij}\left[\delta_{ij}^{r} + \delta_{ij}^{c} - \delta_{i}^{r} - \delta_{j}^{c} + \lambda\left(\delta_{i}^{r} + \delta_{j}^{c}\right)\right]^2 - 4W\lambda^2}}{2W - r_m - c_m}$$

where

$$\delta_{ij}^{r} = \begin{cases} 1 & \text{if } i \text{ is row index for } f_{mj} \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{i}^{r} = \begin{cases} 1 & \text{if } i \text{ is index for } r_m \\ 0 & \text{otherwise} \end{cases}$$

and where

$$\delta_{ij}^{c} = \begin{cases} 1 & \text{if } j \text{ is column index for } f_{im} \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{i}^{c} = \begin{cases} 1 & \text{if } j \text{ is index for } c_m \\ 0 & \text{otherwise} \end{cases}$$

Goodman and Kruskal's Tau (Goodman & Kruskal, 1954)

Similarly defined is Goodman and Kruskal's tau (τ):

$$\tau_{Y|X} = \frac{W \sum_{i,j} \left(f_{ij}^2 / r_i \right) - \sum_{j=1}^{C} c_j^2}{W^2 - \sum_{j=1}^{C} c_j^2}$$

with standard error

$$ASE_1 = \sqrt{\frac{4}{\delta^4} \sum_{i,j} f_{ij} \left\{ (v - \delta) \left(\frac{1}{r_i} \sum_{j=1}^{C} f_{ij} c_j - c_j \right) - W\delta \left(\frac{1}{r_i^2} \sum_{j=1}^{C} f_{ij}^2 - \frac{1}{r_i} f_{ij} \right) \right\}^2}$$

in which

$$\delta = W^2 - \sum_{j=1}^{C} c_j^2 \quad \text{and} \quad v = W \sum_{i,j} f_{ij}^2 / r_i - \sum_{j=1}^{C} c_j^2$$

$\tau_{X|Y}$ and its standard error can be obtained by interchanging the roles of *X* and *Y*.

The significance level is based on the chi-square distribution, since

$$(W-1)(C-1)\tau_{Y|X} \sim \chi^2_{(R-1)(C-1)}$$

$$(W-1)(R-1)\tau_{X|Y} \sim \chi^2_{(R-1)(C-1)}$$

Uncertainty Coefficient

Let $U_{Y|X}$ be the proportional reduction in the uncertainty (entropy) of Y that can be eliminated by knowledge of X. It is computed as

$$U_{Y|X} = \frac{U(X) + U(Y) - U(XY)}{U(Y)}$$

where

$$U(X) = -\sum_{i=1}^{R} \frac{r_i}{W} \ln\left(\frac{r_i}{W}\right)$$

$$U(Y) = -\sum_{j=1}^{C} \frac{c_j}{W} \ln\left(\frac{c_j}{W}\right)$$

and

$$U(XY) = -\sum_{i,j} \frac{f_{ij}}{W} \ln\left(\frac{f_{ij}}{W}\right), \quad \text{for } f_{ij} > 0$$

The asymptotic standard errors are

$$ASE_1 = \frac{1}{WU(Y)^2} \sqrt{\sum_{i,j} f_{ij} \left\{ U(Y) \ln\left(\frac{f_{ij}}{r_i}\right) + \left[U(X) - U(XY)\right] \ln\left(\frac{c_j}{W}\right) \right\}^2}$$

$$ASE_0 = \frac{\sqrt{P - W\left[U(X) + U(Y) - U(XY)\right]^2}}{\left[WU(Y)\right]}$$

where

$$P = \sum_{i,j} f_{ij} \ln\left(\frac{c_j r_i}{W f_{ij}}\right)^2$$

The formulas for $U_{X|Y}$ can be obtained by interchanging the roles of *X* and *Y*.

