ESTIMATION METHOD 13: Simplified Variance of the Cumulative Distribution Function for Proportion (Discrete or Extensive) and for Total Number of a Discrete Resource, and Variance of the Size-Weighted Cumulative Distribution Function for Proportion and Total of a Discrete Resource; Simple Random Sample Variance Estimator

1 Scope and Application

This method calculates the estimated variance of the estimated cumulative distribution function (CDF) for the proportion and total number of a discrete (or in the case of proportion, extensive) resource that has an indicator value equal to or less than a given indicator level. The method also calculates the estimated size-weighted versions of these CDFs for a discrete resource. All of these CDFs are produced using Horvitz-Thompson estimators found in other methods. An estimate can be produced for the entire population or for a geographic subpopulation with unknown size. This size is the total number of units or extent in the subpopulation.

The estimation algorithms have been simplified for use in spreadsheet software such as Lotus 1-2-3 and Quattro Pro; however, because of this simplification, the use of these variability estimates is restricted. This method provides a mechanism for generating quick summaries of indicators to assist in internal research and is distributed with the restriction that results for inclusion in peer-reviewed documents or EPA reports should be cleared by EMAP statisticians. The variance estimates will be produced at the supplied indicator levels of interest. For information on the Horvitz-Thompson estimators of the CDF, refer to Section 7.

2 Statistical Estimation Overview

A sample of size n_a units is selected from subpopulation a with known inclusion probabilities $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_i, \cdots, \pi_{n_a}\}$ and if applicable, the size-weight values $\boldsymbol{w} = \{w_1, \cdots, w_i, \cdots, w_{n_a}\}$. The indicator is evaluated for each unit and represented by $\boldsymbol{y} = \{y_1, \cdots, y_i, \cdots, y_{n_a}\}$. 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 variance estimators of the CDF are calculated for each value of the indicator levels of interest, x_k. The units are assumed to come from an independent sampling design that reduces the usually required joint inclusion probabilities given by π_{ij}, where $i \neq j$, to $\pi_{ij} = (n_a - 1)\pi_i\pi_j / n_a$. Under the independent random sampling model, the Horvitz-Thompson variance estimator simplifies to the usual simple random sampling variance estimator, s^2, applied to a cumulative total. This total differs depending upon whether the CDF is for proportion or for total number. In the case of proportion, the Horvitz-Thompson ratio estimator is used to calculate the CDF and because both the numerator of the proportion are estimated, there is more variability in the estimate. As a result, the variance estimators of the CDF for proportion and the size-weighted CDF for proportion require as input the CDF estimates produced using the Horvitz-Thompson ratio estimator.

The output consists of the estimated variance values.

3 Conditions Under Which This Method Applies

- Independent random sample (IRS) with fixed sample size and known inclusion probabilities or densities
- Discrete resource (or extensive, in the case of proportion)
- Subpopulation is defined geographically, or the number of sites within the subpopulation of interest is known; examples: by ecoregion or first order stream length
- All units sampled from the subpopulation must be accounted for before applying this method; Missing values are excluded

3.1 Restrictions

Variability estimates of the CDF for non-geographic subpopulation estimates cannot be made using the supplied calculation routines. For example, the supplied routine does not apply for the estimates of variability of the percentage of lakes that are hypereutrophic and have ANC < 200, or the estimated number of streams containing a species of fish for a subset of the sample that is determined by a chemistry response. A more sophisticated variance estimator is needed for these cases; contact EMAP Design and Statistics for assistance.

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.
w_i = size-weight value for the i^{th} unit sampled from subpopulation a. This applies to discrete resources only. An example would be area of a lake.

4.2 Additional Components

n_a = number of units sampled from subpopulation a.
x_k = k^{th} indicator level of interest.

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

For the estimated variance of the estimated CDF for proportion, also input

$\hat{F}_a(x_k) = \dfrac{\sum_{i=1}^{n_a} \frac{1}{\pi_i} I(y_i \le x_k)}{\hat{N}_a}$, the estimated CDF for proportion for indicator value x_k in subpopulation a with estimated subpopulation size, $\hat{N}_a = \sum_{i=1}^{n_a} \frac{1}{\pi_i}$.

