METHODS USED TO CALCULATE THE VARIANCES OF THE OSHS CASE AND DEMOGRAPHIC ESTIMATES

FEBRUARY 22, 2002

INTRODUCTION

In an effort to reduce the computer time required to calculate the variances for the Case and Demographic estimates, we decided that the variances would be computed using models. The equations presented here may seem to be surprisingly simple given the complexity of the survey, but we are confident that the statistical basis for the use of these equations is strong.

This paper derives variance and relative standard error equations for the three types of Case and Demographic estimates:
>PROPORTION
>TOTAL
>RATIO

The paper also gives an example of the calculations for each of these types.

When applying these equations, it is important to use the appropriate DAFW case sample size. If, within an area, the estimates are for All Industries, the sample size is the total number of unweighted sample DAFW cases for the area. If the estimate is restricted to a particular industry within the area, the sample size is the number of unweighted sample DAFW cases for that industry within the area.

PROPORTION

In this type of estimator, for an area, we are estimating the proportion, $\hat{p}_{hi}$, of the total number of DAFW cases in industry i that have characteristic h. Under the assumption that the design effect is one and that the total DAFW case sample size for the industry is fixed, we can use the standard variance formula for simple random sampling without replacement to model the variances and relative standard errors for these proportions:

$$V(\hat{p}_{hi}) = \frac{\hat{p}_{hi}(1-\hat{p}_{hi})}{\hat{m}_i}\left(1-\frac{\hat{m}_i}{\hat{C}_i}\right)$$

$$R.S.E.(\hat{p}_{hi}) = \frac{\sqrt{V(\hat{p}_{hi})}}{\hat{p}_{hi}}$$

where for the area

$\hat{C}_i =$ estimate of the total number of DAFW cases for industry i

$\hat{x}_{hi} =$ total number of weighted sample DAFW cases with characteristic h for industry i

$\hat{m}_i =$ total number of unweighted sample DAFW cases for industry i

$$\hat{p}_{hi} = \frac{\hat{x}_{hi}}{\hat{C}_i}$$

TOTAL

In this type of estimator, for an area, we are estimating the number of DAFW cases , E_{hi} hat, in industry i that have characteristic h. Investigation into the micro data file led us to have a high degree of confidence that the estimates of the total number of cases for a particular group (such as SIC 17) and the proportion of the cases in the group with a particular characteristic (such as male) are statistically independent. This simplifies the variance calculation; for, if two random variables are independent, the variance of their product can be expressed in the following way (Quality Control and Industrial Statistics, Duncan, p. 104)

$$\sigma_{XY}{}^2 = \overline{X}^2 \sigma_Y{}^2 + \overline{Y}^2 \sigma_X{}^2 + \sigma_X{}^2 \sigma_Y{}^2$$

A further simplification of the variance calculation is possible because the design effect for a Case and Demographic estimate of proportion is approximately one. This means that, for estimating proportions, the stratified sample of DAFW cases is statistically equivalent to a sample random sample of cases with the same sample size.

Since: $\hat{E}_{hi} = \hat{C}_i \left(\dfrac{\hat{x}_{hi}}{\hat{C}_i} \right)$ and since $\hat{C}_i$ and $\dfrac{\hat{x}_{hi}}{\hat{C}_i}$ are statistically independent:

$$V(\hat{E}_{hi}) = \left(\left(\frac{\hat{x}_{hi}}{\hat{C}_i} \right)^2 V(\hat{C}_i) \right) + \left((\hat{C}_{i^2}) V\left(\frac{\hat{x}_{hi}}{\hat{C}_i} \right) \right) + \left(V(\hat{C}_i) V\left(\frac{\hat{x}_{hi}}{\hat{C}_i} \right) \right)$$

$$R.S.E.(\hat{E}_{hi}) = \frac{\sqrt{V(\hat{E}_{hi})}}{\hat{E}_{hi}}$$

where for the area

$\hat{E}_{hi} = \qquad$ estimate of the total number of DAFW cases with characteristic h for industry i

$\hat{C}_i = \qquad$ estimate of the total number of DAFW cases for industry i

$\hat{x}_{hi} = \qquad$ total number of weighted sample DAFW cases with characteristic h for industry i

$\hat{m}_i = \qquad$ total number of unweighted sample DAFW cases for industry i

$V(\hat{C}_i) = \qquad$ variance for the estimated number of DAFW cases in industry i
This value comes from the summary estimates.

