Statistical Methods Used in EdSurvey

Developed by Paul Bailey and Michael Cohen[*†]

March 29, 2019

This document describes estimation procedures for the `EdSurvey` package. It includes the estimation of means (including regression analysis) and percentages; the estimation of correlation coefficients is covered in a vignette in the `wCorr` package.[1]

Which estimation procedure is used for any statistic appears in the Help file for the function that creates the statistic. For example, to find the estimation procedure used for the standard error of the regression coefficients, use `?lm.sdf` to see the manual entry.

This document uses many symbols; a table of the symbols is shown here as a reference. Terms used only once are defined immediately above or below equations, so they do not appear in this table.

Symbol	Meaning
A	A random variable
B	Another random variable
i	An index used for observations
j	An index used for jackknife replicates
J	The number of jackknife replicates
m	The number of plausible values
m^*	The number of plausible values used in a calculation
n	The number of units in the sample
p	An index used for plausible values
w_i	The ith unit's full sample weight
x_i	The ith unit's value for some variable
$\boldsymbol{X}$	A matrix of predictor variables in a regression
$\boldsymbol{y}$	A vector of predicted variables in a regression
$\boldsymbol{\beta}$	The regression coefficients in a regression
ϵ	The residual term in a regression
γ	The sampling variance multiplier
$\mathcal{A}$	The set of sampled units who are in a population of interest (e.g., Black females)
$\tilde{\mathcal{A}}$	The set of population units who are in a population of interest (e.g., Black females)
$\mathcal{U}$	The set of sampled units who are in a population that contains $\mathcal{A}$ (e.g., Black individuals)
$\tilde{\mathcal{U}}$	The set of population units who are in a population that contains $\mathcal{A}$ (e.g., Black individuals)

The remainder of this document describes estimation procedures that are used in the `EdSurvey` package. The next section describes the estimation of means, and the second section describes the estimation of percentages. Each section starts by describing estimation of the statistic, followed by estimation procedures of the variances of the statistic. Separate sections address situations where plausible values are present and situations where plausible values are not present. For sections on variance estimation, separate sections address the jackknife or Taylor series variance estimators.

*This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

†The authors would like to thank Dan Sherman, Qingshu Xie, and Ting Zhang for reviewing this document.

[1]See `vignette("wCorrFormulas",package="wCorr")`

Estimation of Weighted Means

This section concerns the estimation of means, including regression coefficients and the standard errors of means and regression coefficients.

Estimation of Weighted Means When Plausible Values Are Not Present

Weighted means are estimated according to

$$\mu_x = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where x_i and w_i are the outcome and weight of the ith unit (respectively) and n is the total number of units in the sample.

In the case of regression of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

a weighted regression is used so that the estimated coefficients ($\boldsymbol{\beta}$) minimize the weighted square residuals

$$\boldsymbol{\beta} = \text{ArgMin}_{\boldsymbol{b}} \sum_{i=1}^{n} w_i (y_i - \boldsymbol{X}_i \boldsymbol{b})^2$$

where $\boldsymbol{X}_i$ is the ith row of $\boldsymbol{X}$ and $\text{ArgMin}_{\boldsymbol{b}}$ means the value of $\boldsymbol{b}$ that minimizes the expression that follows it.

For binomial models (logit and probit) with an inverse link function $f(\cdot)$, the likelihood is maximized according to

$$\boldsymbol{\beta} = \text{ArgMin}_{\boldsymbol{b}} \sum_{i=1, y_i=1}^{n} w_i f(X_i \boldsymbol{b}) + \sum_{i=1, y_i=0}^{n} w_i [1 - f(X_i \boldsymbol{b})]$$

where the first sum is over all of the units where the outcome is a success (a one) and the second sum is over all units where the outcome is a failure (a zero).

Estimation of Weighted Means When Plausible Values Are Present

When the variable x has plausible values, then these are used to form the mean estimate (μ) according to

$$\mu = \frac{1}{m} \sum_{p=1}^{m} \frac{\sum_{i=1}^{n} w_i x_{ip}}{\sum_{i=1}^{n} w_i}$$

where x_{ip} is the pth plausible value for the ith unit's outcome, and there are m plausible values for each unit.

For regressions, the coefficient estimates are simply averaged over the plausible values

$$\boldsymbol{\beta} = \frac{1}{m} \sum_{p=1}^{m} \boldsymbol{\beta}_p$$

where $\boldsymbol{\beta}_p$ is the vector of estimated regression coefficients, calculated using the pth set of plausible values.