A symmetric version of the two asymmetric uncertainty coefficients is defined as follows:

$$U = 2\left[\frac{U(X)+U(Y)-U(XY)}{U(X)+U(Y)}\right]$$

with asymptotic standard errors

$$ASE_1 = \frac{2}{W\left[U(X)+U(Y)\right]^2}\sqrt{\sum_{i,j} f_{ij}\left\{U(XY)\ln\left(\frac{r_i c_j}{W^2}\right)-\left[U(X)+U(Y)\right]\ln\left(\frac{f_{ij}}{W}\right)\right\}^2}$$

or

$$ASE_0 = \frac{2}{W\left[U(X)+U(Y)\right]}\sqrt{P-\left[U(X)+U(Y)-U(XY)\right]^2 / W}$$

Cohen's Kappa

Cohen's kappa (κ), defined only for square table $(R = C)$, is computed as

$$\kappa = \frac{W\sum_{i=1}^{R} f_{ii} - \sum_{i=1}^{R} r_i c_i}{W^2 - \sum_{i=1}^{R} r_i c_i}$$

with variance

$$\mathrm{var}_1 = W\left\{\frac{\left(\sum f_{ii}\right)\left(W-\sum f_{ii}\right)}{\left(W^2-\sum r_i c_i\right)^2}+\frac{2\left(W-\sum f_{ii}\right)\left(2\sum f_{ii}\sum r_i c_i - W\sum f_{ii}\left(r_i+c_i\right)\right)}{\left(W^2-\sum r_i c_i\right)^3}\right.$$

$$\left.+\frac{\left(W-\sum f_{ii}\right)^2\left[W\sum_{i,j} f_{ij}\left(r_j+c_i\right)^2-4\left(\sum r_i c_i\right)^2\right]}{\left(W^2-\sum r_i c_i\right)^4}\right\}$$

$$\mathrm{var}_0 = \frac{1}{W\left(W^2-\sum_i r_i c_i\right)^2}\left[W^2\left(\sum_i r_i c_i\right)+\left(\sum_i r_i c_i\right)^2-W\left(\sum_i r_i c_i\left(r_i+c_i\right)\right)\right]$$

Kendall's Tau-*b* and Tau-*c*

Define

$$D_r = W^2 - \sum_{i=1}^{R} r_i^2$$

$$D_c = W^2 - \sum_{j=1}^{C} c_j^2$$

$$C_{ij} = \sum_{h<i} \sum_{k<j} f_{hk} + \sum_{h>i} \sum_{k>j} f_{hk}$$

$$D_{ij} = \sum_{h<i} \sum_{k>j} f_{hk} + \sum_{h>i} \sum_{k<j} f_{hk}$$

$$P = \sum_{i,j} f_{ij} C_{ij}$$

$$Q = \sum_{i,j} f_{ij} D_{ij}$$

Note: the P and Q listed above are double the "usual" P (number of concordant pairs) and Q (number of discordant pairs). Likewise, D_r is double the "usual" $P + Q + X_0$ (the number of concordant pairs, discordant pairs, and pairs on which the row variable is tied) and D_c is double the "usual" $P + Q + Y_0$ (the number of concordant pairs, discordant pairs, and pairs on which the column variable is tied).

Kendall's Tau-*b*

$$\tau_b = \frac{P - Q}{\sqrt{D_r D_c}}$$

with standard error

$$ASE_1 = \frac{1}{(D_r D_c)} \sqrt{\sum_{i,j} f_{ij} \left(2\sqrt{D_r D_c} \left(C_{ij} - D_{ij} \right) + \tau_b v_{ij} \right)^2 - W^3 \tau_b^2 \left(D_r + D_c \right)^2}$$

where

$$v_{ij} = r_i D_c + c_j D_r$$

Under the independence assumption, the standard error is

$$ASE_0 = 2\sqrt{\frac{\sum_{i,j} f_{ij} \left(C_{ij} - D_{ij} \right)^2 - \frac{1}{W} (P - Q)^2}{D_r D_c}}$$

