**ESTIMATION METHOD 8**: Estimation of Variance of the Size-Weighted Cumulative Distribution Function for the Total of a Discrete Resource; Horvitz-Thompson Variance Estimator

## 1 Scope and Application

This method calculates the estimated variance of the estimated size-weighted cumulative distribution function (CDF) for the total of a discrete resource that has an indicator value equal to or less than a given indicator level. The size-weight is a measurement of the discrete resource such as area of a lake. There are two variance estimators presented in this method. An estimate can be produced for the entire population or for an arbitrary subpopulation with known or unknown size. This size is the size-weighted total in the subpopulation. The method applies to any probability sample and the variance estimate will be produced at the supplied indicator levels of interest. This method does not include estimators for the CDF. For information on CDF estimators, refer to Section 7.

## 2 Statistical Estimation Overview

A sample of size $n_a$ units is selected from subpopulation $a$ with known inclusion probabilities $\pi = \{\pi_1, \cdots, \pi_i, \cdots, \pi_{n_a}\}$, joint inclusion probabilities given by $\pi_{ij}$, where $i \neq j$, and size-weight values $\boldsymbol{w} = \{w_1, \cdots, w_i, \cdots, w_{n_a}\}$. The indicator is evaluated for each unit and represented by $\boldsymbol{y} = \{y_1, \cdots, y_i, \cdots, y_{n_a}\}$. The inclusion probabilities are design dependent and should be furnished with the design points. See Section 9 for further discussion.

The Horvitz-Thompson variance estimator of the size-weighted CDF for total, $\hat{V}\,[\hat{F}_a(x_k)]$, is calculated for each value of the indicator levels of interest, $x_k$. There are two Horvitz-Thompson variance estimators presented in this method. The first is a variance estimator of the Horvitz-Thompson estimator of a total. The second is a variance estimator of a Horvitz-Thompson ratio estimator. This variance estimator requires as input the CDF estimates produced using the Horvitz-Thompson ratio estimator of the size-weighted CDF for total, along with the known subpopulation size.

The output consists of the estimated variance values.

## 3 Conditions Under Which This Method Applies

- Probability sample with known inclusion probabilities and joint inclusion probabilities
- Discrete resource
- Arbitrary subpopulation
- All units sampled from the subpopulation must be accounted for before applying this method

## 4 Required Elements

4.1 Input Data

$y_i$ = value of the indicator for the $i^{th}$ unit sampled from subpopulation $a$.
$\pi_i$ = inclusion probability for selecting the $i^{th}$ unit of subpopulation $a$.
$\pi_{ij}$ = joint inclusion probability for selecting both the $i^{th}$ and $j^{th}$ units of subpopulation $a$.
$w_i$ = size-weight value for the $i^{th}$ unit sampled from subpopulation $a$.
$\hat{F}_a(x_k)$ = estimated size-weighted CDF (total) for indicator value $x_k$ in subpopulation $a$.

4.2 Additional Components

$n_a$ = number of units sampled from subpopulation $a$.
$x_k$ = $k^{th}$ indicator level of interest.
$W_a$ = subpopulation size (size-weighted total), if known.

## 5 Formulas and Definitions

The estimated variance of the estimated size-weighted CDF (total) for indicator value $x_k$ in subpopulation $a$, $\hat{V}\,[\hat{F}_a(x_k)]$, with known subpopulation size, $W_a$; Horvitz-Thompson variance estimator of the Horvitz-Thompson estimator of a CDF is

$$\hat{V}\,[\hat{F}_a(x_k)] = \sum_{i=1}^{n_a} w_i^2\, I(y_i \le x_k)\, \frac{1-\pi_i}{\pi_i^2} + \sum_{i=1}^{n_a} \sum_{j \ne i}^{n_a} w_i w_j\, I(y_i \le x_k)\, I(y_j \le x_k) \left( \frac{1}{\pi_i}\frac{1}{\pi_j} - \frac{1}{\pi_{ij}} \right) .$$

The estimated variance of the estimated size-weighted CDF (total) for indicator value $x_k$ in subpopulation $a$, $\hat{V}\,[\hat{F}_a(x_k)]$, with estimated subpopulation size, $\hat{W}_a$; Horvitz-Thompson variance estimator of the Horvitz-Thompson ratio estimator of a CDF is

$$\hat{V}\,[\hat{F}_a(x_k)] = \frac{\displaystyle\sum_{i=1}^{n_a} d_i^2\, \frac{1-\pi_i}{\pi_i^2} + \sum_{i=1}^{n_a} \sum_{j \ne i}^{n_a} d_i\, d_j \left( \frac{1}{\pi_i}\frac{1}{\pi_j} - \frac{1}{\pi_{ij}} \right)}{\hat{W}_a^2}\, W_a^2 ;$$

$$\hat{W}_a = \sum_{i=1}^{n_a} \frac{w_i}{\pi_i}, \quad d_i = w_i \left[ I(y_i \le x_k) - \frac{\hat{F}_a(x_k)}{W_a} \right], \quad d_j = w_j \left[ I(y_j \le x_k) - \frac{\hat{F}_a(x_k)}{W_a} \right].$$