Estimation of the Coefficient of Determination in a Weighted Linear Regression

In regression analysis, statistics such as the coefficient of determination (or R-squared) are estimated across all observations. For these statistics, their values are averaged across the regression runs (one per set of plausible values). For example,

$$R^2 = \frac{1}{m}\sum_{p=1}^{m} R_p^2$$

where R_p^2 is the R-squared value for the regression run with the pth set of plausible values.

For a particular regression, the R-squared is defined in Weisberg (1985, eq. 2.31) as

$$R^2 = 1 - \frac{RSS}{SYY}$$

where $RSS = \boldsymbol{e}^T\boldsymbol{W}\boldsymbol{e}$ (Weisberg, 1985, eq. 4.2), and $SYY = (\boldsymbol{y} - \bar{y})^T\boldsymbol{W}(\boldsymbol{y} - \bar{y})$, $\bar{y}$ is the weighted mean of the outcome, and $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Estimation of Standard Deviations

Weighted variance estimates dispersion in the variable in question and are calculated according to

$$\hat{s}^2 = \frac{\sum_{i=1}^{n} w_i(x_i - \mu_x)^2}{\sum_{i=1}^{n} w_i}$$

where μ_x is the weighted mean. When there are several plausible values, the variance is averaged across the plausible values (calculating μ_x per plausible value).

The estimate of the standard deviation is the square root of the estimated variance.

Estimation of Standardized Regression Coefficients

Using the definition of the standardized regression coefficients (b)

$$b_j = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_{X_j}}\beta_j$$

where b_j is the standardized regression coefficient associated with the jth regressor, $\tilde{\sigma}_y$ is the weighted standard deviation of the outcome variable, and $\tilde{\sigma}_{X_j}$ is the standard deviation of the jth regressor.

Default Variance Estimation of Standardized Regression Coefficients

The default standard error of the standardized regression coefficients then treats the standard error estimates as constants and is

$$\sigma_{b_j} = \frac{\sigma_y}{\sigma_{X_j}}\sigma_{\beta_j}$$

Sampling Variance Estimation of Standardized Regression Coefficients

An alternative method estimates the standardized regression coefficients using the same process but estimates their standard error accounting for the design-based sampling variance.

In this method, the standardized regression coefficients are estimated per plausible value and jackknife replicate. When the standardized regression coefficient is estimated for a plausible value, the overall variance of the outcome ($\tilde{\sigma}y$) and the regressors ($\tilde{\sigma}_{X_j}$) are used. For a jackknife replicate, the values of $\tilde{\sigma}_y$ and $\tilde{\sigma}_{X_j}$ are updated with the jackknife replicate weights.

Estimating the variance of the standardized regression coefficient proceeds identically to estimating the variance of the regressors, and the weighted standard deviations are also updated with the jackknife replicate weights.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Jackknife Method

When the predicted value does not have plausible values and the requested variance method is jackknife, the variance of the coefficients ($\boldsymbol{V}_J$) is estimated as

$$\boldsymbol{V}_J = \boldsymbol{V}_{jrr,0} = \gamma \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)^2$$

where γ is a constant equal to 1 for the jackknife variance estimation method; its inclusion allows us to extend the equations to other variance estimation methods, such as balanced repeated replication (Wolter, 2007) or Fay's method (Judkins, 1990); $\boldsymbol{\beta}_j$ are the coefficients estimated with the jth jackknife replicate weights, and $\boldsymbol{\beta}_0$ are the coefficients estimated with the sample weights; and J is the total number of jackknife replicate weights.

The covariance between β_l and β_m ($C_{J;lm}$) is estimated as

$$C_{J;lm} = C_{jrr,0;lm} = \gamma \sum_{j=1}^{J} (\beta_{j;l} - \beta_{0;l})(\beta_{j;m} - \beta_{0;m})$$

where subscripts after the semicolon indicate the matrix element (two subscripts) of the covariance matrix ($\boldsymbol{C}$) or the vector element (one subscript) for the estimate vector $\boldsymbol{\beta}$. The other subscripts are as with the variance estimation.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Present, Using the Jackknife Method

