S-31322 SAS - Calculate Summary Variances and Relative Standard Errors (RSE)

As a NO User,

I need the system to have the control/function to execute actions to calculate Summary Variances (SV) and Summary Relative Standard Errors (RSE) using the stored data and algorithm, so that the publishablility and reliability of the estimates can be determined.

1. Starting Point
   This process will generate sampling errors and variances for all levels of estimates after State and national summary estimates have been calculated.

2. Definitions and assumptions
   The SOII uses a Taylor series linearization methodology to calculate estimates of standard errors for published estimates.

   a. **Variance Estimation:**
      Calculate the variance for **totals of case counts** for each case type (TRC, DART, DAFW, DJTR, INJU, ORC, ILLN, SKIN, RESP, POIS, HEAR, OTHR), hours, and employment using the following unbiased formula for weighted samples

      i. **Where the number of usable units in the estimation stratum and the number of units in the sampling frame (which is the sum of final weights for this estimation stratum) equal 1**, set the variance equal to 0

      ii. **Where the number of usable units in the estimation stratum is greater than 1**, the variance is calculated by the following:

$$var(\hat{x}) = var(N\bar{x}) = N^2 * var(\bar{x}) = N^2 \frac{var(x)}{n}$$

$$¿ N^2 \frac{1}{n} ¿$$

$$¿ \frac{N^2}{n} ¿$$

where

- $\hat{x}$is the total weighted estimate of the variable (case counts for each case type, hours, employment) for all usable units in the estimation stratum: $\hat{x} = \sum_{i=1}^{n} w_i x_i$
- N is the sum of final weights for usable units in the estimation stratum: $N = \sum_{i=1}^{n} w_i$
- n is the number of usable units reporting in the estimation stratum
- ow_i is the original state sampling weight of establishment *i* in the estimation
- w_i is the final weight of establishment *I* in the estimation stratum
- x_i is the reported value of the variable (case counts for each case type, hours, employment) for establishment *i* in the estimation stratum

- $\bar{x}$ is the weighted average value of the variable (case counts for each case type, hours, employment) for all usable units in the estimation stratum: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

iii. Where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1, the mean of the stratum, $\bar{x}$, cannot be used because it will yield an incorrect variance of 0. To correct for this, use the mean from the roll-up hierarchy. That is,

- If there is only one usable unit in the estimation stratum and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1, calculate the variance as the variance of the roll-up level where the mean is the mean of the survey year, state, ownership, size class, and parent TEI of the usable units in this roll-up level. However, if the one usable unit in the estimation stratum is also the only usable unit for the survey year, state, ownership, size class, and parent TEI, continue to roll-up by parent TEI level as far as the industry domain level until the number of usable units is greater than 1.

- If the one usable unit in the estimation stratum is also the only usable unit for the survey year, state, ownership, size class, and industry domain, calculate the variance as the variance of the roll-up level where the mean is the mean of survey year, state, ownership, size class 0 (all sizes) and TEI. Continue to roll-up the parent TEI level as far as TEI000000 (all industries) until the number of usable units is greater than 1.

Therefore, **where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1**, we assign the roll-up level variance for this single usable unit; in other words, we use the roll-up level variance to approximate the variance for this single unit estimation stratum. The variance is calculated in the same way as those strata with more than one usable unit; it just includes all usable units at the roll-up level. Specifically the variance for the stratum with single usable unit is calculated as the following:

$$var(\hat{x}) = \text{¿} var(\hat{x}_{RU}) = var(N\bar{x}_{RU}) = N^2 * var(\bar{x}_{RU}) = N^2 \frac{var(x)_{RU}}{n_{RU}}$$