For these equations:

$\hat{F}_a(x_k)$ = estimated size-weighted CDF (total) for indicator value $x_k$ in subpopulation $a$.

$$I(y_i \le x_k) = \begin{cases} 1, & y_i \le x_k \\ 0, & \text{otherwise} \end{cases}.$$

$x_k$ = $k^{th}$ indicator level of interest.
$y_i$ = value of the indicator for the $i^{th}$ unit sampled from subpopulation $a$.
$\pi_i$ = inclusion probability for selecting the $i^{th}$ unit of subpopulation $a$.
$\pi_{ij}$ = joint inclusion probability for selecting both the $i^{th}$ and $j^{th}$ units of subpopulation $a$.
$w_i$ = size-weight value for the $i^{th}$ unit sampled from subpopulation $a$.
$n_a$ = number of units sampled from subpopulation $a$.

## 6 Procedure

### 6.1 Enter Data

Input the sample data consisting of the indicator values, $y_i$ , and their associated inclusion probabilities, $\pi_i$ and size-weights, $w_i$. For example,

| Calcium $y_i$ | Inclusion Probability $\pi_i$ | Lake Area $w_i$ |
|---|---|---|
| 1.5992 | .07734 | 24.249 |
| 2.3707 | .00375 | 92.251 |
| 1.5992 | .75000 | 28.018 |
| 2.0000 | .75000 | 52.953 |
| 7.0000 | .00375 | 362.254 |
| 2.8196 | .02227 | 140.671 |
| 1.2204 | .01406 | 7.758 |
| 1.5992 | .03750 | 29.702 |
| 2.9399 | .00586 | 149.276 |
| .7395 | .00375 | 1.081 |

## 6.2 Sort Data

Sort the sample data in nondecreasing order based on the $y_i$ indicator values. Keep all occurrences of an indicator value to obtain correct results.

| Calcium $y_i$ | Inclusion Probability $\pi_i$ | Lake Area $w_i$ |
|---|---|---|
| .7395 | .00375 | 1.081 |
| 1.2204 | .01406 | 7.758 |
| 1.5992 | .07734 | 24.249 |
| 1.5992 | .75000 | 28.018 |
| 1.5992 | .03750 | 29.702 |
| 2.0000 | .75000 | 52.953 |
| 2.3707 | .00375 | 92.251 |
| 2.8196 | .02227 | 140.671 |
| 2.9399 | .00586 | 149.276 |
| 7.0000 | .00375 | 362.254 |

## 6.3 Compute or Input Joint Inclusion Probabilities

The required joint inclusion probabilities are in the following table. For this example, they were computed by the formula $\pi_{ij} = [2(n_a - 1)\pi_i\pi_j] / [2n_a - \pi_i - \pi_j]$ and are displayed in the following table.

| Joint Inclusion Probability $\pi_{ij} = \pi_{ji}$, $\pi_{ii} = \pi_i$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $j$ / $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | |
| 2 | .000047 | | | | | | | | |
| 3 | .000262 | .000983 | | | | | | | |
| 4 | .002630 | .009867 | .054457 | | | | | | |
| 5 | .000127 | .000476 | .002625 | .026350 | | | | | |
| 6 | .002630 | .009867 | .054457 | .547297 | .026350 | | | | |
| 7 | .000013 | .000047 | .000262 | .002630 | .000127 | .002630 | | | |
| 8 | .000075 | .000282 | .001558 | .015636 | .000754 | .015636 | .000075 | | |
| 9 | .000020 | .000074 | .000410 | .004111 | .000198 | .004111 | .000020 | .000118 | |
| 10 | .000013 | .000047 | .000262 | .002630 | .000127 | .002630 | .000013 | .000075 | .000020 |

## 6.4 Obtain Subpopulation Size

Input $W_a$ if using a known subpopulation size. $W_a = 156000$ for this dataset.