When the predicted value has plausible values and the requested variance method is jackknife, the variance ($\boldsymbol{V}_{JP}$) is estimated as the sum of a variance component from the plausible values (also called imputation values, so that the variance term is called $\boldsymbol{V}_{imp}$) and the sampling variance using plausible values ($\boldsymbol{V}_{jrr,P}$) is estimated according to the following formula:

$$\boldsymbol{V}_{JP} = \boldsymbol{V}_{imp} + \boldsymbol{V}_{jrr,P}$$

The sampling variance is

$$\boldsymbol{V}_{jrr,P} = \frac{1}{m^*} \sum_{i=1}^{m^*} \boldsymbol{V}_{jrr,p} \ .$$

Note that in this equation, m^* is a number that can be as small as 1 or as large as the number of plausible values.[2] In the previous equation, $\boldsymbol{V}_{jrr,P}$ is the average of $\boldsymbol{V}_{jrr,p}$ over the plausible values, and the values of $\boldsymbol{V}_{jrr,p}$ are calculated in a way analogous to $\boldsymbol{V}_{jrr,0}$ in the previous section, except that the pth plausible values are used within each step:

$$\boldsymbol{V}_{jrr,p} = \gamma \sum_{j=1}^{J} (\boldsymbol{\beta}_{jp} - \boldsymbol{\beta}_{0p})^2$$

The imputation variance is estimated according to Rubin (1987):

$$\boldsymbol{V}_{imp} = \frac{m+1}{m(m-1)} \sum_{p=1}^{m} (\boldsymbol{\beta}_p - \boldsymbol{\beta})^2$$

where m is the number of plausible values, $\boldsymbol{\beta}_p$ is the vector of coefficients calculated with the pth set of plausible values, and $\boldsymbol{\beta}$ is the estimated coefficient vector averaged over all plausible values.

Covariance terms between β_l and β_m are estimated according to

$$C_{JP;lm} = C_{imp;lm} + C_{jrr,P;lm}$$

where subscripts after a semicolon indicate the indexes of the covariance term being identified,

$$C_{jrr,p;lm} = \gamma \sum_{j=1}^{J} (\beta_{jp;l} - \beta_{0p;l})(\beta_{jp;m} - \beta_{0p;m})$$

and

$$\boldsymbol{C}_{imp;lm} = \frac{m+1}{m(m-1)} \sum_{p=1}^{m} (\beta_{p;l} - \beta_l)(\beta_{p;m} - \beta_m)$$

where β_l and β_m are the estimates averaged across all of the plausible values.

Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Taylor Series Method

When the predicted value does not have plausible values and the requested variance method is the Taylor series, the variance of the coefficients ($\boldsymbol{V}_T$) is estimated as is shown in this section.

When an outcome does not have plausible values and the estimator can be rewritten as a parameter estimated by maximum likelihood using a sum of individual-level scores, such as

$$\sum_i w_i \ell_i(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = 0 \tag{1}$$

where w_i, y_i, $\boldsymbol{X}_i$ is the ith unit's full sample weight, outcome, and covariates, respectively, while $\boldsymbol{\beta}$ is the parameters of interest; then, the variance of the parameters can be estimated by Taylor series estimation as described in this section (Binder, 1983). Note that because they are weighted by sample weights, the likelihoods here are pseudo-likelihoods.

As an example, in the case of linear regression, the maximum likelihood estimator can be written in the form of eq. 1 with[3]

$$\ell_i = -\frac{1}{2} ln\left(2\pi\sigma^2\right) - \frac{(y_i - \boldsymbol{X}_i\boldsymbol{\beta})}{2\sigma^2} \tag{2}$$

[2]This option is included because any value for m^* gives an estimate of $\boldsymbol{V}_{jrr}$ with the same properties as larger values of m^* (they are unbiased under the same conditions), but larger values of m^* can take substantially longer to compute. The value of m^* is set with the `jrrIMax` argument; note that `jrrIMax` affects the estimation of V_{jrr} only.

[3]This formula ignores the fact that σ is estimated because the EdSurvey package does not estimate the variance of the variance (the variance of σ^2).

The Taylor series variance estimate of the coefficients ($\boldsymbol{V}_T$) is estimated as[4]

$$\boldsymbol{V}_T = \boldsymbol{V}_{Taylor,0}(\boldsymbol{\beta}) = \boldsymbol{D}^T \boldsymbol{Z} \boldsymbol{D} \tag{3}$$

where $\boldsymbol{D}$ is the inverse Hessian (the typical variance estimator)

$$\boldsymbol{D} = \left[\frac{\partial^2 \sum_i w_i \ell_i(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \right]^{-1} \tag{4}$$

evaluated at the maximum likelihood estimate for $\boldsymbol{\beta}$; the $\boldsymbol{Z}$ matrix is a sample-design based estimator of the variance of the sum of the scores (the sum of the ℓ_i terms).