$$V\left(\frac{\hat{x}_{hi}}{\hat{C}_i} \right) = \frac{\left(\dfrac{\hat{x}_{hi}}{\hat{C}_i} \right)\left(1 - \dfrac{\hat{x}_{hi}}{\hat{C}_{hi}} \right)}{\hat{m}_i} \left(1 - \frac{\hat{m}_i}{\hat{C}_i} \right)$$

= variance of the proportion of the cases with characteristic h for industry i

RATIO

In this type of estimator, for industry i in an area, we are estimating the ratio, R_{hki} hat, of the total number of DAFW cases that have characteristic h to the number of cases that that have both characteristic h and characteristic k. For example, the proportion of the total number of DAFW cases in an industry that are male that fall within a certain range of number of days lost. We can express this ratio of two totals as the quotient of two statistically independent proportions:

$$\hat{R}_{hki}=\frac{\hat{x}_{hki}}{\hat{x}_{hi}}=\frac{\dfrac{\hat{x}_{hki}}{\hat{C}_i}}{\dfrac{\hat{x}_{hi}}{\hat{C}_i}}=\frac{\hat{p}_{hki}}{\hat{p}_{hi}}$$

From Quality Control and Industrial Statistics, Duncan, p. 104, the following variance formula is a valid approximation for the ratio.

$$\sigma_{X/Y}^{2}=\frac{\overline{X}^{2}\sigma_{Y}^{2}+\overline{Y}^{2}\sigma_{X}^{2}}{\overline{Y}^{4}}$$

Therefore

$$V(\hat{R}_{hki})=\frac{(\hat{p}_{hki}^{2})V(\hat{p}_{hi})+(\hat{p}_{hi}^{2})V(\hat{p}_{hki})}{\hat{p}_{hi}^{4}}$$

$$R.S.E.(\hat{R}_{hki})=\frac{\sqrt{V(\hat{R}_{hki})}}{\hat{R}_{hki}}$$

where for the area

$\hat{C}_i=$ estimate of the total number of DAFW cases for industry i

$\hat{x}_{hi}=$ total number of weighted sample DAFW cases with characteristic h for industry

i

$\hat{x}_{hki}=$ total number of weighted sample DAFW cases with characteristics h and k for industry i

$\hat{m}_i=$ total number of unweighted sample DAFW cases for industry i

$$V(\hat{p}_{hi})=\frac{\hat{p}_{hi}(1-\hat{p}_{hi})}{\hat{m}_i}(1-\frac{\hat{m}_i}{\hat{C}_i})$$

$$V(\hat{p}_{hki})=\frac{\hat{p}_{hki}(1-\hat{p}_{hki})}{\hat{m}_i}(1-\frac{\hat{m}_i}{\hat{C}_i})$$

NUMERICAL EXAMPLE 1: PROPORTION

Here we are estimating the proportion of Delaware DAFW cases in SIC 17 that occurred to males. In this example industry i is SIC 17 and characteristic h is male.

$\hat{C}_i =$ estimated number of Delaware DAFW cases in SIC 17 = 318

$\hat{x}_{hi} =$ total number of weighted Delaware DAFW cases in SIC 17 that occurred to males = 299

$\hat{m}_i =$ total number of unweighted Delaware DAFW cases in SIC 17 = 189

$$\hat{p}_{hi} = \frac{\hat{x}_{hi}}{\hat{C}_i} = \frac{299}{318} = 0.940$$

Therefore:

$$V(\hat{p}_{hi}) = \frac{\hat{p}_{hi}(1-\hat{p}_{hi})}{\hat{m}_i}(1-\frac{\hat{m}_i}{\hat{C}_i}) = \frac{.940*.060}{189}*(1-\frac{189}{318}) = .000298*(1-.594) = 0.000121$$

$$R.S.E.(\hat{p}_{hi}) = \frac{\sqrt{V(\hat{p}_{hi})}}{\hat{p}_{hi}} = \frac{0.0110}{.940} = 0.0117$$

NUMERICAL EXAMPLE 2:   TOTAL

We are estimating the total number of Delaware DAFW cases that occurred to males; therefore, industry i is All Industries and characteristic h is male

$\hat{E}_{hi}=$ estimate of the total number of Delaware DAFW cases for males in All Industries = 3237

$\hat{C}_i=$ estimate of the total number of Delaware DAFW cases in All Industries = 5128

