ESTIMATION METHOD 12: Estimation of Variance of the Cumulative Distribution Function for the Proportion of a Discrete or an Extensive Resource; Yates-Grundy Variance Estimator

1 Scope and Application

This method calculates the estimated variance of the estimated cumulative distribution function (CDF) for the proportion of a discrete or 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 population with known or unknown size. In the discrete case, this size is the number of units in the population. In the extensive case, this size is the population extent. The method applies to any probability sample with fixed sample size 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 units is selected from population a with known inclusion probabilities $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_i, \cdots, \pi_n\}$ and joint inclusion probabilities given by π_{ij}, where $i \neq j$. The indicator is evaluated for each unit and represented by $\boldsymbol{y} = \{y_1, \cdots, y_i, \cdots, y_n\}$. When sampling an extensive resource, the inclusion probabilities are replaced by the inclusion density function evaluated at the sample locations. The inclusion probabilities are design dependent and should be furnished with the design points. See Section 9 for further discussion.

The Yates-Grundy 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 Yates-Grundy 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 population size squared, N^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 probabilities (or densities) and joint inclusion probabilities (or densities)
- Discrete or Extensive resource
- Arbitrary population
- All units sampled from the population 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 population a.

π_i = For discrete resources, the inclusion probability for selecting the i^{th} unit of population a. For extensive resources, the inclusion density evaluated at the location of the i^{th} sample point in population a.

π_{ij} = For discrete resources, the inclusion probability for selecting both the i^{th} and j^{th} units of population a. For extensive resources, the inclusion density evaluated at the locations of the i^{th} and j^{th} sample points in population a.

$\hat{F}_a(x_k)$ = estimated CDF (proportion) for indicator value x_k in population a.

4.2 Additional Components

n = number of units sampled from population a.

x_k = k^{th} indicator level of interest.

N = population size, if known.

5 Formulas and Definitions

The estimated variance of the estimated CDF (proportion) for indicator value x_k in population a, $\hat{V}[\hat{F}_a(x_k)]$, with known population size, N; Yates-Grundy variance estimator of the Horvitz-Thompson estimator of a CDF is

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

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

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

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

For these equations:

$\hat{F}_a(x_k)$ = estimated CDF (proportion) for indicator value x_k in population 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 population a.

π_i = For discrete resources, the inclusion probability for selecting the i^{th} unit of population a. For extensive resources, the inclusion density evaluated at the location of the i^{th} sample point in population a.

π_{ij} = For discrete resources, the inclusion probability for selecting both the i^{th} and j^{th} units of population a. For extensive resources, the inclusion density evaluated at the locations of the i^{th} and j^{th} sample points in population a.

n = number of units sampled from population a.

6 Procedure

6.1 Enter Data

Input the sample data consisting of the indicator values, y_i, and their associated inclusion probabilities (or densities), 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 (or Densities)

The required joint inclusion probabilities 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 Probability $\pi_{ij} = \pi_{ji}$, $\pi_{ii} = \pi_i$									
i \ *j*	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 Population Size

Input *N* if using a known population size. $N = 1130$ for this data set.

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

$\hat{N}$ = (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 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 $I(y_i \le x_k) = 1$ if $y_i \le x_k$. If this is not the case, set this term equal to zero.

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

Do this for each x_k. Results for the example data are in Section 6.7. For the example using a known population size, $N = 1130$ 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, $N = 1130$ $\hat{V}\,[\hat{F}_a(x_k)]$
.7395	.044710	.055482
1.2204	.046005	.056116
1.5992	.046453	.054400
2.0000	.046467	.054346
2.3707	.052579	.092363
2.8196	.052209	.088936
2.9399	.044710	.091247
7.0000	0	.106996

7 Associated Methods

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

8 Validation Data

Actual data with results, EMAP Design and Statistics Dataset #12, are available for comparing results from other versions of these algorithms.

9 Notes

Inclusion probabilities (or densities), π_i, and joint inclusion probabilities (or densities), 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. 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 the formula $\pi_{ij} = (n - 1)\pi_i\pi_j / n$ that assumes simple random sampling.

10 References

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

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.

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.

Stevens, Jr., D. L. 1995. A family of designs for sampling continuous spatial populations. *Environmetrics*. Submitted.