Kendall's Tau-*c*

$$\tau_c = \frac{q(P - Q)}{W^2 (q - 1)}$$

with standard error

$$ASE_1 = \frac{2q}{(q-1)W^2} \sqrt{\sum_{i,j} f_{ij} \left(C_{ij} - D_{ij} \right)^2 - \frac{1}{W} (P - Q)^2}$$

or, under the independence assumption,

$$ASE_0 = \frac{2q}{(q-1)W^2} \sqrt{\sum_{i,j} f_{ij} \left(C_{ij} - D_{ij} \right)^2 - \frac{1}{W} (P - Q)^2}$$

where

$$q = \min\{R, C\}$$

Gamma

Gamma (γ) is estimated by

$$\gamma = \frac{P-Q}{P+Q}$$

with standard error

$$ASE_1 = \frac{4}{(P+Q)^2}\sqrt{\sum_{i,j} f_{ij}\left(QC_{ij} - PD_{ij}\right)^2}$$

or, under the hypothesis of independence,

$$ASE_0 = \frac{2}{(P+Q)}\sqrt{\sum_{i,j} f_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{W}(P-Q)^2}$$

Somers' *d*

Somers' *d* with row variable *X* as the independent variable is calculated as

$$d_{Y|X} = \frac{P-Q}{D_r}$$

with standard error

$$ASE_1 = \frac{2}{D_r^2}\sqrt{\sum_{i,j} f_{ij}\left\{D_r\left(C_{ij} - D_{ij}\right) - (P-Q)(W-R_i)\right\}^2}$$

or, under the hypothesis of independence,

$$ASE_0 = \frac{2}{D_r}\sqrt{\sum_{i,j} f_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{W}\left(P - Q\right)^2}$$

By interchanging the roles of *X* and *Y*, the formulas for Somers' *d* with *X* as the dependent variable can be obtained.

Symmetric version of Somers' *d* is

$$d = \frac{\left(P - Q\right)}{\frac{1}{2}\left(D_c + D_r\right)}$$

The standard error is

$$ASE_1 = \frac{2\sigma_{\tau_b}^2}{\left(D_r + D_c\right)}\sqrt{D_r D_c}$$

where $\sigma_{\tau_b}^2$ is the variance of Kendall's τ_b,

$$ASE_0 = \frac{4}{\left(D_c + D_r\right)}\sqrt{\sum_{i,j} f_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{W}\left(P - Q\right)^2}$$

Pearson's *r*

The Pearson's product moment correlation *r* is computed as

$$r = \frac{\mathrm{cov}\left(X, Y\right)}{\sqrt{S\left(X\right)S\left(Y\right)}} \equiv \frac{S}{T}$$

where

$$\operatorname{cov}(X, Y) = \sum_{i,j} X_i Y_j f_{ij} - \left(\sum_{i=1}^{R} X_i r_i \right) \left(\sum_{j=1}^{C} Y_j c_j \right) \Big/ W$$

$$S(X) = \sum_{i=1}^{R} X_i^2 r_i - \left(\sum_{i=1}^{R} X_i r_i \right)^2 \Big/ W$$

and

$$S(Y) = \sum_{j=1}^{C} Y_j^2 c_j - \left(\sum_{j=1}^{C} Y_j c_j \right)^2 \Big/ W$$

The variance of r is

$$\operatorname{var}_1 = \frac{1}{T^4} \sum_{i,j} f_{ij} \left\{ T\left(X_i - \overline{X} \right)\left(Y_j - \overline{Y} \right) - \frac{S}{2T} \left[\left(X_i - \overline{X} \right)^2 S(Y) + \left(Y_j - \overline{Y} \right)^2 S(X) \right] \right\}^2$$

If the null hypothesis is true,

$$\operatorname{var}_0 = \frac{\sum_{i,j} f_{ij} X_i^2 Y_j^2 - \left(\sum_{i,j} f_{ij} X_i Y_j \right)^2 \Big/ W}{\left(\sum_i r_i X_i^2 \right) \left(\sum_j c_j Y_j^2 \right)}$$

where

$$\overline{X} = \sum_{i=1}^{R} X_i r_i / W$$

and

$$\bar{Y} = \sum_{j=1}^{C} Y_j c_j \Big/ W$$

Under the hypothesis that $\rho = 0$,

$$t = \frac{r\sqrt{W-2}}{\sqrt{1-r^2}}$$

is distributed as a *t* with $W-2$ degrees of freedom.