$$\text{¿} N^2 \frac{1}{n_{RU}} \text{¿}$$

$$\text{¿} \frac{N^2}{n_{RU}} \text{¿}$$

where

- $\hat{x}_{RU}$ is the total weighted estimate of the variable (case counts for each case type, hours, employment) for all usable units in the roll-up level: $\hat{x}_{RU}=\sum_{i=1}^{n_{RU}} w_i x_i$
- N is the sum of final weights for usable units in the estimation stratum: $N=\sum_{i=1}^{n} w_i$, in this case it is the single usable unit estimation stratum, so N= w_i
- N_{RU} is the sum of final weights for usable units in the roll-up level: $N_{RU}=\sum_{i=1}^{n_{RU}} w_i$
- n_{RU} is the number of usable units reporting in the roll-up level
- ow_i is the original state sampling weight of establishment *i* in the roll-up level
- w_i is the final weight of establishment *i* in the roll-up level
- x_i is the reported value of the variable (case counts for each case type, hours, employment) for establishment *i* in the roll-up level
- $\overline{x}_{RU}$ is the weighted average value of the variable (case counts for each case type, hours, employment) for all usable units in the roll-up level:

  $$\overline{x}_{RU}=\frac{\sum_{i=1}^{n_{RU}} w_i x_i}{\sum_{i=1}^{n_{RU}} w_i}$$

b. Covariance Estimation:

Calculate the **covariance for total case counts** for each case type with hours

i. **Where the number of usable units in the estimation stratum and the number of units in the sampling frame (which is the sum of final weight in this estimation stratum) equal 1**, set the covariance equal to 0.

ii. **Where the number of usable units in the estimation stratum is greater than 1**, the covariance is calculated by the following:

$$cov(\hat{x},\hat{y})=N^2\frac{1}{n}¿$$

$$¿\frac{N^2}{n}¿$$

where

- $\hat{x}$is the total weighted estimate of the variable (case counts for each case type) for all usable units in the estimation stratum: $\hat{x}=\sum_{i=1}^{n} w_i x_i$
- $\hat{y}$is the total weighted estimate of the variable (hours) for all usable units in the estimation stratum: $\hat{y}=\sum_{i=1}^{n} w_i y_i$

- N is the sum of final weights for usable units in the estimation stratum: $N=\sum_{i=1}^{n} w_i$
- n is the number of usable units reporting in the estimation stratum
- ow_i is the original state sampling weight of establishment *i* in the estimation stratum
- w_i is the final weight of establishment *i* in the estimation stratum
- x_i is the reported value of the variable (case counts for each case type) for establishment *i* in the estimation stratum
- $\bar{x}$ is the weighted average value of the variable (case counts for each case type, hours, employment) for all usable units in the estimation stratum: $\bar{x}=\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- y_i is the reported value of the variable (hours) for establishment *i* in the estimation stratum
- $\bar{y}$ is the weighted average value of the variable (hours) for all usable units in the estimation stratum:$\bar{y}=\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$

iii. Where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1, the means of the stratum, $\bar{x}$ and $\bar{y}$, cannot be used because it will yield an incorrect covariance of 0. To correct for this, use the means from the roll-up hierarchy:

- If there is only one usable unit in the estimation stratum and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1, calculate the variance as the variance of the roll-up level where the mean is the mean of the survey year, state, ownership, size class, and parent TEI of the usable units in this roll-up level. However, if the one usable unit in the estimation stratum is also the only usable unit for the survey year, state, ownership, size class, and parent TEI, continue to roll-up by parent TEI level as far as the industry domain level until the number of usable units is greater than 1.

- If the one usable unit in the estimation stratum is also the only usable unit for the survey year, state, ownership, size class, and industry domain, calculate the variance as the variance of the roll-up level where the mean is the mean of survey year, state, ownership, size class 0 (all sizes) and TEI. Continue to roll-up the parent TEI level as far as TEI000000 (all industries) until the number of usable units is greater than 1.

Therefore, **where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1**, similarly we use the roll-up level covariance as the covariance for the singe usable unit estimation stratum. The approximate covariance for the stratum with single usable unit is calculated by the following:

$$cov(\hat{x},\hat{y}) = \text{¿}\, cov(\hat{x}_{RU},\hat{y}_{RU}) = N^2 \frac{1}{n_{RU}}\text{¿}$$

