ESTIMATION METHOD 10: Estimation of Variance of the Cumulative Distribution Function for the Proportion of an Extensive Resource; Horvitz-Thompson Variance Estimator

1 Scope and Application

This method calculates the estimated variance of the estimated cumulative distribution function (CDF) for the proportion of an extensive 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 extent of the resource 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 densities $\pi = \{\pi_1, \cdots, \pi_i, \cdots, \pi_{n_a}\}$ and joint inclusion densities given by π_{ij}, where $i \neq j$. The indicator is evaluated for each unit and represented by $\boldsymbol{y} = \{y_1, \cdots, y_i, \cdots, y_{n_a}\}$. The inclusion densities 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, $\hat{V}\,[\hat{F}_a(x_k)]$, 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 proportion. The second is a variance estimator of a Horvitz-Thompson ratio estimator. The former estimator calculates the variance of the Horvitz-Thompson estimator of a total and divides this variance by the known subpopulation size squared, N_a^2. The latter estimator requires as input the CDF estimates produced using the Horvitz-Thompson ratio estimator of the CDF for proportion.

The output consists of the estimated variance values.

3 Conditions Under Which This Method Applies

- Probability sample with known inclusion densities and joint inclusion densities
- Extensive 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 density evaluated at the location of the i^{th} sample point in subpopulation a.

π_{ij} = joint inclusion density evaluated at the locations of the i^{th} and j^{th} sample points in subpopulation a.

$\hat{F}_a(x_k)$ = estimated CDF (proportion) 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 (proportion) 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

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

The estimated variance of the estimated CDF (proportion) 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

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

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

For these equations:

$\hat{F}_a(x_k)$ = estimated CDF (proportion) 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 density evaluated at the location of the i^{th} sample point in subpopulation a.

π_{ij} = joint inclusion density evaluated at the locations of the i^{th} and j^{th} sample points in 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 densities, π_i. For example,

Calcium y_i	Inclusion Density 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 Density 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 Densities

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

Joint Inclusion Density $\pi_{ij} = \pi_{ji}$, $\pi_{ii} = \pi_i$									
j / i	1	2	3	4	5	6	7	8	9
1									
2	.000047								
3	.000261	.000979							
4	.002531	.009491	.052205						
5	.000127	.000475	.002610	.025313					
6	.002531	.009491	.052205	.506250	.025313				
7	.000013	.000047	.000261	.002531	.000127	.002531			
8	.000075	.000282	.001550	.015032	.000752	.015032	.000075		
9	.000020	.000074	.000408	.003955	.000198	.003955	.000020	.000117	
10	.000013	.000047	.000261	.002531	.000127	.002531	.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 of the Horvitz-Thompson ratio estimator of a CDF. Sum the reciprocals of the inclusion densities, π_i, for all units in the sample a to obtain $\hat{N}_a$.

$\hat{N}_a = (1/.00375) + (1/.01406) + (1/.07734) + \ldots + (1/.00375) = 1128.939$ 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 Proportion, Ratio Estimator $\hat{F}_a(x_k)$
.7395	.2362
1.2204	.2992
1.5992	.3355
2.0000	.3366
2.3707	.5729
2.8196	.6126
2.9399	.7638
7.0000	1

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 numerator of the variance formula by summing across all the y_i data values. Divide by the applicable subpopulation size squared.

When the variance of the non-ratio form of the CDF estimator is used, the calculation may be simplified. 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. Divide this sum by the subpopulation size squared.

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 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 x_k	Estimated Variance of CDF for Proportion, Ratio Estimator $\hat{V}\,[\hat{F}_a(x_k)]$	Estimated Variance of CDF for Proportion, $N_a = 1130$ $\hat{V}\,[\hat{F}_a(x_k)]$
.7395	.044888	.055690
1.2204	.046211	.056351
1.5992	.046672	.054565
2.0000	.046687	.054479
2.3707	.052820	.092531
2.8196	.052442	.089057
2.9399	.044888	.091322
7.0000	0	.106996

7 Associated Methods

An appropriate estimator for the estimated CDF for extensive resources may be found in Method 1 (Horvitz-Thompson Estimator).

8 Validation Data

Actual data with results, EMAP Design and Statistics Dataset #10, are available for comparing results from other versions of these algorithms.

9 Notes

Inclusion densities, π_i, and joint inclusion densities, π_{ij}, are determined by the design and should be furnished with the design points. In some instances, the joint inclusion densities may be calculated from a formula that uses the location of the design points or they may be approximated by a formula that assumes simple random sampling. This simple random sampling formula, $\pi_{ij} = (n_a - 1)\pi_i\pi_j / n_a$, is used in Section 6.3.

10 References

Cordy, C. B. 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters* 18:353–362.

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.

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