Spearman Correlation

The Spearman's rank correlation coefficient r_s is computed by using rank scores R_i for X_i and C_i for Y_j. These rank scores are defined as follows:

$$R_i = \sum_{k<i} r_k + (r_i + 1)\big/2 \qquad \text{for } i = 1, 2, \ldots, R$$

$$C_j = \sum_{h<j} c_h + (c_j + 1)\big/2 \qquad \text{for } j = 1, 2, \ldots, C$$

The formulas for r_s and its asymptotic variance can be obtained from the Pearson formulas by substituting R_i and C_j for X_i and Y_j, respectively.

Eta

Asymmetric η with the column variable *Y* as dependent is

$$\eta_Y = \sqrt{1 - \frac{S_{YW}}{S(Y)}}$$

where

$$S_{YW} = \sum_{i,j} Y_j^2 f_{ij} - \sum_{i=1}^{R} \frac{1}{r_i} \left(\sum_{j=1}^{C} Y_j f_{ij} \right)^2$$

Relative Risk

Consider a 2×2 table (that is, $R = C = 2$). In a case-control study, the relative risk is estimated as

$$R_0 = \frac{f_{11} f_{22}}{f_{12} f_{21}}$$

The $100(1-\alpha)$ percent *CI* for the relative risk is obtained as

$$\left[R_0 \exp\left(-z_{1-\alpha/2} v\right), \quad R_0 \exp\left(z_{1-\alpha/2} v\right) \right]$$

where

$$v = \left(\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}} \right)^{1/2}$$

The relative risk ratios in a cohort study are computed for both columns. For column 1, the risk is

$$R_1 = \frac{f_{11}(f_{21} + f_{22})}{f_{21}(f_{11} + f_{12})}$$

and the corresponding $100(1-\alpha)$ percent *CI* is

$$\left[R_1 \exp\left(-z_{1-\alpha/2} v\right), \quad R_1 \exp\left(z_{1-\alpha/2} v\right) \right]$$

where

$$v=\left(\frac{f_{12}}{f_{11}(f_{11}+f_{12})}+\frac{f_{22}}{f_{21}(f_{21}+f_{22})}\right)^{1/2}$$

The relative risk for column 2 and the confidence interval are computed similarly.

McNemar-Bowker's Test

This statistic is used to test if a square table is symmetric.

Notations

n	Dimension of the table (both row and column)
p_{ij}	Unknown population cell probability of row i and column j
n_{ij}	Observed counts cell count of row i and column j

Algorithm

Given a $n \times n$ square table, the McNemar-Bowker's statistic is used to test the hypothesis $H_0 : p_{ij} = p_{ji}$ for all (i<j) v.s. $H_1 : p_{ij} \neq p_{ji}$ for at least one pair of (i,j). The statistic is defined by the formula

$$\chi^2 = \sum_{i<j} \frac{I(n_{ij} + n_{ji} > 0)(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

Where *I()* is the indicator function. Under the null hypothesis, χ^2 has an asymptotic Chi-square distribution with $n(n-1)/2$ degrees of freedom. The null hypothesis will be rejected if χ^2 has a large value. The two-sided p-value is equal to $1-F(n(n-1)/2,\chi^2)$, where $F(df,\cdot)$ is the CDF of Chi-square distribution with df degrees of freedom.

