ESTIMATION METHOD 6: Estimation of Variance of the Cumulative Distribution Function for the Total Number of a Discrete Resource; Horvitz-Thompson Variance Estimator

1 Scope and Application

This method calculates the estimated variance of the estimated cumulative distribution function (CDF) for the total number of a discrete resource that has an indicator value equal to or less than a given indicator level. 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 total number of units 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 $\pi = \{\pi_1, \cdots, \pi_i, \cdots, \pi_{n_a}\}$ and joint inclusion probabilities given by π_{ij}, where $i \neq j$. The indicator is evaluated for each unit and represented by $y = \{y_1, \cdots, y_i, \cdots, y_{n_a}\}$. 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 CDF for total number, $\hat{V}\,[\hat{F}_a(x_k)]$, 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 CDF for total number, 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.
π_i = inclusion probability for selecting the i^{th} unit of subpopulation a.
π_{ij} = joint inclusion probability for selecting both the i^{th} and j^{th} units of subpopulation a.
$\hat{F}_a(x_k)$ = estimated CDF (total number) 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.
N_a = subpopulation size, if known.

5 Formulas and Definitions

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

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

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

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

$$\hat{N}_a = \sum_{i=1}^{n_a}\frac{1}{\pi_i}\;, \quad d_i = I(y_i \le x_k) - \frac{\hat{F}_a(x_k)}{N_a}\;, \quad d_j = I(y_j \le x_k) - \frac{\hat{F}_a(x_k)}{N_a}\;.$$

For these equations:

$\hat{F}_a(x_k)$ = estimated CDF (total number) 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.
π_i = inclusion probability for selecting the i^{th} unit of subpopulation a.
π_{ij} = joint inclusion probability for selecting both the i^{th} and j^{th} units of 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, π_i. For example,

Calcium y_i	Inclusion Probability π_i
1.5992	.07734
2.3707	.00375
1.5992	.75000
2.0000	.75000
7.0000	.00375
2.8196	.02227
1.2204	.01406
1.5992	.03750
2.9399	.00586
.7395	.00375

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 π_i
.7395	.00375
1.2204	.01406
1.5992	.07734
1.5992	.75000
1.5992	.03750
2.0000	.75000
2.3707	.00375
2.8196	.02227
2.9399	.00586
7.0000	.00375

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 N_a if using a known subpopulation size. N_a = 1130 for this dataset.

Calculate $\hat{N}_a$ from the sample data only if using the variance estimator of the Horvitz-Thompson ratio estimator of a CDF. Sum the reciprocals of the inclusion probabilities, π_i , for all units in the sample *a* to obtain $\hat{N}_a$.

$\hat{N}_a$ = (1/.00375) + (1/.01406) + (1/.07734) + . . . + (1/.00375) = 1128.939 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	CDF for Total Number, Ratio Estimator $\hat{F}_a(x_k)$
.7395	266.91
1.2204	338.10
1.5992	379.12
2.0000	380.36
2.3707	647.38
2.8196	692.24
2.9399	863.09
7.0000	1130

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 $I(y_i \le x_k) = 1$ if $y_i \le x_k$. 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 $N_a^2/\hat{N}_a^2$ 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, N_a = 1130.

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 x_k	Estimated Variance of CDF for Total Number, Ratio Estimator $\hat{V}\ [\hat{F}_a(x_k)]$	Estimated Variance of CDF for Total Number, $N_a = 1130$ $\hat{V}\ [\hat{F}_a(x_k)]$
.7395	57090	70845
1.2204	58744	71655
1.5992	59316	69463
2.0000	59334	69394
2.3707	67138	117938
2.8196	66666	113562
2.9399	57090	116513
7.0000	0	136623

7 Associated Methods

An appropriate estimator for the estimated CDF may be found in Method 2 (Horvitz-Thompson Estimator).

8 Validation Data

Actual data with results, EMAP Design and Statistics Dataset #6, are available for comparing results from other versions of these algorithms.

9 Notes

Inclusion probabilities, π_i, and joint inclusion probabilities, π_{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.