CSTABULATE

This document describes the algorithms used in the complex sampling estimation procedure CSTABULATE.

Complex sample data must contain both the values of the variables to be analyzed and the information on the current sampling design. The sampling design includes the sampling method, strata and clustering information, inclusion probabilities and the overall sampling weights.

The sampling design specification for CSTABULATE may include up to three stages of sampling. Any of the following general sampling methods may be assumed in the first stage: random sampling with replacement, random sampling without replacement and equal probabilities and random sampling without replacement and unequal probabilities. The first two sampling methods can also be specified for the second and the third sampling stage.

Notation

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

H	Number of strata.
n_h	Sampled number of primary sampling units (PSU) per stratum.
f_h	Sampling rate per stratum.
m_{hi}	Number of elements in the i^{th} sampled unit in stratum h, $i = 1, \ldots, n_h$.
y_{hij}	Value of variable y for the j^{th} element in the i^{th} sampled unit in stratum h.
w_{hij}	Overall sampling weight for the j^{th} element in the i^{th} sampled unit in stratum h.
n	Total number of elements in the sample.
N	Total number of elements in the population.
Y	Population total sum for variable y.

Weights

Overall weights specified for each ultimate element are processed as given. They can be obtained as a product of weights for corresponding units computed in each sampling stage.

When sampling without replacement in a given stage, substituting $w_{hi} = 1/\pi_{hi}$ for unit i in stratum h results in the application of the estimator for the population totals due to Horvitz and Thompson (1952). The corresponding variance estimator (2) or (3) will also be unbiased. π_{hi} is the probability of unit i from stratum h being selected in the given stage.

If sampling with replacement in a given stage, substituting $w_{hi} = 1/(n_h p_{hi})$ yields the estimator for the population totals due to Hansen and Hurwitz (1943). Repeatedly selected units should be replicated in the data. The corresponding variance estimator (1) will be unbiased. p_{hi} is the probability of selecting unit i in a single draw from stratum h in the given stage.

Weights obtained in each sampling stage need to be multiplied when processing multi-stage samples. The resulting overall weights for the elements in the final stage are used in all expressions and formulas below.

Z expressions

$$z_{hij} = w_{hij} y_{hij} \qquad z'_{hij} = w_{hij} y'_{hij}$$

$$z_{hi} = \sum_{j=1}^{m_{hi}} z_{hij} \qquad z'_{hi} = \sum_{j=1}^{m_{hi}} z'_{hij}$$

$$\bar{z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi} \qquad \bar{z}'_h = \frac{1}{n_h} \sum_{i=1}^{n_h} z'_{hi}$$

$$S_h^2(y, y') = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)(z'_{hi} - \bar{z}'_h)$$

For multi-stage samples, the index h denotes a stratum in the given stage, and i stands for unit from h in the same stage. The index j runs over all final stage elements contained in unit hi.

Variable Total

An estimate for the population total of variable y in a single-stage sample is the weighted sum over all the strata and all the clusters:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

Alternatively, we compute the weighted sum over all the elements in the sample:

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i$$

The latter expression is more general as it also applies to multi-stage samples.

Variables Total Covariance

For a multi-stage sample containing a with replacement sampling stage, all specifications other than weights are ignored for the subsequent stages. They make no contribution to the variance estimates.

Single stage sample

The covariance of the total for variables y and y' in a single-stage sample is estimated by the following:

$$\hat{C}(\hat{Y}, \hat{Y}') = \hat{C}_1(\hat{Y}, \hat{Y}') = \sum_{h=1}^{H} U_h(\hat{Y}, \hat{Y}')$$

where $U_h(\hat{Y}, \hat{Y}')$ is an estimate contribution from stratum $h = 1, \ldots, H$ and depends on the sampling method as follows:

- For sampling with replacement

$$U_h(\hat{Y}, \hat{Y}') = n_h S_h^2(y, y') \quad \textbf{(1)}$$

- For simple random sampling

$$U_h(\hat{Y}, \hat{Y}') = (1 - f_h) n_h S_h^2(y, y') \quad \textbf{(2)}$$

- For sampling without replacement and unequal probabilities

$$U_h(\hat{Y}, \hat{Y}') = \sum_{i=1}^{n_h} \sum_{i>j}^{n_h} \left(\frac{\pi_{hi}\pi_{hj}}{\pi_{hij}} - 1\right)(z_{hi} - z_{hj})(z'_{hi} - z'_{hj}) \quad \textbf{(3)}$$

In the variance estimator (3), π_{hi} and π_{hj} are the inclusion probabilities for units i and j in stratum h, and π_{hij} is the joint inclusion probability for the same units. This estimator is due to Yates and Grundy (1953) and Sen (1953).