A Special Case: 2x2 Tables

For 2x2 table, the statistic reduces to the classical McNemar (1947) statistic for which exact p-value can be computed. The two-tailed probability level is

$$2\sum_{i=0}^{\min(n_{12},n_{21})}\binom{n_{12}+n_{21}}{i}(1/2)^{n_{12}+n_{21}}$$

Conditional Independence and Homogeneity

The Cochran's and Mantel-Haenzel statistics test the independence of two dichotomous variables, controlling for one or more other categorical variables. These "other" categorical variables define a number of strata, across which these statistics are computed.

The Breslow-Day statistic is used to test homogeneity of the common odds ratio, which is a weaker condition than the conditional independence (i.e., homogeneity with the common odds ratio of 1) tested by Cochran's and Mantel-Haenszel statistics. Tarone's statistic is the Breslow-Day statistic adjusted for the consistent but inefficient estimator such as the Mantel-Haenszel estimator of the common odds ratio.

Notation and Definitions

The addition of strata requires the following modifications to the notation:

K	The number of strata.
f_{ijk}	Sum of cell weights for cases in the *i*th row of the *j*th column of the *k*th strata.
c_{jk}	$\sum_{i=1}^{R} f_{ijk}$, the *j*th column of the *k*th strata subtotal.
r_{ik}	$\sum_{j=1}^{C} f_{ijk}$, the *i*th row of the kth strata subtotal.
n_k	$\sum_{j=1}^{C} c_{jk} = \sum_{i=1}^{R} r_{ik}$, the grand total of the *k*th strata.

E_{ijk} $\quad$ $E\left(f_{ijk}\right) = \frac{r_{ik} c_{jk}}{n_k}$, the expected cell count of the *i*th row of the *j*th column of the *k*th strata.

A stratum such that $n_k = 0$ is omitted from the analysis. (K must be modified accordingly.) If $n_k = 0$ for all k, then no computation is done.

Preliminarily, define for each k

$$\hat{p}_{ik} = \frac{f_{i1k}}{r_{ik}},$$

$$d_k = \hat{p}_{1k} - \hat{p}_{2k},$$

$$\hat{p}_k = \frac{c_{1k}}{n_k},$$

and

$$w_k = \frac{r_{1k} r_{2k}}{n_k}.$$

Cochran's Statistic

Cochran's (1954) statistic is

$$C = \frac{\sum_{k=1}^{K} w_k d_k \Big/ \sum_{k=1}^{K} w_k}{\sqrt{\sum_{k=1}^{K} w_k \hat{p}_k (1 - \hat{p}_k)} \Big/ \sum_{k=1}^{K} w_k} = \frac{\sum_{k=1}^{K} w_k d_k}{\sqrt{\sum_{k=1}^{K} w_k \hat{p}_k (1 - \hat{p}_k)}}.$$

All stratum such that $r_{1k} = 0$ or $r_{2k} = 0$ are excluded, because d_k is undefined. If every stratum is such, C is undefined. Note that a stratum such that $r_{1k} > 0$ and $r_{2k} > 0$ but that $c_{1k} = 0$ or $c_{2k} = 0$ is a valid stratum, although it contributes nothing to the denominator or numerator. However, if every stratum is such, C is again undefined. So, in order to compute a non system missing value of C, at least one stratum must have all non-zero marginal totals.

Alternatively, Cochran's statistic can be written as

$$C = \frac{\sum_{k=1}^{K} (f_{11k} - E_{11k})}{\sqrt{\sum_{k=1}^{K} w_k \hat{p}_k (1 - \hat{p}_k)}}.$$

When the number of strata is fixed as the sample sizes within each stratum increase, Cochran's statistic is asymptotically standard normal, and thus its square is asymptotically distributed as a chi-squared distribution with 1 d.f.