For a two-staged survey sample design frequently used by NCES, the estimator of the variance of a sum, such as the ℓ_i terms is

$$\boldsymbol{Z} = \sum_{j=1}^{J} \frac{n_s}{n_s - 1} \sum_{u=1}^{n_s} \boldsymbol{z}_{uj} \boldsymbol{z}_{uj}^T \tag{5}$$

where each $\boldsymbol{z}$ vector is the mean deviation in the primary sampling unit's (PSU's) score vector from the stratum average (defined below); the inner sum is over the sampled PSUs (u), of which there are n_s, and the outer sum is over all units in the jackknife replicate strata (j), of which there are J. Note that only strata with at least two PSUs that have students are included in the sum; others are simply excluded.[5]

To find $\boldsymbol{z}$, first define the term

$$U_{ik} = \frac{\partial \ell_i}{\partial \beta_k} \tag{6}$$

in the case of linear regression, that is

$$\frac{\partial \ell_i}{\partial \beta_k} = (y_i - \boldsymbol{X}_i \boldsymbol{\beta}) X_{ik} w_i \tag{7}$$

for generalized linear models with an inverse link function $f(\eta_i)$ where $\eta_i = X_i \beta$, then

$$\frac{\partial \ell_i}{\partial \beta_k} = \frac{\partial f(\eta_i)}{\partial \eta_i} X_{ik} w_i \tag{8}$$

where i indexes the matrix row (observation) and k indexes the matrix column (regressor). For logistic regression, the inverse link function and its derivative are

$$f(\eta) = \frac{1}{1 - e^{-x}} \qquad \frac{\partial f(\eta_i)}{\partial \eta_i} = \frac{e^{-x}}{(1 - e^{-x})^2} \tag{9}$$

while for probit, the inverse link function and its derivative are

$$f(\eta) = \Phi(\eta)^{-1} \qquad \frac{\partial f(\eta_i)}{\partial \eta_i} = \phi(\eta) \tag{10}$$

where $\Phi(\eta)^{-1}$ is the distribution function of the standard normal and $\phi(\eta)$ is the density function of the standard normal.

[4]This is a slight generalization of Binder (1983, sec. 4.2) to the weighted case. This is derived in more detail and with notation more closely aligned to Binder in the AM Manual (Cohen, 2002; see Tools: Procedures: Other Available Procedures: Regression: Details).

[5]This leads to a downward bias in the estimated variance. When the number of units excluded by this rule is proportionally small, the bias also should be proportionally small.

The kth entry of $\boldsymbol{z}_{uj}$ is then given by

$$z_{ujk} = \sum_{i \in \mathcal{Q}_{uj}}^{n_s} \left[U_{ik} - \left(\frac{1}{n_s} \sum_{i' \in \mathcal{Q}_j}^{n_s} U_{i'k} \right) \right] \tag{11}$$

where $\mathcal{Q}_{uj}$ is the indices for observations in the uth PSU of the jth stratum, and $\mathcal{Q}_j$ is the indices for observations in the jth stratum (across all PSUs). Thus, when two PSUs are selected per stratum, the value of $\boldsymbol{z}$ will be related by $\boldsymbol{z}_{1j} = -\boldsymbol{z}_{2j}$.

Notice that $\boldsymbol{V}_T$ is defined as a matrix, so the variance estimates of the $\boldsymbol{\beta}$ parameters are along the diagonal of the matrix, while covariances of the ith and jth elements appear in the i, j element (and j, i element).

Estimation of Standard Errors of Weighted Means When Plausible Values Are Present, Using the Taylor Series Method

When the predicted value has plausible values and the requested variance method is the Taylor series, the variance of the coefficients ($\boldsymbol{V}_{TP}$) is estimated as

$$\boldsymbol{V}_{TP} = \boldsymbol{V}_{Taylor,P}(\boldsymbol{\beta}) + \boldsymbol{V}_{imp}$$

where the equation for $\boldsymbol{V}_{imp}$ (and $\boldsymbol{C}_{imp}$) are given in the section on the jackknife variance estimator and where the $\boldsymbol{V}_{Taylor,P}$ are averaged over the plausible values according to

$$\boldsymbol{V}_{Taylor,P} = \frac{1}{m} \sum_{p=1}^{m} \boldsymbol{V}_{Taylor}(\boldsymbol{\beta}_p)$$