For each stratum h containing a single element, the covariance contribution $U_h(\hat{Y},\hat{Y}')$ is always set to zero.

Two-stage sample

When the sample is obtained in two stages and sampling without replacement is applied in the first stage, we use the following estimate for the covariance of the total for variables y and y':

$$\hat{C}(\hat{Y},\hat{Y}') = \hat{C}_2(\hat{Y},\hat{Y}') = \hat{C}_1(\hat{Y},\hat{Y}') + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}\sum_{k=1}^{K_{hi}}U_{hik}(\hat{Y},\hat{Y}').$$

where

- π_{hi} is the first stage inclusion probability for the primary sampling unit i in stratum h. In the case of simple random sampling, the inclusion probability is equal to the sampling rate f_h for stratum *h*.
- K_{hi} is the number of second stage strata in the primary sampling unit i within the first stage stratum *h*.
- $U_{hik}(\hat{Y},\hat{Y}')$ is a covariance contribution from the second stage stratum k from the primary sampling unit hi. It depends on the second stage sampling method. The corresponding formula (1) or (2) applies.

Three-stage sample

When the sample is obtained in three stages where sampling in the first stage is done without replacement and simple random sampling is applied in the second stage, we use the following estimate for the covariance of the total for variables y and y':

$$\hat{C}(\hat{Y},\hat{Y}') = \hat{C}_2(\hat{Y},\hat{Y}') + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}\sum_{k=1}^{K_{hi}}f_{hik}\sum_{j=1}^{n_{hik}}\sum_{l=1}^{L_{hikj}}U_{hikjl}(\hat{Y},\hat{Y}')$$

where

- f_{hik} is the sampling rate for the secondary sampling units in the second stage stratum hik.
- L_{hikj} is the number of the third stage strata in the secondary sampling unit $hikj$.
- $U_{hikjl}(\hat{Y},\hat{Y}')$ is a covariance contribution from the third stage stratum l contained in the secondary sampling unit $hikj$. It depends on the third stage sampling method. The corresponding formula (1) or (2) applies.

Variable total variance

The variance of the total for variable y in a complex sample is estimated by

$$\hat{V}(\hat{Y}) = \hat{C}(\hat{Y}, \hat{Y})$$

with $\hat{C}(\hat{Y}, \hat{Y})$ defined above.

Population Size Estimation

An estimate for the population size corresponds to the estimate for the variable total; it is sum of the sampling weights. We have the following estimate for the single-stage samples:

$$\hat{N} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} .$$

More generally,

$$\hat{N} = \sum_{i=1}^{n} w_i .$$

The variance of $\hat{N}$ is obtained by replacing y_{hij} with 1; that is, by replacing z_{hij} with w_{hij} in the corresponding variance estimator formula for $\hat{V}(\hat{Y})$.

Cell Estimates: One-Way Tables

Let the population be classified according to the values of single categorical row variable and possibly one or more categorical variables in the layer. Categories for the row variable are enumerated by $r = 1, \ldots, R$ and categories for the layer variables are given by $l = 1, \ldots, L$. Each combination of the values (r, l) defines a domain and a cell in the one-way table $(r, l), r = 1, \ldots, R$. For each cell (r, l) we define a corresponding indicator variable:

$$\delta_{hij}(r, l) = \begin{cases} 1 & \text{if the sample unit } hij \text{ is in the cell } (r,l) \\ 0 & \text{otherwise} \end{cases}$$

Sizes

To estimate a cell population size or a table population size, we replace y_i with $\delta_i(r, l)$ in the formula for the population total and obtain the following expressions:

- Cell population size

$$\hat{N}(r,l) = \sum_{i=1}^{n} w_i \delta_i(r,l)$$

- Table population size

$$\hat{N}(+,l) = \sum_{i=1}^{n} \sum_{r=1}^{R} w_i \delta_i(r,l).$$

Similarly, in order to estimate variances of the above estimators, we substitute y_{hij} with $\delta_{hij}(r,l)$ in the corresponding formula for the whole population. The following substitutions of z_{hij} in the formulas for $\hat{V}(\hat{Y})$ are used for estimating the variances of these estimators:

- Cell population size

$$z_{hij}(r,l) = w_{hij} \delta_{hij}(r,l)$$

- Table population size

$$z_{hij}(+,l) = \sum_{r=1}^{R} w_{hij} \delta_{hij}(r,l)$$

Proportions

A table proportion estimate is computed at each layer category as follows:

$$\hat{P}_{tab}(r,l) = \hat{N}(r,l) / \hat{N}(+,l)$$

This estimator is a ratio and we apply Taylor linearization formulas as suggested by Woodruff (1971). The following substitution of z_{hij} in the formulas for $\hat{V}(\hat{Y})$ are used for estimating the variance of the table proportion at a given layer:

$$z_{hij}(r,l) = w_{hij} \frac{\delta_{hij}(r,l) - \delta_{hij}(+,l)\hat{P}_{tab}(r,l)}{\hat{N}(+,l)}$$

Cell Estimates: Two-Way Tables

Let the population be cross-classified according to the values of a categorical row variable, a categorical column variable and possibly one or more categorical variables in the layer. Categories for the row variable are enumerated by $r = 1, \ldots, R$, while categories for the column variable are denoted by $c = 1, \ldots, C$ and categories for the layer variables are given by $l = 1, \ldots, L$. Each combination of values (r, c, l) defines a domain and a cell in the two-way table (r, c, l), where $r = 1, \ldots, R$ and $c = 1, \ldots, C$. For each cell (r, c, l) we define a corresponding indicator variable:

$$\delta_{hij}(r,c,l) = \begin{cases} 1 & \text{if the sample unit } hij \text{ is in the cell } (r,c,l) \\ 0 & \text{otherwise} \end{cases}$$

We will also use the following indicator notation:

- Row indicator

$$\delta_i(r,+,l) = \sum_{c=1}^{C} \delta_i(r,c,l)$$

- Column indicator

$$\delta_i(+,c,l) = \sum_{r=1}^{R} \delta_i(r,c,l)$$

- Table indicator

$$\delta_i(+,+,l) = \sum_{r=1}^{R} \sum_{c=1}^{C} \delta_i(r,c,l)$$

Sizes

To estimate various domain sizes, we substitute y_i with δ_i in the corresponding formula for the whole population as follows:

- Cell population size

$$\hat{N}(r,c,l) = \sum_{i=1}^{n} w_i \delta_i(r,c,l)$$

- Row population size

$$\hat{N}(r,+,l) = \sum_{i=1}^{n} w_i \delta_i (r,+,l)$$

- Column population size

$$\hat{N}(+,c,l) = \sum_{i=1}^{n} w_i \delta_i (+,c,l)$$

- Table population size

$$\hat{N}(+,+,l) = \sum_{i=1}^{n} w_i \delta_i (+,+,l).$$

Similarly, in order to estimate variance of the above estimators, we substitute y_{hij} with δ_{hij} in the corresponding formula for the whole population. The following substitutions of z_{hij} in the formulas for $\hat{V}(\hat{Y})$ are used for estimating variances of:

- Cell population size

$$z_{hij}(r,c,l) = w_{hij}\delta_{hij}(r,c,l)$$

- Row population size

$$z_{hij}(r,+,l) = w_{hij}\delta_{hij}(r,+,l)$$

- Column population size

$$z_{hij}(+,c,l) = w_{hij}\delta_{hij}(+,c,l)$$

- Table population size

$$z_{hij}(+,+,l) = w_{hij}\delta_{hij}(+,+,l)$$

Proportions

We define various proportion estimates to be computed as follows:

- Row population proportion

$$\hat{P}_{row}(r,c,l) = \hat{N}(r,c,l) \big/ \hat{N}(r,+,l)$$

- Column population proportion

$$\hat{P}_{col}(r,c,l) = \hat{N}(r,c,l) \big/ \hat{N}(+,c,l)$$

- Table population proportion

$$\hat{P}_{tab}(r,c,l) = \hat{N}(r,c,l) \big/ \hat{N}(+,+,l)$$

- Marginal column population proportion

$$\hat{P}_{mcol}(+,c,l) = \hat{N}(+,c,l) \big/ \hat{N}(+,+,l)$$

- Marginal row population proportion

$$\hat{P}_{mrow}(r,+,l) = \hat{N}(r,+,l) \big/ \hat{N}(+,+,l)$$

In order to estimate variances of the above estimators, again apply the Taylor linearization formulas as for the one-way tables. The following substitutions of z_{ij} in the formulas for $\hat{V}(\hat{Y})$ are used for estimating variances of:

- Row population proportion

$$z_{hij}(r,c,l) = w_{hij} \frac{\delta_{hij}(r,c,l) - \delta_{hij}(r,+,l)\hat{P}_{row}(r,c,l)}{\hat{N}(r,+,l)}$$

- Column population proportion

$$z_{hij}(r,c,l) = w_{hij} \frac{\delta_{hij}(r,c,l) - \delta_{hij}(+,c,l)\hat{P}_{col}(r,c,l)}{\hat{N}(+,c,l)}$$

- Table population proportion

$$z_{hij}(r,c,l) = w_{hij}\,\frac{\delta_{hij}(r,c,l) - \delta_{hij}(+,+,l)\hat{P}_{tab}(r,c,l)}{\hat{N}(+,+,l)}$$

- Marginal column population proportion

$$z_{hij}(+,c,l) = w_{hij}\,\frac{\delta_{hij}(+,c,l) - \delta_{hij}(+,+,l)\hat{P}_{mcol}(+,c,l)}{\hat{N}(+,+,l)}$$

- Marginal row population proportion

$$z_{hij}(r,+,l) = w_{hij}\,\frac{\delta_{hij}(r,+,l) - \delta_{hij}(+,+,l)\hat{P}_{mrow}(r,+,l)}{\hat{N}(+,+,l)}$$

Standard Errors

Let Z denote any of the domain quantities defined above: cell population sizes or proportions. Then the standard error of an estimator $\hat{Z}$ is the square root of its estimated variance:

$$SE(\hat{Z}) = \sqrt{\hat{V}(\hat{Z})}\,.$$

Coefficient of variation

The coefficient of variation of the estimator $\hat{Z}$ is the ratio of its standard error and its value:

$$CV(\hat{Z}) = \frac{SE(\hat{Z})}{\hat{Z}}\,.$$

The coefficient of variation is undefined when $\hat{Z} = 0$.

Confidence Limits

Sizes

A level $1-\alpha$ confidence interval is constructed for a given $0 \le \alpha \le 1$ for any domain size N_d defined earlier. The confidence bounds are then given by

$$\hat{N}_d \pm SE(\hat{N}_d) t_\nu (1-\alpha/2)$$

where $SE(\hat{N}_d)$ is the estimated standard error of $\hat{N}_d$, and $t_\nu(1-\alpha/2)$ is the $100(1-\alpha/2)$ percentile of the t distribution with ν degrees of freedom.

Proportions

For any domain proportion P_d, we use the logistic transformation $f(p) = \ln(p/(1-p))$ and obtain the following $1-\alpha$ level confidence bounds for the transformed estimate:

$$\ln\left(\frac{\hat{P}_d}{1-\hat{P}_d}\right) \pm \frac{SE(\hat{P}_d)}{\hat{P}_d(1-\hat{P}_d)} t_\nu(1-\alpha/2).$$

These bounds are transformed back to the original metric using the logistic inverse $f^{-1}(y) = \exp(y)/(1+\exp(y))$.

Degrees of freedom

The degrees of freedom ν for the t distributions above is calculated as the difference between the number of primary sampling units and the number of strata in the first stage of sampling. We shall also refer to this quantity as the sample design degrees of freedom.