$$\text{¿}\frac{N^2}{n_{RU}}\text{¿}$$

where

- $\hat{x}_{RU}$is the total weighted estimate of the variable (case counts for each case type) for all usable units in the roll-up level: $\hat{x}_{RU} = \sum_{i=1}^{n_{RU}} w_i x_i$
- $\hat{y}_{RU}$is the total weighted estimate of the variable (hours) for all usable units in the roll-up level: $\hat{y}_{RU} = \sum_{i=1}^{n_{RU}} w_i y_i$
- N is the sum of final weights for usable units in the estimation stratum: $N = \sum_{i=1}^{n} w_i$, in this case it is the single usable unit estimation stratum, so N= w_i
- N_{RU} is the sum of final weights for usable units in the roll-up level: $N_{RU} = \sum_{i=1}^{n_{RU}} w_i$
- n_{RU} is the number of usable units reporting in the roll-up level
- ow_i is the original state sampling weight of establishment *i* in the roll-up level
- w_i is the final weight of establishment *i* in the roll-up level
- x_i is the reported value of the variable (case counts for each case type) for establishment *i* in the roll-up level
- $\bar{x}_{RU}$ is the weighted average value of the variable (case counts for each case type, hours, employment) for all usable units in the roll-up level:

$$\bar{x}_{RU} = \frac{\sum_{i=1}^{n_{RU}} w_i x_i}{\sum_{i=1}^{n_{RU}} w_i}$$

- y_i is the reported value of the variable (hours) for establishment *i* in the roll-up level
- $\bar{y}_{RU}$ is the weighted average value of the variable (hours) for all usable units in the roll-up level:

$$\overline{y}_{RU}=\frac{\sum_{i=1}^{n_{RU}} w_i\, y_i}{\sum_{i=1}^{n_{RU}} w_i}$$

c. **Variance of Ratios:**

i. **Where the number of usable units in the estimation stratum and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) equal 1**, set the variance equal to 0.

ii. **Where the number of usable units in the estimation stratum is greater than 1**, calculate the **variance of the ratio** of total case counts for each case type (x) and hours (y) ($Rate=\frac{\hat{x}}{\hat{y}}$) using the following unbiased formula for weighted samples

$$Var\left(\frac{\hat{x}}{\hat{y}}\right)=\left(\frac{\hat{x}}{\hat{y}}\right)^2*\left(\frac{var(\hat{x})}{\hat{x}^2}+\frac{var(\hat{y})}{\hat{y}^2}-\frac{2*cov(\hat{x},\hat{y})}{\hat{x}*\hat{y}}\right)$$

where

- $\hat{x}$is the total weighted estimate of the variable (case counts for each case type) for all usable units in the estimation stratum: $\hat{x}=\sum_{i=1}^{n} w_i x_i$
- $\hat{y}$is the total weighted estimate of the variable (hours) for all usable units in the estimation stratum: $\hat{y}=\sum_{i=1}^{n} w_i\, y_i$
- $var(\hat{x})$ is the variance estimate of the variable in the estimation stratum (total case counts for each case type ($\hat{x}$))
- $var(\hat{y})$ is the variance estimate of the variable in the estimation stratum (total hours ($\hat{y}$))
- $cov(\hat{x},\hat{y})$ is the covariance estimate between two variables in the estimation stratum (case counts of each case type and hours)

If a standard rate per 100 employees (equivalent 200,000 employee hours per year) is reported ($SRate=\frac{\hat{x}}{\hat{y}}*200{,}000$), then the **variance of standard rate** for case types TRC, DART, DAFW, DJTR, ORC, and INJU is

$$var\left(SRate_1\right)=var\left(\frac{\hat{x}}{\hat{y}}*200{,}000\right)=\left(200{,}000\right)^2\left(\frac{\hat{x}}{\hat{y}}\right)^2*\left(\frac{var(\hat{x})}{\hat{x}^2}+\frac{var(\hat{y})}{\hat{y}^2}-\frac{2*cov(\hat{x},\hat{y})}{\hat{x}*\hat{y}}\right)$$