For the estimated variance of the estimated size-weighted CDF for proportion, also input $\hat{G}_a(x_k) = \dfrac{\sum_{i=1}^{n_a} \frac{w_i}{\pi_i} I(y_i \le x_k)}{\hat{W}_a}$, the estimated size-weighted CDF for proportion for indicator value x_k in subpopulation a with estimated subpopulation size-weighted total, $\hat{W}_a = \sum_{i=1}^{n_a} \frac{w_i}{\pi_i}$.

5 Formulas and Definitions

The estimated variance of the estimated CDF (proportion) for indicator value x_k in subpopulation a, $\hat{V}\,[\hat{F}_a(x_k)]$; Simple random sample variance estimator of the Horvitz-Thompson ratio estimator of a CDF is

$$\hat{V}\,[\hat{F}_a(x_k)] = \frac{n_a\, s^2}{\hat{N}_a^{\,2}}; \quad s^2 = \frac{\sum_{i=1}^{n_a} \left(t_i - \bar{t} \right)}{n_a - 1}; \quad t_i = \left(I(y_i \le x_k) - \hat{F}_a(x_k) \right) \frac{1}{\pi_i}.$$

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

$$\hat{V}\,[\hat{N}_a\, \hat{F}_a(x_k)] = n_a\, s^2; \quad s^2 = \frac{\sum_{i=1}^{n_a} \left(t_i - \bar{t} \right)}{n_a - 1}; \quad t_i = I(y_i \le x_k) \frac{1}{\pi_i}.$$

The estimated variance of the estimated size-weighted CDF (proportion) for indicator value x_k in subpopulation a, $\hat{V}\,[\hat{G}_a(x_k)]$; Simple random sample variance estimator of the Horvitz-Thompson ratio estimator of a CDF is

$$\hat{V}\,[\hat{G}_a(x_k)] = \frac{n_a\, s^2}{\hat{W}_a^{\,2}}; \quad s^2 = \frac{\sum_{i=1}^{n_a} \left(t_i - \bar{t} \right)}{n_a - 1}; \quad t_i = \left(I(y_i \le x_k) - \hat{G}_a(x_k) \right) \frac{w_i}{\pi_i}.$$

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

$$\hat{V}\,[\hat{W}_a\, \hat{G}_a(x_k)] = n_a s^2; \quad s^2 = \frac{\sum_{i=1}^{n_a} \left(t_i - \bar{t} \right)}{n_a - 1}; \quad t_i = I(y_i \le x_k) \frac{w_i}{\pi_i}.$$

For these equations:

$\hat{W}_a$ = estimated subpopulation size-weighted total.

$\hat{N}_a$ = estimated subpopulation size.

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

$\hat{G}_a(x_k)$ = estimated size-weighted 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 probability for selecting the i^{th} unit of subpopulation a.

w_i = size-weight value for the i^{th} unit sampled from subpopulation a. This applies to discrete resources only. An example would be area of a lake.

s^2 = sample variance of t.

n_a = number of units sampled from subpopulation a.

$\bar{t} = \dfrac{\sum_{i=1}^{n_a} t_i}{n_a}$, the arithmetic mean of t.

6 Procedure

6.1 Enter Data

Input the sample data consisting of the indicator values, y_i , their associated inclusion probabilities, π_i, and if applicable, the size-weight values, w_i and CDF estimates. For this example data, the variance of the empirical CDF is of interest; x_k values are equal to y_i .

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. Our sample data is

Indicator y_i (1)	Inclusion Probability π_i (2)	Size-weight (ex. area) w_i (3)	$\hat{F}_a(x_k)$ (4)	$\hat{G}_a(x_k)$ (5)
1.9	.042201	117.85	.1219	.0945
6.0	.059245	147.30	.2087	.1786
9.8	.023847	185.55	.4244	.4418
10.9	.060562	55.55	.5093	.4728
11.0	.037023	239.91	.6482	.6920
11.8	.055115	165.09	.7415	.7933
12.0	.102785	129.83	.7916	.8360
12.3	.059545	51.42	.8779	.8652
13.6	.084789	262.33	.9386	.9699
14.2	.083752	74.58	1.0000	1.0000

$$\hat{N}_a = \sum_{i=1}^{n_a} \frac{1}{\pi_i} = 194.43 \; ; \quad \hat{W}_a = \sum_{i=1}^{n_a} \frac{w_i}{\pi_i} = 29563.60 \; ; \quad n_a = 10.$$

6.3 Compute estimated variance of the estimated CDF for proportion, $\hat{V}\,[\hat{F}_a(x_k)]$, and for total number, $\hat{V}\,[\hat{N}_a\,\hat{F}_a(x_k)]$.