Mantel and Haeszel's Statistic

Mantel and Haenszel's (1959) statistic is simply Cochran's statistic with small-sample corrections for continuity and variance "inflation". These corrections are desirable when r_{1k} and r_{2k} are small, but the corrections can make a noticeable difference even for relatively large r_{1k} and r_{2k} (Snedecor and Cochran, 1980, p. 213). The statistic is defined as:

$$M = \frac{\{|\sum_{k=1}^{K} (f_{11k} - E_{11k})| - 0.5\} \operatorname{sgn}\{\sum_{k=1}^{K} (f_{11k} - E_{11k})\}}{\sqrt{\sum_{k=1}^{K} \frac{r_{1k} r_{2k}}{n_k - 1} \hat{p}_k (1 - \hat{p}_k)}},$$

where sgn is the signum function

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

Any stratum in which $n_k = 1$ is excluded from the computation. If every stratum is such, then *M* is undefined. *M* is also undefined if every stratum is such that $r_{1k} = 0$, $r_{2k} = 0$, $c_{1k} = 0$, or $c_{2k} = 0$. In order to compute a non system missing value of *M*, at least one stratum must have all non-zero marginal totals, just as for *C*.

When the number of strata is fixed as the sample sizes within each stratum increase, or when the sample sizes within each strata are fixed as the number of strata increases, this statistic is asymptotically standard normal, and thus its square is asymptotically distributed as a chi-squared distribution with 1 d.f.

The Breslow-Day Statistic

The Breslow-Day statistic for any estimator $\hat{\theta}$ is

$$\sum_{k=1}^{K} \frac{\{f_{11k} - \text{E}(f_{11k} | c_{1k}; \hat{\theta})\}^2}{\text{V}(f_{11k} | c_{1k}; \hat{\theta})}.$$

E and V are based on the exact moments, but it is customary to replace them with the asymptotic expectation and variance. Let $\hat{\text{E}}$ and $\hat{\text{V}}$ mean the estimated asymptotic expectation and the estimated asymptotic variance, respectively. Given the Mantel-Haenszel common odds ratio estimator $\hat{\theta}_{\text{MH}}$, we use the following statistic as the Breslow-Day statistic:

$$B = \sum_{k=1}^{K} \frac{\{f_{11k} - \hat{\text{E}}(f_{11k} | c_{1k}; \hat{\theta}_{\text{MH}})\}^2}{\hat{\text{V}}(f_{11k} | c_{1k}; \hat{\theta}_{\text{MH}})},$$

where

$$\hat{\text{E}}(f_{11k} | c_{1k}; \hat{\theta}_{\text{MH}}) = \hat{f}_{11k}$$

satisfies the equations

$$\frac{\hat{f}_{11k}(n_k - r_{1k} - c_{1k} + \hat{f}_{11k})}{(r_{1k} - \hat{f}_{11k})(c_{1k} - \hat{f}_{11k})} = \hat{\theta}_{\text{MH}},$$

with constraints such that

$$\hat{f}_{11k} \geq 0,$$
$$r_{1k} - \hat{f}_{11k} > 0,$$
$$c_{1k} - \hat{f}_{11k} > 0,$$
$$n_k - r_{1k} - c_{1k} + \hat{f}_{11k} \geq 0;$$

and

$$\hat{\text{V}}(f_{11k} | c_{1k}; \hat{\theta}_{\text{MH}}) = \left(\frac{1}{\hat{f}_{11k}} + \frac{1}{\hat{f}_{12k}} + \frac{1}{\hat{f}_{21k}} + \frac{1}{\hat{f}_{22k}} \right)^{-1}$$

with constraints such that

$$\hat{f}_{11k} > 0,$$
$$\hat{f}_{12k} = r_{1k} - \hat{f}_{11k} > 0,$$
$$\hat{f}_{21k} = c_{1k} - \hat{f}_{11k} > 0,$$
$$\hat{f}_{22k} = n_k - r_{1k} - c_{1k} + \hat{f}_{11k} > 0;$$

All stratum such that $r_{1k} = 0$ or $c_{1k} = 0$ are excluded. If every stratum is such, B is undefined. Stratum such that $\hat{f}_{11k} = 0$ are also excluded. If every stratum is such, then B is undefined.