Design Effects

Sizes

The design effect $Deff$ for a two-way table cell population size is estimated by

$$Deff = \frac{\hat{V}(\hat{N}(r,c,l))}{\hat{V}_{srs}(\hat{N}(r,c,l))}.$$

$\hat{V}(\hat{N}(r,c,l))$ is an estimate of the variance of $\hat{N}(r,c,l)$ under the complex sample design, while $\hat{V}_{srs}(\hat{N}(r,c,l))$ is its estimate of variance under the simple random sampling assumption as follows:

$$\hat{V}_{srs}(\hat{N}(r,c,l)) = (1-\frac{n}{\hat{N}})\frac{1}{n-1}\hat{N}(r,c,l)(\hat{N}-\hat{N}(r,c,l)).$$

Computations of the design effects for the one-way table cells, as well as for the row, column and table population sizes are analogous to the one above.

Proportions

$Deff$ for a two-way table population proportion is estimated by

$$Deff = \frac{\hat{V}(\hat{P}_{tab}(r,c,l))}{\hat{V}_{srs}(\hat{P}_{tab}(r,c,l))}.$$

$\hat{V}(\hat{P}_{tab}(r,c,l))$ is an estimate of the variance of $\hat{P}_{tab}(r,c,l)$ under the complex sample design, while $\hat{V}_{srs}(\hat{P}_{tab}(r,c,l))$ is its estimate of variance under the simple random sampling assumption:

$$\hat{V}_{srs}(\hat{P}_{tab}(r,c,l)) = (1-\frac{n}{\hat{N}})\frac{\hat{N}}{n-1}\frac{\hat{P}_{tab}(r,c,l)(1-\hat{P}_{tab}(r,c,l))}{\hat{N}(+,+,l)}.$$

Computations of the design effects for one-way table proportions, as well as for the row, column, marginal row and marginal column population proportions are analogous to the one above.

Design effects for various estimates are computed only when the condition $\frac{n}{\hat{N}} < 1$ is satisfied.

Design effect square root

We also compute the square root of a design effect $\sqrt{Deff}$.

Design effects and their applications have been discussed by Kish (1965) and Kish (1995).

Tests of Independence for Two-Way Tables

Let the population be cross-classified according to the values of a categorical row variable, a categorical column variable and possibly one a more categorical variables in the layer. Categories for the row variable are enumerated by $r = 1, \ldots, R$, while categories for the column variable are denoted by $c = 1, \ldots, C$. When the layer variables are given we assume that their categories coincide with the strata in the first sampling stage. In the following we omit reference to the layers as the formulas apply for each stratum separately when needed.

We use a contrast matrix $\mathbf{C}$ defined as follows. Let $\mathbf{A}_R$ be the contrast matrix given by

$$\mathbf{A}_R = [\mathbf{I}_{R-1} \mid -\mathbf{1}_{R-1}]'.$$

$\mathbf{I}_{R-1}$ is an identity matrix of size $R-1$ and $\mathbf{1}_{R-1}$ is a vector with $R-1$ elements equal to 1.

Define $\mathbf{C}$ to be a $RC \times (R-1)(C-1)$ matrix defined by the following Kronecker product:

$$\mathbf{C} = \mathbf{A}_R \otimes \mathbf{A}_C.$$

Adjusted Pearson statistic test of independence

We provide an adjusted Pearson statistic test. The Pearson statistic is computed according to the following standard formula:

$$X^2 = n \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(\hat{P}(r,c) - \hat{P}(r,+)\hat{P}(+,c))^2}{\hat{P}(r,+)\hat{P}(+,c)}$$

Since under the null hypothesis, the asymptotic distribution of X^2 is generally not a chi-square distribution, we perform an adjustment using the following $\hat{\mathbf{\Delta}}$ matrix:

$$\hat{\mathbf{\Delta}} = n(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{P}}}^{-1}\hat{\mathbf{M}}\mathbf{D}_{\hat{\mathbf{P}}}^{-1}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{P}}}^{-1}\hat{\mathbf{V}}(\hat{\mathbf{P}})\mathbf{D}_{\hat{\mathbf{P}}}^{-1}\mathbf{C}).$$

$\hat{\mathbf{P}}$ is a vector and $\mathbf{D}_{\hat{\mathbf{P}}}$ is a diagonal matrix of size RC containing elements $\hat{P}(r,c)$. $\hat{\mathbf{M}} = [\hat{\mathbf{D}}_{\hat{\mathbf{P}}} - \hat{\mathbf{P}}\hat{\mathbf{P}}']$ is a multinomial covariance matrix estimating the asymptotic covariance of $\hat{\mathbf{P}}$ under the simple random sampling design, while $\hat{\mathbf{V}}(\hat{\mathbf{P}})$ estimates covariance matrix of $\hat{\mathbf{P}}$ under the complex sampling design.