$\hat{x}_{hi}=$ total number of weighted sample Delaware DAFW cases for males in All Industries = 3237

$\hat{m}_i=$ total number of unweighted Delaware DAFW cases for All Industries = 2497

$V(\hat{C}_i)=$ variance for the estimated number of Delaware DAFW cases in All Industries = 12472
     This value comes from the summary estimates.

$$V\left(\frac{\hat{x}_{hi}}{\hat{C}_i}\right)=\frac{\left(\frac{\hat{x}_{hi}}{\hat{C}_i}\right)\left(1-\frac{\hat{x}_{hi}}{\hat{C}_{hi}}\right)}{\hat{m}_i}\left(1-\frac{\hat{m}_i}{\hat{C}_i}\right)=\frac{\left(\frac{3237}{5128}\right)\left(1-\frac{3237}{5128}\right)}{2497}\left(1-\frac{2497}{5128}\right)=0.0000478$$

Therefore:

$$V(\hat{E}_{hi})=\left(\left(\frac{\hat{x}_{hi}}{\hat{C}_i}\right)^2 V(\hat{C}_i)\right)+\left((\hat{C}_{i^2})V\left(\frac{\hat{x}_{hi}}{\hat{C}_i}\right)\right)+\left(V(\hat{C}_i)V\left(\frac{\hat{x}_{hi}}{\hat{C}_i}\right)\right)$$

$$=\left(\left(\frac{3237}{5128}\right)^2*12472\right)+\left(5128^2*.0000478\right)+\left(12472*.0000478\right)=6227$$

$$R.S.E.(\hat{E}_{hi})=\frac{\sqrt{V(\hat{E}_{hi})}}{\hat{E}_{hi}}=\frac{78.9}{3237}=0.0244$$

NUMERICAL EXAMPLE 3: RATIO

We are estimating the ratio of the number of Delaware DAFW cases that occurred to males that had 1 to 5 days away from work to the number of DAFW cases that occurred to males. In this example, industry i is All Industries, characteristic h is male, and characteristic k is 1 to 5 days away from work.

$\hat{C}_i =$ estimate of the total number of Delaware DAFW cases for All Industries = 5128

$\hat{x}_{hi} =$ total number of weighted sample Delaware DAFW cases for males in All Industries = 3237

$\hat{x}_{hki} =$ total number of weighted sample Delaware DAFW cases for males and 1 to 5 days away from work in All Industries = 1607

$\hat{m}_i =$ total number of unweighted sample Delaware DAFW cases for All Industries = 2497

$$\hat{p}_{hi} = \frac{\hat{x}_{hi}}{\hat{C}_i} = \frac{3237}{5128} = .631 \qquad\qquad \hat{p}_{hki} = \frac{\hat{x}_{hki}}{\hat{C}_i} = \frac{1607}{5128} = .313$$

$$V(\hat{p}_{hi}) = \frac{\hat{p}_{hi}(1-\hat{p}_{hi})}{m_i}(1-\frac{m_i}{C_i}) = \frac{.631(1-.631)}{2497}(1-\frac{2497}{5128}) = .0000478$$

$$V(\hat{p}_{hki}) = \frac{\hat{p}_{hki}(1-\hat{p}_{hki})}{m_i}(1-\frac{m_i}{C_i}) = \frac{.313(1-.313)}{2497}(1-\frac{2497}{5128}) = .0000442$$

$$\hat{R}_{hki} = \frac{\hat{x}_{hki}}{\hat{x}_{hi}} = \frac{\dfrac{\hat{x}_{hki}}{\hat{C}_i}}{\dfrac{\hat{x}_{hi}}{\hat{C}_i}} = \frac{\hat{p}_{hki}}{\hat{p}_{hi}} = \frac{.313}{.631} = 0.496$$

Therefore:

$$V(\hat{R}_{hki}) = \frac{(\hat{p}_{hki}^2)V(\hat{p}_{hi}) + (\hat{p}_{hi}^2)V(\hat{p}_{hki})}{\hat{p}_{hi}^4} = \frac{(.313^2 * .0000478) + (.631^2 * .0000442)}{.631^4} = .000141$$

$$R.S.E.(\hat{R}_{hki}) = \frac{\sqrt{V(\hat{R}_{hki})}}{\hat{R}_{hki}} = \frac{.0119}{.496} = .0240$$