Where the standard rate per 10,000 employees (equivalent 20,000,000 employee hours per year) is reported ($SRate=\frac{\hat{x}}{\hat{y}}*20{,}000{,}000$), then the **variance of standard rate** for case types ILLN, SKIN, RESP, POIS, HEAR, OTHR is

$$var(SRate_2)=var\left(\frac{\hat{x}}{\hat{y}}*20{,}000{,}000\right)=(20{,}000{,}000)^2\left(\frac{\hat{x}}{\hat{y}}\right)^2*\left(\frac{var(\hat{x})}{\hat{x}^2}+\frac{var(\hat{y})}{\hat{y}^2}-\frac{2*cov(\hat{x},\hat{y})}{\hat{x}*\hat{y}}\right)$$

iii. **Where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling frame (which is the sum of final weights in this estimation stratum) does not equal 1,** calculate the **variance of the ratio** as the variance of total case counts for each case type (x) and hours (y) of the roll-up level used above, using the following unbiased formula for weighted samples

$$Var\left(\frac{\hat{x}}{\hat{y}}\right)=(200{,}000)^2\left(\frac{\hat{x}}{\hat{y}}\right)^2*\left(\frac{var(\hat{x}_{RU})}{\hat{x}^2}+\frac{var(\hat{y}_{RU})}{\hat{y}^2}-\frac{2*cov(\hat{x}_{RU},\hat{y}_{RU})}{\hat{x}*\hat{y}}\right)$$

where

- $\hat{x}$is the total weighted estimate of the variable (case counts for each case type) for all usable units in the estimation stratum: $\hat{x}=\sum_{i=1}^{n} w_i x_i$
- $\hat{y}$is the total weighted estimate of the variable (hours) for all usable units in the estimation stratum: $\hat{y}=\sum_{i=1}^{n} w_i y_i$
- $\hat{x}_{RU}$is the total weighted estimate of the variable (case counts for each case type) for all usable units in the roll-up level: $\hat{x}_{RU}=\sum_{i=1}^{n_{RU}} w_i x_i$
- $\hat{y}_{RU}$is the total weighted estimate of the variable (hours) for all usable units in the roll-up level: $\hat{y}_{RU}=\sum_{i=1}^{n_{RU}} w_i y_i$
- $var(\hat{x}_{RU})=$¿is the variance estimate of the variable (total case counts for each case type ($\widehat{x_{RU}}$)) for the roll-up level
- $var(\hat{y}_{RU})$ is the variance estimate of the variable (total hours ($\widehat{y_{RU}}$)) for the roll-up level
- $cov(\hat{x}_{RU},\hat{y}_{RU})$is the covariance estimate between two variables (case counts of each case type and hours) for the roll-up level

If a standard rate per 100 employees (equivalent 200,000 employee hours per year) is reported ($SRate=\frac{\hat{x}}{\hat{y}}*200{,}000$), then the **variance of standard rate** for the roll-up level for case types TRC, DART, DAFW, DJTR, ORC, and INJU is

$$var\left(SRate_1\right)=\left(200{,}000\right)^2\left(\frac{\hat{x}}{\hat{y}}\right)^2*\left(\frac{var\left(\hat{x}_{RU}\right)}{\hat{x}^2}+\frac{var\left(\hat{y}_{RU}\right)}{\hat{y}^2}-\frac{2*cov\left(\hat{x}_{RU},\hat{y}_{RU}\right)}{\hat{x}*\hat{y}}\right)$$

Where the standard rate per 10,000 employees (equivalent 20,000,000 employee hours per year) is reported ($SRate=\frac{\hat{x}}{\hat{y}}*20{,}000{,}000$), then the **variance of standard rate** for the roll-up level for case types ILLN, SKIN, RESP, POIS, HEAR, OTHR is

$$var\left(SRate_2\right)=\left(200{,}000{,}000\right)^2\left(\frac{\hat{x}}{\hat{y}}\right)^2*\left(\frac{var\left(\hat{x}_{RU}\right)}{\hat{x}^2}+\frac{var\left(\hat{y}_{RU}\right)}{\hat{y}^2}-\frac{2*cov\left(\hat{x}_{RU},\hat{y}_{RU}\right)}{\hat{x}*\hat{y}}\right)$$

d. Percent Relative Standard Error (RSE) Estimation:

Calculate the **percent relative standard error** using the following formulas

$$\%RSE\left(\hat{x}\right)=\frac{100*\sqrt[\square]{Var\left(\hat{x}\right)}}{\hat{x}} \quad \text{for totals}$$

where

- $\hat{x}$is the total estimate of the variable (case counts for each case type, hours, and employment) for all usable units in the stratum. Note that the estimate may be derived from a single usable unit or multiple usable units.
- $var\left(\hat{x}\right)$ is the variance estimate of the variable (case counts for each case type, hours, and employment). Note where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling stratum is greater than 1, the roll-up level variance was used to approximate the variance for the single unit stratum.