We use the F-based variant of the Rao and Scott's (1984) second-order adjustment

$$FX^2 = \frac{X^2}{tr\hat{\mathbf{\Delta}}}$$

with

$$d = \frac{(tr\hat{\mathbf{\Delta}})^2}{tr\hat{\mathbf{\Delta}}^2}.$$

The asymptotic distribution of FX^2 is approximated by the $F(d, d\nu)$ distribution where ν is the number of the sample design degrees of freedom.

Properties of this test are given in a review of simulation studies by Rao and Thomas (2003).

Adjusted likelihood ratio test of independence

The likelihood ratio test statistic is given by

$$G^2 = 2n\sum_{r=1}^{R}\sum_{c=1}^{C}\hat{P}(r,c)\ln(\frac{\hat{P}(r,c)}{\hat{P}(r,+)\hat{P}(+,c)})$$

The adjusted likelihood ratio statistic is computed in an analogous manner to the Pearson adjustment where $\hat{\mathbf{\Delta}}$ is the same as before and

$$FG^2 = \frac{G^2}{tr\hat{\mathbf{\Delta}}}$$

with

$$d = \frac{(tr\hat{\mathbf{\Delta}})^2}{tr\hat{\mathbf{\Delta}}^2}.$$

Again, the asymptotic distribution of adjusted statistic FG^2 is approximated by the $F(d, d\nu)$ distribution where ν is the number of the sample design degrees of freedom.

Residuals

Under the independence hypothesis, the expected table proportion estimates are given by $\hat{E}(r,c) = \hat{P}(r,+)\hat{P}(+,c)$ and residual are defined as $\hat{R}(r,c) = \hat{P}(r,c) - \hat{E}(r,c)$ for $r = 1,\ldots,R$ and $c = 1,\ldots,C$.

Standardized residuals are computed by

$$\frac{\hat{R}(r,c)}{\sqrt{\hat{V}(\hat{R}(r,c))}},$$

$\hat{V}(\hat{R}(r,c))$ denotes the estimated residual variance for $r = 1,\ldots,R$ and $c = 1,\ldots,C$.

Let $\hat{\mathbf{M}} = [\hat{\mathbf{D}}_{\hat{\mathbf{P}}} - \hat{\mathbf{P}}\hat{\mathbf{P}}']$ estimate the asymptotic covariance matrix under simple random sampling where $\hat{\mathbf{P}}$ and $\mathbf{D}_{\hat{\mathbf{P}}}$ are defined as above. $\mathbf{X}$ is another contrast matrix specified by

$$\mathbf{X} = [\mathbf{A}_{\mathbf{R}} \otimes \mathbf{1}_{\mathbf{C}} \mid \mathbf{1}_{\mathbf{R}} \otimes \mathbf{A}_{\mathbf{C}}].$$

Contrast matrices $\mathbf{A}_R$ and $\mathbf{A}_C$, as well as the unit vectors $\mathbf{1}_R$ and $\mathbf{1}_C$, are defined as earlier.

Variance estimates for residuals are obtained from the diagonal of the following matrix:

$$\hat{\mathbf{V}}(\hat{\mathbf{R}}) = [\mathbf{I} - \hat{\mathbf{M}}\mathbf{X}(\mathbf{X}'\hat{\mathbf{M}}\mathbf{X})^{-1}\mathbf{X}']\hat{\mathbf{V}}(\hat{\mathbf{P}})[\mathbf{I} - \mathbf{X}(\mathbf{X}'\hat{\mathbf{M}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{M}}].$$

Odds Ratios and Risks

These statistics are computed only for 2×2 tables. If any layers are specified, they must correspond to the first stage strata.