Calculate the variance estimates for $x_k = 1.9$, by completing the following steps.

Create a new table of 6 columns using columns (1) and (2) from the table in Section 6.2. Insert the $I(y_i \le x_k)$ values into column (3) by setting $I(y_i \le x_k) = 1$ if $y_i \le 1.9$. If this is not the case, set $I(y_i \le x_k) = 0$. In column (4), insert the difference between column (3) and the CDF value corresponding to $x_k = 1.9$. This CDF value is .1219, from the table in Section 6.2. In column (5), put the result from dividing column (4) by column (2). In column (6), put the result from dividing column (3) by column (2). Results are in the following table; $I(y_i \le 1.9)$ is abbreviated as $I(1.9)$.

Indicator y_i (1)	Inclusion Probability π_i (2)	$I(1.9)$ (3)	$I(1.9) - \hat{F}_a(1.9)$ (4) = (3) − .2087	$\frac{I(1.9) - \hat{F}_a(1.9)}{\pi_i}$ (5) = (4) ÷ (2)	$\frac{I(1.9)}{\pi_i}$ (6) = (3) ÷ (2)
1.9	.042201	1	.8781	20.8	23.696
6.0	.059245	0	-.1219	-2.1	0
9.8	.023847	0	-.1219	-5.1	0
10.9	.060562	0	-.1219	-2.0	0
11.0	.037023	0	-.1219	-3.3	0
11.8	.055115	0	-.1219	-2.2	0
12.0	.102785	0	-.1219	-1.2	0
12.3	.059545	0	-.1219	-2.0	0
13.6	.084789	0	-.1219	-1.4	0
14.2	.083752	0	-.1219	-1.5	0

For estimating the variance of the estimated CDF for proportion for $x_k = 1.9$, calculate the sample variance, s^2, of column (5). (In Excel, use the VAR( ) function). $s^2 = 54.765$.

$$\hat{V}\,[\hat{F}_a(1.9)] = \hat{V}\,[.1219] = \frac{n_a s^2}{\hat{N}_a^{\,2}} = \frac{(10)(54.765)}{194.43^2} = 0.0145.$$

For estimating the variance of the estimated CDF for total number for $x_k = 1.9$, calculate the sample variance, s^2, of column (6). $s^2 = 56.15$.

$$\hat{V}\,[\hat{N}_a\hat{F}_a(1.9)] = \hat{V}\,[23.70] = n_a s^2 = (10)(56.15) = 561.5.$$

Do this same procedure for the next x_k value, x_k = 6.0. The table now becomes

Indicator y_i	Inclusion Probability π_i	$I(6.0)$	$I(6.0) - \hat{F}_a(6.0)$	$\frac{I(6.0) - \hat{F}_a(6.0)}{\pi_i}$	$\frac{I(6.0)}{\pi_i}$
(1)	(2)	(3)	(4) = (3) – .2087	(5) = (4) ÷ (2)	(6) = (3) ÷ (2)
1.9	.042201	1	.7913	18.751	23.696
6.0	.059245	1	.7913	13.357	16.879
9.8	.023847	0	-.2087	-8.751	0
10.9	.060562	0	-.2087	-3.446	0
11.0	.037023	0	-.2087	-5.637	0
11.8	.055115	0	-.2087	-3.786	0
12.0	.102785	0	-.2087	-2.030	0
12.3	.059545	0	-.2087	-3.505	0
13.6	.084789	0	-.2087	-2.461	0
14.2	.083752	0	-.2087	-2.492	0

For estimating the variance of the estimated CDF for proportion for x_k = 6.0, calculate the sample variance, s^2, of column (5). s^2 = 77.026.

$$\hat{V}\,[\hat{F}_a(6.0)] = \hat{V}\,[.2087] = \frac{n_a s^2}{\hat{N}_a^{\,2}} = \frac{(10)(77.026)}{194.43^2} = 0.0204.$$

For estimating the variance of the estimated CDF for total number for x_k = 6.0, calculate the sample variance, s^2, of column (6). s^2 = 75.75.

$$\hat{V}\,[\hat{N}_a\hat{F}_a(6.0)] = \hat{V}\,[40.58] = n_a s^2 = (10)(75.75) = 757.5.$$

Repeat this process for the remaining x_k values.