Calculate $\hat{W}_a$ from the sample data only if using the variance estimator of the Horvitz-Thompson ratio estimator of a CDF. Divide each $w_i$ by the inclusion probability, $\pi_i$ , for all units in the sample $a$. Sum each of these quantities to obtain $\hat{W}_a$ .

$\hat{W}_a = (1.081/.00375) + (7.758/.01406) + (24.249/.07734) + . . . + (362.254/.00375) = 155045.265$ for this data set.

## 6.5 Input Indicator Levels of Interest and Estimated CDF Values

For this example data, the variance of the empirical CDF is of interest; $x_k$ values = (.7395, 1.2204, 1.5992, 2, 2.3707, 2.8196, 2.9399, 7).

Input $\hat{F}_a(x_k)$ for each $x_k$ if the Horvitz-Thompson ratio estimator was used to estimate the CDF.

| Calcium $x_k$ | Size-Weighted CDF for Total, Ratio Estimator $\hat{F}_a(x_k)$ |
|---|---|
| .7395 | 290 |
| 1.2204 | 845 |
| 1.5992 | 1995 |
| 2.0000 | 2066 |
| 2.3707 | 26818 |
| 2.8196 | 33174 |
| 2.9399 | 58804 |
| 7.0000 | 156000 |

## 6.6 Compute Estimated Variance Values

Calculate $\hat{V}\ [\hat{F}_a(x_k)]$ for $x_k$ using the formulas from Section 5.

Compare each $y_i$ to $x_k$. Set $I(y_i \leq x_k) = 1$ if $y_i \leq x_k$. If this is not the case, set this term equal to zero.

Calculate the variance of the Horvitz-Thompson ratio estimator of the CDF by calculating the numerator portion of the equation that sums across all the $y_i$ data values. Multiply this quantity by $W_a^2 / \hat{W}_a^2$ to obtain the variance.

When the variance of the non-ratio form of the CDF estimator is used, the calculation is simpler. Sum across the $y_i$ data values until $y_i$ exceeds $x_k$ (when using sorted data) instead of across all the $y_i$ data values, because each additional term will contribute zero to the sum.

Do this for each $x_k$. Results for the example data are in Section 6.7. For the example using a known subpopulation size, $W_a = 156000$ is used.

6.7 Output Results

Output the indicator levels of interest and at least the associated estimated variance, $\hat{V}\,[\hat{F}_a(x_k)]$.

| Calcium<br><br>$x_k$ | Estimated Variance of Size-Weighted CDF for Total, Ratio Estimator (x 100,000)<br><br>$\hat{V}\,[\hat{F}_a(x_k)]$ | Estimated Variance of Size-Weighted CDF for Total, $W_a$ = 156000 (x 100,000)<br><br>$\hat{V}\,[\hat{F}_a(x_k)]$ |
|---:|---:|---:|
| .7395 | 1.33932 | .82786 |
| 1.2204 | 7.75618 | 3.47933 |
| 1.5992 | 31.28985 | 7.80689 |
| 2.0000 | 32.70403 | 7.63202 |
| 2.3707 | 7976.63932 | 5928.54471 |
| 2.8196 | 9540.63204 | 5950.33953 |
| 2.9399 | 20483.46898 | 10551.14618 |
| 7.0000 | 0 | 91052.68363 |

## 7 Associated Methods

An appropriate estimator for the estimated CDF may be found in Method 4 (Horvitz-Thompson Estimator).

## 8 Validation Data

Actual data with results, EMAP Design and Statistics Dataset #8, are available for comparing results from other versions of these algorithms.

## 9 Notes

Inclusion probabilities, $\pi_i$, and joint inclusion probabilities, $\pi_{ij}$, are determined by the design and should be furnished with the design points. In some instances, the joint inclusion probabilities may be calculated from a formula such as Overton's approximation where $\pi_{ij} = [2(n_a - 1)\pi_i\pi_j] \,/\, [2n_a - \pi_i - \pi_j]$, which is used in Section 6.3.

## 10 References


Cochran, W. G. 1977. *Sampling techniques*. 3rd Edition. New York: John Wiley & Sons.

Lesser, V. M., and W. S. Overton. 1994. *EMAP status estimation: Statistical procedures and algorithms*. EPA/620/R-94/008. Washington, DC: U.S. Environmental Protection Agency.

Overton, W. S., D. White, and D. L. Stevens Jr. 1990. *Design report for EMAP, Environmental Monitoring and Assessment Program*. EPA 600/3-91/053. Corvallis, OR: U.S. Environmental Protection Agency, Environmental Research Laboratory.

Särndal, C. E., B. Swensson, and J. Wretman, 1992. *Model assisted survey sampling*. New York: Springer-Verlag.