where $\boldsymbol{V}_{Taylor}(\boldsymbol{\beta}_p)$ is calculated as in the previous section, using the pth plausible values to form $\frac{\partial \ell_i}{\partial \beta_k}$. For example, in the case of linear regression,

$$\frac{\partial \ell_{ip}}{\partial \beta_{kp}} = (y_{ip} - \boldsymbol{X}_i \boldsymbol{\beta}_p) X_{ik} w_i \tag{12}$$

where ℓ_{ip} and y_{ip} are the likelihood and outcome of the ith observation for the pth plausible value, and β_{kp} is the kth regression coefficient for the pth plausible value. The remainder of the calculation of U_{ik} and z_{ujk} are otherwise identical.

Estimation of Standard Errors of Differences of Means

In some cases, two means (μ_1 and μ_2) are calculated for potentially overlapping groups. When this is done, the difference ($\Delta = \mu_1 - \mu_2$) may be of interest. When a covariance term is available, the variance of this difference is then estimated using the formula

$$\sigma_\Delta^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

where σ_{12} is the covariance of μ_1 and μ_2.

When there is not a covariance term available, the approximate equation is used

$$\sigma_\Delta^2 = \sigma_1^2 + \sigma_2^2 - 2p\sigma_m^2$$

where a subscript of m is used to indicate the subset that has more observations, and p is the ratio of the number of observations used in calculating both μ_1 and μ_2 divided by the n-size of the larger of the two samples.

When one group is a subset of the other, this is a part-whole comparison. Part-whole comparisons in the context of the National Assessment of Educational Progress (NAEP) apply to comparisons between jurisdictions. Please use with caution when implementing this method with other types of gap comparisons.

Estimation of Weighted Percentages

Percentages are used to estimate the proportion of individuals in a group who have some characteristic; for example, the percentage of Blacks who are female. This often is called a "domain." In the population, the universe is the set $\tilde{\mathcal{U}}$; in the example, $\tilde{\mathcal{U}}$ is Blacks who are eligible for sampling. The tilde is used to indicate that this set is in the population.[6] The sought-after percentage is then the percentage of individuals in the subset $\tilde{\mathcal{A}} \subseteq \tilde{\mathcal{U}}$. In the example, $\tilde{\mathcal{A}}$ is the set of Black females who are eligible for sampling. The percentage for which an estimate is desired is then 100 times the number of individuals in $\tilde{\mathcal{A}}$ divided by the number of individuals in $\tilde{\mathcal{U}}$. Mathematically,

$$\Pi = 100 \times \frac{|\tilde{\mathcal{A}}|}{|\tilde{\mathcal{U}}|}$$

where $|\cdot|$ is the cardinality, or the count of the number of members in a set. Note that in this example, $\tilde{\mathcal{U}}$ was itself a subset of the entire eligible population. In other cases, $\tilde{\mathcal{U}}$ simply could be the population of eligible individuals. Then, the value Π would represent the percentage of eligible individuals who were Black females.

The remainder of this section describes statistics meant to estimate Π and the variance of those estimates.

Estimation of Weighted Percentages When Plausible Values Are Not Present

In the sample, units are identified as in $\mathcal{A}$ and $\mathcal{U}$ (where the tilde is dropped to indicate they are the sampled sets) and the estimator is[7]

$$\pi = 100 \times \frac{\sum_{i \in \mathcal{A}} w_i}{\sum_{i \in \mathcal{U}} w_i}$$

where π is the estimated percentage.

Another statistic of interest is the weighted sample size of $\mathcal{A}$, or an estimate of the number of individuals in the population who are members of $\tilde{\mathcal{A}}$. This is calculated with $\sum_{i \in \mathcal{A}} w_i$.

Estimation of Weighted Percentages When Plausible Values Are Present

If membership in $\mathcal{A}$ or both $\mathcal{A}$ and $\mathcal{U}$ is dependent on a measured score being in a range, then the value of Π is estimated once for each set of plausible values (indexed by p) by

$$\pi = 100 \times \frac{1}{m} \sum_{p=1}^{m} \frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i}$$

In the case where membership in $\mathcal{U}$ is not associated with the plausible value, $\mathcal{U}_p$ will be the same for all sets of plausible values. The same applies for $\mathcal{A}_p$.

Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Not Present, Using the Jackknife Method

When membership in $\mathcal{A}$ and $\mathcal{U}$ are not dependent on plausible values and the requested variance method is jackknife, the variance of the percentage ($\boldsymbol{V}_{\pi,J}$) is estimated as

$$\boldsymbol{V}_{\pi,J} = 100^2 \times \boldsymbol{V}_{jrr,f,0}$$

[6] When the tilde is not present, the set is just of individuals in the sample.

[7] The notation $i \in \mathcal{A}$ is a bit of an abuse of notation. Strictly speaking, it is the unit that is in $\mathcal{A}$ and $\mathcal{U}$, not the indices.

where the jackknife variance of the fraction is given by

$$\boldsymbol{V}_{jrr,f,0} = \gamma \sum_{j=1}^{J} \left(\frac{\sum_{i \in \mathcal{A}} w_{ij}}{\sum_{i \in \mathcal{U}} w_{ij}} - \frac{\sum_{i \in \mathcal{A}} w_i}{\sum_{i \in \mathcal{U}} w_i} \right)^2$$

The subscript j is used to indicate that the weights for the jth jackknife replicates are being used, and weights that do not contain a second subscript are the student full sample weights.

Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Present, Using the Jackknife Method

When membership in $\mathcal{A}$ and $\mathcal{U}$ are dependent on plausible values and the requested variance method is jackknife, the variance of the percentage ($\boldsymbol{V}_{\pi,JP}$) is estimated as

$$\boldsymbol{V}_{\pi,TP} = 100^2 \times (\boldsymbol{V}_{jrr,f,P} + \boldsymbol{V}_{imp,f})$$

Here, the only modification to $\boldsymbol{V}_{jrr,f}$ to make it $\boldsymbol{V}_{jrr,f,P}$ is that the sets $\mathcal{A}$ and $\mathcal{U}$ must be modified to regard one set of plausible values.

$$\boldsymbol{V}_{jrr,f,P} = \gamma \frac{1}{m^*} \sum_{p=1}^{m^*} \sum_{j=1}^{J} \left(\frac{\sum_{i \in \mathcal{A}_p} w_{ij}}{\sum_{i \in \mathcal{U}_p} w_{ij}} - \frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} \right)^2$$

where the subscript j is used to indicate that the weights for the jth jackknife replicates are being used, weights that do not contain a second subscript are the student full sample weights, and the subscript p indicates the plausible values being used. Note that in some situations, the $\mathcal{A}_p$ will be identical to each other across all plausible values, and the $\mathcal{U}_p$ will be identical to each other in a broader set of situations.

The value of $V_{imp,f}$ is given by

$$V_{imp,f} = \frac{m+1}{m(m-1)} \sum_{p=1}^{m} \left(\frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} - \frac{1}{m} \sum_{p'=1}^{m} \frac{\sum_{i \in \mathcal{A}_{p'}} w_i}{\sum_{i \in \mathcal{U}_{p'}} w_i} \right)^2$$

so that the second sum is simply the average over all plausible values and represents the estimate itself (π), and the expression could be rewritten slightly more compactly as

$$V_{imp,f} = \frac{m+1}{m(m-1)} \sum_{p=1}^{m} \left(\frac{\sum_{i \in \mathcal{A}_p} w_i}{\sum_{i \in \mathcal{U}_p} w_i} - \frac{\pi}{100} \right)^2$$

Estimation of the Standard Error of Weighted Percentages When Plausible Values Are Not Present, Using the Taylor Series Method

When membership in $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{U}}$ are not dependent on plausible values and the requested variance method is the Taylor series, the variance-covariance matrix $\boldsymbol{V}_{\pi,JP}$ of the coefficients is estimated as

$$\boldsymbol{V}_{\pi,T} = 100^2 \times \begin{bmatrix} \boldsymbol{DZD} & -\boldsymbol{DZD1} \\ -\mathbf{1}^T\boldsymbol{DZD} & \mathbf{1}^T\boldsymbol{DZD1} \end{bmatrix}$$

where the block matrix has elements $DZD \in \mathbb{R}^{(c-1)\times(c-1)}$; the cth row and column are then products of DZD, and the vector $\mathbf{1} \in \mathbb{R}^{c-1}$ has a 1 in every element; the definition of $\boldsymbol{D}$ is the inverse of a matrix of derivatives of a score vector, taken with respect to $\boldsymbol{\pi}$; and Z is a variance estimate of the proportions based on the sample survey. This is based on results derived here, following Binder (1983, sec. 3.2).