Breslow-Day's statistic is asymptotically distributed as a chi-squared random variable with K-1 degrees of freedom under the null hypothesis of a constant odds ratio.

Tarone's Statistic

Tarone (1985) proposes an adjustment to the Breslow-Day statistic when the common odds ratio estimator is consistent but inefficient, specifically when we have the Mantel-Haenszel common odds ratio estimator. The adjusted statistic, Tarone's statistic, for $\hat{\theta}_{\text{MH}}$ is

$$T = \sum_{k=1}^{K} \frac{\{f_{11k} - \hat{\text{E}}(f_{11k}|c_{1k};\hat{\theta}_{\text{MH}})\}^2}{\hat{\text{V}}(f_{11k}|c_{1k};\hat{\theta}_{\text{MH}})} - \frac{\left[\sum_{k=1}^{K}\{f_{11k} - \hat{\text{E}}(f_{11k}|c_{1k};\hat{\theta}_{\text{MH}})\}\right]^2}{\sum_{k=1}^{K}\hat{\text{V}}(f_{11k}|c_{1k};\hat{\theta}_{\text{MH}})}$$

$$= B - \frac{\left[\sum_{k=1}^{K}\{f_{11k} - \hat{\text{E}}(f_{11k}|c_{1k};\hat{\theta}_{\text{MH}})\}\right]^2}{\sum_{k=1}^{K}\hat{\text{V}}(f_{11k}|c_{1k};\hat{\theta}_{\text{MH}})},$$

where $\hat{\text{E}}$ and $\hat{\text{V}}$ are as before.

The required data conditions are the same as for the Breslow-Day statistic computation. T is, of course, undefined, when B is undefined.

T is also asymptotically distributed as a chi-squared random variable with K-1 degrees of freedom under the null hypothesis of a constant odds ratio.

Estimation of the Common Odds Ratio

For K strata of 2×2 tables, write the true odds ratios as

$$\theta_k = \frac{p_{1k}(1-p_{2k})}{(1-p_{1k})p_{2k}}$$

for $k = 1, \ldots, K$. And, assuming that the true common odds ratio exists, $\theta = \theta_1 = \ldots = \theta_K$, Mantel and Haenszel's (1959) estimator of this common odds ratio is

$$\hat{\theta}_{\mathrm{MH}} = \frac{\sum_{k=1}^{K} \frac{f_{11k} f_{22k}}{n_k}}{\sum_{k=1}^{K} \frac{f_{12k} f_{21k}}{n_k}}.$$

If every stratum is such that $f_{12k} = 0$ or $f_{21k} = 0$, then $\hat{\theta}_{\mathrm{MH}}$ is undefined.

The (natural) log of the estimated common odds ratio is asymptotically normal. Note, however, that if $f_{11k} = 0$ or $f_{22k} = 0$ in every stratum, then $\hat{\theta}_{\mathrm{MH}}$ is zero and $\log\left(\hat{\theta}_{\mathrm{MH}}\right)$ is undefined.

The Asymptotic Confidence Interval

Robins et al. (1986) give an estimated asymptotic variance for $\log\left(\hat{\theta}_{\mathrm{MH}}\right)$ that is appropriate in both asymptotic cases:

$$\hat{\sigma}^2[\log(\hat{\theta}_{\mathrm{MH}})] = \frac{\sum_{k=1}^{K} \frac{(f_{11k} + f_{22k}) f_{11k} f_{22k}}{n_k^2}}{2(\sum_{k=1}^{K} \frac{f_{11k} f_{22k}}{n_k})^2} + \frac{\sum_{k=1}^{K} \frac{(f_{11k} + f_{22k}) f_{12k} f_{21k} + (f_{12k} + f_{21k}) f_{11k} f_{22k}}{n_k^2}}{2(\sum_{k=1}^{K} \frac{f_{11k} f_{22k}}{n_k})(\sum_{k=1}^{K} \frac{f_{12k} f_{21k}}{n_k})} + \frac{\sum_{k=1}^{K} \frac{(f_{12k} + f_{21k}) f_{12k} f_{21k}}{n_k^2}}{2(\sum_{k=1}^{K} \frac{f_{12k} f_{21k}}{n_k})^2}.$$