Let $\hat{N}_{11}$, $\hat{N}_{12}$, $\hat{N}_{21}$ and $\hat{N}_{22}$ be the cell population size estimates, $\hat{N}_{1+}$, $\hat{N}_{2+}$, $\hat{N}_{+1}$, $\hat{N}_{+2}$ marginal estimates and $\hat{N}_{++}$ the population size estimate.

Estimates and variances

The odds ratio is defined by the following expression:

$$OR = \frac{\hat{N}_{11}\hat{N}_{22}}{\hat{N}_{12}\hat{N}_{21}}.$$

Relative risks are defined by

$$RR_1 = \frac{\hat{N}_{11}/\hat{N}_{1+}}{\hat{N}_{21}/\hat{N}_{2+}} \text{ and } RR_2 = \frac{\hat{N}_{12}/\hat{N}_{1+}}{\hat{N}_{22}/\hat{N}_{2+}}.$$

Risk differences are given by

$$D_1 = \frac{\hat{N}_{11}}{\hat{N}_{1+}} - \frac{\hat{N}_{21}}{\hat{N}_{2+}} \text{ and } D_2 = \frac{\hat{N}_{12}}{\hat{N}_{1+}} - \frac{\hat{N}_{22}}{\hat{N}_{2+}}.$$

The following substitutions of z_{ij} in the formulas for $\hat{V}(\hat{Y})$ are used for estimating variances:

- Odds ratio

$$z_{hij}(r,c) = w_{hij}\left(\frac{\delta_{hij}(1,1)}{\hat{N}_{11}} - \frac{\delta_{hij}(1,2)}{\hat{N}_{12}} - \frac{\delta_{hij}(2,1)}{\hat{N}_{21}} + \frac{\delta_{hij}(2,2)}{\hat{N}_{22}}\right) \times OR$$

- Risk ratio RR_1

$$z_{hij}(r,c) = w_{hij}\left(\frac{\delta_{hij}(1,1)\hat{N}_{12}}{\hat{N}_{11}\hat{N}_{1+}} - \frac{\delta_{hij}(1,2)}{\hat{N}_{1+}} - \frac{\delta_{hij}(2,1)\hat{N}_{22}}{\hat{N}_{21}\hat{N}_{2+}} + \frac{\delta_{hij}(2,2)}{\hat{N}_{2+}}\right) \times RR_1$$

- Risk difference D_1

$$z_{hij}(r,c) = w_{hij}\left(\frac{\delta_{hij}(1,1)\hat{N}_{12} - \delta_{hij}(1,2)\hat{N}_{11}}{\hat{N}_{1+}^2} - \frac{\delta_{hij}(2,1)\hat{N}_{22} - \delta_{hij}(2,2)\hat{N}_{21}}{\hat{N}_{2+}^2}\right)$$

The estimations of variance for RR_2 and D_2 are performed using similar substitutions.

Confidence limits

A level $1-\alpha$ confidence interval is constructed for a given $0 \le \alpha \le 1$ for odds ratio, risk ratio and risk difference in every table.

For the odds ratio or risk ratio R we use the logarithm transformation and obtain the confidence bounds

$$\ln(\hat{R}) \pm \frac{SE(\hat{R})}{\hat{R}} t_\nu(1-\alpha/2).$$

These bounds are transformed back to the original metric using the exponential function.

No transformations are used when estimating confidence bounds for a risk difference D:

$$\hat{D} \pm SE(\hat{D}) t_\nu(1-\alpha/2).$$

Tests of Homogeneity for One-Way Tables

Let the population be classified according to the values of a categorical row variable and possibly one a more categorical variables in the layer. Categories for the row variable are enumerated by $r = 1, \ldots, R$. When the layer variables are given we assume that their categories coincide with the strata in the first sampling stage. In the following we omit references to the layers as the formulas apply for each stratum separately when needed.

We study proportions $P(r) = N(r)/N(+)$. Test of homogeneity consists in testing the null hypothesis $\mathbf{H}_0 : P(r) = 1/R$ for $r = 1, \ldots, R-1$.