6.4 Compute estimated variance of the estimated size-weighted CDF for proportion, $\hat{V}\,[\hat{G}_a(x_k)]$, and for total, $\hat{V}\,[\hat{W}_a\hat{G}_a(x_k)]$.

The procedure for calculating the variance estimates for the size-weighted CDFs is the same as the one used in Section 6.3. The only difference between the estimates is that w_i/π_i is substituted for $1/\pi_i$ in every part of the calculation. The following example is for x_k = 6.0.

Create a new table of 6 columns. Use column (1) from the table in Section 6.2 for the first column. In the second column, enter the result from dividing column (3) by column (2) of the table in Section 6.2. Insert the $I(y_i \le x_k)$ values into column (3) by setting $I(y_i \le x_k) = 1$ if $y_i \le 6.0$. If this is not the case, set $I(y_i \le x_k) = 0$. In column (4), insert the difference between column (3) and the size-weighted CDF value corresponding to $x_k = 6.0$. This CDF value is .1786, from the table in Section 6.2. In column (5), put the result from multiplying column (4) by column (2). In column (6), put the result from multiplying column (3) by column (2). Results are in the following table; $I(y_i \le 6.0)$ is abbreviated as $I(6.0)$.

Indicator y_i (1)	w_i/π_i (2)	$I(6.0)$ (3)	$I(6.0) - \hat{G}_a(6.0)$ (4) = (3) − .1786	$[I(6.0) - \hat{G}_a(6.0)]\frac{w_i}{\pi_i}$ (5) = (4) × (2)	$[I(6.0)]\frac{w_i}{\pi_i}$ (6) = (3)×(2)
1.9	2792.5879	1	.8214	2293.8317	2792.5879
6.0	2486.2858	1	.8214	2042.2351	2486.2858
9.8	7780.8529	0	-.1786	-1389.6603	0
10.9	917.2418	0	-.1786	-163.8194	0
11.0	6480.0259	0	-.1786	-1157.3326	0
11.8	2995.3733	0	-.1786	-534.9737	0
12.0	1263.1221	0	-.1786	-225.5936	0
12.3	863.5486	0	-.1786	-154.2298	0
13.6	3093.9155	0	-.1786	-552.5733	0
14.2	890.4862	0	-.1786	-159.0408	0

For estimating the variance of the estimated size-weighted CDF for proportion for $x_k = 6.0$, calculate the sample variance, s^2, of column (5). $s^2 = 1491256.3$.

$$\hat{V}\,[\hat{G}_a(6.0)] = \hat{V}\,[.1786] = \frac{n_a\, s^2}{\hat{W}_a^{\,2}} = \frac{(10)(1491256.3)}{29563.60^2} = 0.0171.$$

For estimating the variance of the estimated size-weighted CDF for total number for $x_k = 6.0$, calculate the sample variance, s^2, of column (6). $s^2 = 1243723.68$.

$$\hat{V}\,[\hat{W}_a\,\hat{G}_a(6.0)] = \hat{V}\,[5280.06] = n_a\, s^2 = (10)(1243723.68) = 12{,}437{,}237.$$

Repeat this process for the remaining x_k values.

7 Associated Methods

An appropriate estimator for the estimated CDF for proportion for discrete or extensive resources may be found in Method 1 (Horvitz-Thompson Estimator). For the estimated CDF for total number, size-weighted CDF for proportion, and size-weighted CDF for total (which apply only to discrete resources), see Methods 2, 3, and 4, respectively.

8 Validation Data

Actual data with results, EMAP Design and Statistics Dataset #13, are available for comparing results from other versions of these algorithms.

9 Notes

Inclusion probabilities, π_i, are determined by the design and should be furnished with the design points.

Population estimates are calculated using inclusion probabilities or densities and differ by indicator. For example, in the 1993 stream sample, periphyton, and full physical habitat (P-hab) were measured only on the 1X grid streams, requiring use of the 1X inclusion probabilities. Water chemistry measurements were taken on both 1X and 7X streams, and in this case, the 7X inclusion probabilities should be used. Reference/test sites (both lakes and streams) were hand picked, and can not be used to make population estimates. These restrictions apply to all sampling years.

If estimates across multiple years are required, responses for sites sampled in multiple years should only be included for the initial year and the inclusion probabilities should be multiplied by the number of years of data.

10 References

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.

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