The score function in question is

$$S(\pi_h) = \left(\sum_{i=1}^{n} w_i \mathbb{1}(\text{unit } i \text{ is in class } h) \right) - \left(\sum_{i=1}^{n} \pi_h w_i \right)$$

Setting the score function to zero and solving yields the parameter estimator shown in the section "Estimation of Weighted Percentages When Plausible Values Are Present," less the factor of 100 that converts a proportion to a percentage.

For the first $c-1$ elements of $\boldsymbol{\pi}$, when this function is solved for π_h, the solution is the estimate of π_h shown earlier:

$$\pi_h = \frac{\sum_{i=1}^{n} w_i \mathbb{1}(\text{unit } i \text{ is in class } h)}{\sum_{i=1}^{n} w_i}$$

For π_c, the definition is that

$$\pi_c = 1 - \sum_{k=1}^{c-1} \pi_k$$

and with some algebraic rearrangement, this becomes

$$= \frac{\sum_{i=1}^{n} w_i \mathbb{1}(\text{unit } i \text{ is in class } c)}{\sum_{i=1}^{n} w_i}$$

The value of D is then the derivative of $S(\pi)$ with respect to π. Because this derivative must be calculated in total equilibrium (so that all the percentages add up to 100), this is done for the first $c-1$ items, and the variance of π_c is separately calculated. Taking the derivative of $S(\pi)$ and then inverting it shows that $\boldsymbol{D} \in \mathbb{R}^{(c-1)\times(c-1)}$ is a diagonal matrix with entries $\frac{1}{\sum_{i=1}^{n} w_i}$.

Then, the $\boldsymbol{Z}$ matrix accounts is given by

$$\boldsymbol{Z} = \sum_{s=1}^{N_s} \frac{n_s}{n_s - 1} \sum_{j=1}^{n_s} \boldsymbol{U}_{sk}^T \boldsymbol{U}_{sk}$$

where N_s is the number of strata, n_s is the number of PSUs in a stratum, and $\boldsymbol{U}_{sk}$ is the vector of mean score deviates given by

$$\boldsymbol{U}_{sk} = \sum_{l=1}^{n_{sk}} \boldsymbol{S}_{skl}(\boldsymbol{\pi}) - \frac{1}{n_s} \sum_{j=1}^{n_s} \sum_{l=1}^{n_{sj}} \boldsymbol{S}_{sjl}(\boldsymbol{\pi})$$

where n_{sk} is the number of observations in PSU k and in stratum s, l is an index for individuals within the stratum and PSU, and the score vector is given by

$$\boldsymbol{S}_{skl}(\boldsymbol{\pi}) = w_{skl} \boldsymbol{e}_{skl} - w_{skl} \boldsymbol{\pi}$$

where $\boldsymbol{e}_{skl}$ is a vector that is 0 in all entries except for a single 1 for the class that the unit is in. For example, if a respondent is a male and the possible levels are ("Female", "Male"), then their level of $\boldsymbol{e}_{skl}$ would be $(0,1)^T$.

This gives the covariance matrix for the first $c-1$ elements of the $\boldsymbol{\pi}$ vector. Using the usual formula for variance and covariance, it is easy to see that the variance for the final row and column are as shown at the beginning of this section.[8]

[8] However, the matrix need not be calculated in this fashion. Instead, the final row and column (the covariance terms associated with the value π_c) need not be dropped. They can be simply included in the formulation of D and S along with every other term.

Two heuristic arguments are offered for this. First, the variance terms are all exchangeable, so the same formula that applies to the first term applies to the final term under reordering. Thus, any term in the covariance matrix can be found by simply permuting the covariance matrix so that the term is not in the cth row or column. Because of that, the method for calculating the upper left portion of the block matrix clearly applies to the cth row and column, which can be calculated directly. Some experiments with NAEP data show that the two methods agree.

The second heuristic argument is that the values of π_h already meet the requirement of summing up to 1 when the score vector is set equal to zero and solved. This means that the constraint does not need to be imposed a second time.

Estimation of Degrees of Freedom

One method of estimating the degrees of freedom for education survey results is to find the sum of the number of PSUs (schools) and subtract from it the number of strata. For NAEP surveys, this results in an estimate of 62.

However, many estimates do not use variation across all PSUs, so the degrees of freedom should be expected to be smaller.