Adjusted Pearson statistic test

We perform an adjusted Pearson statistic test for testing the homogeneity. The Pearson test statistic is computed according to the following standard formula:

$$X^2 = n\sum_{r=1}^{R} R(\hat{P}(r) - 1/R)^2.$$

Since the asymptotic distribution of X^2 is generally not the chi-square distribution, we apply an adjustment using the $\hat{\mathbf{\Delta}}$ matrix given by:

$$\hat{\mathbf{\Delta}} = n(\hat{\mathbf{M}}(\hat{\mathbf{P}}_0))^{-1}\hat{\mathbf{V}}(\hat{\mathbf{P}}_0).$$

$\hat{\mathbf{V}}(\hat{\mathbf{P}}_0)$ is the estimated covariance matrix under the complex sample design, while $\hat{\mathbf{M}}(\hat{\mathbf{P}}_0)$ is an estimated asymptotic covariance matrix under the simple random sampling given by

$$\hat{\mathbf{M}}(\hat{\mathbf{P}}_0) = [diag(\hat{\mathbf{P}}_0) - \hat{\mathbf{P}}_0\hat{\mathbf{P}}_0'],$$

where $\hat{\mathbf{P}}_0$ is a vector and $diag(\hat{\mathbf{P}}_0)$ is a diagonal matrix of size $R-1$ containing elements $\hat{P}(r)$, $r = 1, \ldots, R-1$.

We use the F-based variant of the Rao and Scott's (1984) second-order adjustment

$$FX^2 = \frac{X^2}{tr\hat{\mathbf{\Delta}}}$$

with

$$d = \frac{(tr\hat{\mathbf{\Delta}})^2}{tr\hat{\mathbf{\Delta}}^2}.$$

The asymptotic distribution of FX^2 is approximated by the $F(d, d\nu)$ distribution where ν is the number of the sample design degrees of freedom.

Adjusted likelihood ratio test

The likelihood ratio test statistic is given by

$$G^2 = 2n\sum_{r=1}^{R} \hat{P}(r)\ln(R\hat{P}(r))$$

The adjusted likelihood ratio statistic is computed in an identical way as the adjustment for the Pearson statistic:

$$FG^2 = \frac{G^2}{tr\hat{\mathbf{\Delta}}}.$$

d and $\hat{\mathbf{\Delta}}$ are the same as specified before. Again, the asymptotic distribution of adjusted statistic FG^2 is approximated by the $F(d, d\nu)$ distribution where ν is the number of the sample design degrees of freedom.

References

Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.

Hansen, M. H., and Hurwitz, W. N. (1943), "On the theory of sampling from finite populations", *Annals of Mathematical Statistics*, volume 14, pages 333 - 362.

Horvitz, D. G., and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, volume 47, pages 663 - 685.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Kish, L. (1995), "Methods for Design Effects", *Journal of Official Statistics*, volume 11, pages 119 - 127.

Rao, J. N. K., and Scott, A. J. (1981), "The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables", *Journal of the American Statistical Association*, volume 76, pages 221-230.

Rao, J. N. K., and Scott, A. J. (1984),"On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data", *Anals of Statistics*, volume 12, pages 46-60.

Rao,p J. N. K., and Thomas, D. R. (2003), “Analysis of categorical response data from complex surveys: an upraisal and update”, In *Analysis of Survey Data*, ed. R.Chambers and C. Skinner. New York: John Wiley & Sons.

Särndal, C. E., Swenson, B., and Wretman, J. H. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Sen, A. R. (1953), “On the estimate of the variance in sampling with varying probabilities”, *Journal of the Indian Society of Agricultural Statistics*, volume 5, pages 55 - 77.

Thomas, D. R., and Rao, J. N. K. (1987), “Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling”, *Journal of the American Statistical Association*, volume 82, pages 630-636.

Woodruff, R. S. (1971), “A Simple Method for Approximating the Variance of a Complicated Estimate,” *Journal of the American Statistical Association*, volume 66, pages 411 – 414.

Yates, F., and Grundy, P. M. (1953), “Selection without replacement from within strata with probability proportional to size”, *Journal of the Royal Statistical Society Series B*, volume 15, pages 253 - 261.