An asymptotic (100 - α)% confidence interval for $\log(\theta)$ is

$$\log(\hat{\theta}_{\mathrm{MH}}) \pm z(\alpha / 2)\hat{\sigma}[\log(\hat{\theta}_{\mathrm{MH}})],$$

where $z(\alpha / 2)$ is the upper $\alpha / 2$ critical value for the standard normal distribution.

All these computations are valid only if $\hat{\theta}_{MH}$ is defined and greater than 0.

The Asymptotic *P*-value

We compute an asymptotic *P*-value under the null hypothesis that $\theta \, (= \theta_k \;\forall k) \; = \theta_o \; (> 0)$ against a 2-sided alternative hypothesis $(\theta \neq \theta_o)$, using the standard normal variate, as follows

$$\Pr\left(|Z| > \left| \frac{\log(\hat{\theta}_{MH}) - \log(\theta_o)}{\hat{\sigma}[\log(\hat{\theta}_{MH})]} \right| \right) = 2\Pr\left(Z > \left| \frac{\log(\hat{\theta}_{MH}) - \log(\theta_o)}{\hat{\sigma}[\log(\hat{\theta}_{MH})]} \right| \right),$$

given that $\log\left(\hat{\theta}_{MH}\right)$ is defined.

Alternatively, we can consider using $\hat{\theta}_{MH}$ and the estimated exact variance of $\hat{\theta}_{MH}$, which is still consistent in both limiting cases:

$$\hat{\sigma}^2[\log(\hat{\theta}_{MH})]\hat{\theta}^2_{MH}.$$

Then, the asymptotic *P*-value may be approximated by

$$\Pr\left(|Z| > \left| \frac{\hat{\theta}_{MH} - \theta_o}{\hat{\sigma}[\log(\hat{\theta}_{MH})]\theta_o} \right| \right).$$

The caveat for this formula is that $\hat{\theta}_{MH}$ may be quite skewed even in moderate sample sizes (Robins et al., 1986, p. 314).

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley, New York.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley, New York.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. 1975. *Discrete multivariate analysis*: *Theory and practice*. Cambridge, Mass.: MIT Press.

Bowker, A.H. 1948. A test for symmetry in contingency tables. *Journal of the American Statistical Association*, V43, 572-574.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*, **1**, *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

Brown, M. B. 1975. The asymptotic standard errors of some estimates of uncertainty in the two-way contingency table. *Psychometrika*, **40**(3): 291.

Brown, M. B., and Benedetti, J. K. 1977. Sampling behavior of tests for correlation in two-way contingency tables. *Journal of the American Statistical Association*, 72: 309–315.

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.

Goodman, L. A., and Kruskal, W. H. 1954. Measures of association for cross-classification. *Journal of the American Statistical Association*, 49: 732–764.

Goodman, L. A., and Kruskal, W. H. 1972. Measures of association for cross-classification, IV: simplification and asymptotic variances, *Journal of the American Statistical Association*, 67: 415–421.

Hauck, W. (1989). Odds ratio inference from stratified samples. *Commun. Statist.-Theory Meth.*, **18**, 767-800.

Somes, G. W. and O'Brien, K. F. (1985). Mantel-Haenszel statistic. In *Encyclopedia of Statistical Sciences, Vol. 5* (S. Kotz and N. L. Johnson, eds.) 214-217. John Wiley, New York.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719-748.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, V12, 153-157.

Robins, J., Breslow, N., and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**, 311-323.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*, 7th ed. Iowa State University Press, Ames, Iowa.

Tarone, R. E. (1985). On heterogeneity tests based on efficient scores. *Biometrika*, **72**, 91-95.