$$\%RSE\left(\frac{\hat{x}}{\hat{y}}\right)=\frac{\sqrt[\square]{var\left(SRate\right)}}{SRate}*100 \quad \text{for rates}$$

where

- $\hat{x}$is the total estimate of the variable (total case counts for each case type) for all usable units in the estimation stratum
- $\hat{y}$is the total estimate of the variable (total hours) for all usable units in the estimation stratum
- $var\left(SRate\right)$is the variance estimate of the standard rate. Note where the number of usable units in the estimation stratum equals 1 and the number of units in the sampling stratum is greater than 1, the roll-up level variance of the ratio was used to approximate the variance of the ratio for the single unit stratum.

- $SRate$ is the standard rate of case counts to hours depending on which case type is being used.

e. Estimation for domain levels

Estimation domains for both national and state estimates are combinations of year, state, ownership, TEI, and reported size class.

For variance estimates, calculate the estimate at the finest detail: survey year, state, ownership, TEI, and reported size class. The variances of *counts* of broader levels are calculated as the sum of the finer levels that comprise the broad level. For example, the variance of counts of the survey year, state, ownership, TEI (i.e., reported size class 0) is the sum of the variances of the survey year, state, ownership, TEI of reported size classes 1-5.

The variances of *rates* of broader levels cannot be calculated as the sum of variances of the finer levels that comprise the broad level. Instead, they are calculated using all the usable units in the estimation domain. For example, the variance of the rate of the survey year, state, ownership, TEI (i.e., reported size class 0) is calculated using all the usable units from size class 1-5.

If $\hat{x}$is equal to 0, then the variance, and covariance and RSE for $\hat{x}$and $\frac{\hat{x}}{\hat{y}}$ are equal to 0.

Estimates are based on the reported size class of the establishment. The number of usable units in an estimation stratum is based on the reported size class, not the size class used for sampling.

Data for the mining (NAICS 212) and railroad industries (NAICS 482) are obtained from MSHA an FRA, respectively, and are considered a census. Since these data are obtained from outside sources for which there is no ability to assess reliability, set the variance and covariance estimates for these industries to NULL/missing values before summarization for any estimation domain of interest. This means MHSA and FRA strata only contribute to point estimates (total and rate).

3. Inputs
   a. Microdata
      i. Summary case counts
   b. Final state and national summary weights (output from 8.1.5)
   c. List of Target Estimation Industries (TEIs)

4. Processing steps

   Calculate Variances and RSEs based on the formulas described in the definition for 14 for W weighted count estimates (results for the W Estimates for the twelve case types, hours, and employment are copied to the 14 T estimates) and 12 for R Estimates (for the twelve case types)

5. Outputs
   a. Variance (14 for W weighted count estimates (results for the W Estimates for the twelve case types, hours, and employment are copied to the 14 T estimates) and 12 for R Estimates (for the twelve case types))

    b. Percent Relative Standard Errors (14 for W weighted count estimates (results for the W Estimates for the twelve case types, hours, and employment are copied to the 14 T estimates) and 12 for R Estimates (for the twelve case types))

S-31323 SAS - Summary Variances (SV) and SV Relative Standard Errors (RSE) errors

As a NO User,

I need the system to have the ability to feed any errors and details relating to Summary Variances (SV) and SV Relative Standard Errors (RSE) calculation to the Error Monitoring GUI

So that errors in Estimation Runs can be view

Error checking for Variance and RSE:

1. Check that rate (R), weighted count (W), and weighted count in thousands (T), have RSEs. Census industries (railroad and mining) we don't calculate RSEs for. Also, we don't calculate RSE for quartile estimates.
    a. select count(*), estimate_type from soii.estimates where rse is not null and survey_year=XXXX group by estimate_type
    b. expect W, R, and T rows to have >0 for count
2. System should not calculate size class estimates for state estimates unless it is a sector level TEI (current system: LEVEL_CODE<3).
3. Very small variances should be set to zero. Set variances less than the absolute value of 0.000005 to zero.