Estimation of Degrees of Freedom, Using the Jackknife

When the jackknife estimator is used, the Welch-Satterthwaite (Satterthwaite 1946; Welch, 1947) degrees of freedom estimate (dof_{WS}) is

$$dof_{WS} = \frac{1}{m^*} \sum_{p=1}^{m^*} \frac{\left[\sum_{j=1}^{J} \left[(A_{jp} - A_{0p}) - (B_{jp} - B_{0p})\right]^2\right]^2}{\sum_{j=1}^{J} \left[(A_{jp} - A_{0p}) - (B_{jp} - B_{0p})\right]^4}$$

where A_{jp} is the estimate of A using the jth jackknife replicate value and the pth plausible value, and A_{0p} is the estimate of A using the full sample weights and the pth plausible value; the same is true for the B subscripts. In the case of a regression coefficient, $(A_{jp} - A_{0p}) - (B_{jp} - B_{0p})$ is replaced by the jth deviate from the full sample weight coefficient $\beta_{jp} - \beta_{0p}$.

The Johnson Rust corrected degrees of freedom (dof_{JR}) is

$$dof_{JR} = \left(3.16 - \frac{2.77}{\sqrt{J}}\right) dof_{WS}$$

Estimation of Degrees of Freedom, Using the Taylor Series

For the Taylor series estimator, the degrees of freedom estimator also uses the Welch-Satterthwaite (WS) degrees of freedom estimate (dof_{WS}). However, because the jackknife replicate estimates are not available, a different equation is used. The WS weights require an estimate of the degrees of freedom per group; when using the jackknife variance estimator, this was estimated using the jackknife replicate weights. For the Taylor series, this is estimated per stratum.

The WS (Satterthwaite, 1946) equation states that when

$$v = \sum_{j=1}^{n} s_j^2 \tag{13}$$

where v is the variance estimator for a statistic, then

$$\text{dof}(v) = \frac{\left(\sum_{j=1}^{n} s_j^2\right)^2}{\sum_{j=1}^{n} s_j^4} \tag{14}$$

For the Taylor series variance estimator, the vector of variance estimators ($\boldsymbol{v}$) is

$$\boldsymbol{v} = \text{diag}(\boldsymbol{DZD}) \tag{15}$$

where diag$(\cdot)$ is the function that extracts the diagonal of a matrix, and the matrices $\boldsymbol{D}$ and $\boldsymbol{Z}$ are defined in the section "Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Taylor Series Method."

Because eq. 15 is not obviously a sum, it must be rewritten as one. The matrix $\boldsymbol{Z}$ is a sum, which can be simplified from eq. 5 to a sum of matrices by stratum, indexed by j

$$\boldsymbol{Z} = \sum_{j=1}^{J} \boldsymbol{Z}_j \tag{16}$$

Then

$$\boldsymbol{v} = \text{diag}\left[\boldsymbol{D}\left(\sum_{j=1}^{J} \boldsymbol{Z}_j\right)\boldsymbol{D}\right] \tag{17}$$

$$= \text{diag}\left(\sum_{j=1}^{J} \boldsymbol{D}\boldsymbol{Z}_j\boldsymbol{D}\right) \tag{18}$$

using

$$s_j^2 = (\boldsymbol{D}\boldsymbol{Z}_j\boldsymbol{D})_{ii} \tag{19}$$

Letting the subscript ii indicate the ith diagonal element of the matrix, eq. 18 can then be plugged into eq. 14

$$\text{dof}(v_i) = \frac{\left[\sum_{j=1}^{J} (\boldsymbol{D}\boldsymbol{Z}_j\boldsymbol{D})_{ii}\right]^2}{\sum_{j=1}^{J} (\boldsymbol{D}\boldsymbol{Z}_j\boldsymbol{D})_{ii}^2} \tag{20}$$

With multiple plausible values, the average of dof_{WS} across the plausible values can also be used.

The same equation is used for the Johnson Rust corrected degrees of freedom as with the jackknife.

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*(3), 279–292.

Cohen, J. (2002). *AM manual*. Washington, DC: American Institutes for Research.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Education Statistics*, *17*(2), 175–190.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, *6*(5), 223–239.

National Center for Education Statistics. (2009). t *test for partly overlapping groups*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/2004_2005/infer_compare2_overlap.aspx

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.

Weisberg, S. (1985). *Applied linear regression*. New York, NY: Wiley.

Welch, B. L. (1947) The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, *34*, 28–35.

Wolters, K. M. (2007). *Introduction to variance estimation* (2nd ed.) New York, NY: Springer.