ESTIMATION METHOD 3: Estimation of the Size-Weighted Cumulative Distribution Function for Proportion of a Discrete Resource; Horvitz-Thompson Estimator, Normal Approximation

1 Scope and Application

This method calculates the estimate of the size-weighted cumulative distribution function (CDF) for the proportion of a discrete resource that has an indicator value equal to or less than a given indicator level. The size-weight value is a measurement of the discrete resource such as area of a lake. The method applies to any probability sample and presents two estimators. An estimate can be produced for the entire population or for an arbitrary subpopulation with known or unknown size, where this size is the size-weighted total in the subpopulation. Suggestions for estimating the CDF over the range of the indicator are included. Alternatively, the CDF can be calculated at the indicator levels found in the probability sample. The method uses the Normal approximation to provide confidence bounds or intervals for the true cumulative distribution function. This method does not include variance estimators for the estimated CDF. For information on appropriate variance estimators, refer to Section 7.

This method has been applied in:

The 1991 Surface Waters Pilot Report

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 size-weight values $\boldsymbol{w} = \{w_1, \cdots, w_i, \cdots, w_{n_a}\}$. The indicator is evaluated for each unit and represented by $\boldsymbol{y} = \{y_1, \cdots, y_i, \cdots, y_{n_a}\}$.

Estimates of the cumulative distribution function are obtained for the indicator levels of interest, $\boldsymbol{x} = \{x_1, \cdots, x_k, \cdots, x_m\}$. Several alternatives are available for choosing $\boldsymbol{x}$. The recommended alternative is the use of equally spaced values across the range of the indicator. Ideally, this range is known *a priori* and extends beyond the range of any particular data set. A second alternative is to use the set of unique values in the data set. This alternative gives the classical empirical cumulative distribution function. A third alternative is to use the midpoints of adjacent ordered values in $\boldsymbol{y}$ for the levels $\boldsymbol{x}$.

To obtain the estimated size-weighted cumulative distribution function, $\hat{F}_a(x_k)$, the Horvitz-Thompson estimator of a cumulative total is calculated for each x_k by summing up the number of indicators which are less than or equal to the x_k value, weighted by the size-weight values w_i. This total is then divided by the subpopulation size (size-weighted total), W_a. When this subpopulation size is unknown, the estimated subpopulation size, $\hat{W}_a$, is substituted for the known subpopulation size, W_a, to form the Horvitz-Thompson ratio estimator.

The Horvitz-Thompson ratio estimator may perform better than the estimator using the known subpopulation size, W_a , and may be used even if the subpopulation size is known. Some of the conditions under which this ratio estimator is recommended are given in Section 9. This ratio estimator should always be used in the case of missing data.

Confidence limits for $F_a(x_k)$ are produced by assuming a Normal distribution. These limits may be used to construct either a lower confidence bound, an upper confidence bound, or a confidence interval for $F_a(x_k)$. Computation of these limits requires an estimated variance of $\hat{F}_a(x_k)$ which is not provided in this method. Details for computing a suitable estimated variance of $\hat{F}_a(x_k)$ are found in other methods referenced in Section 7.

The output consists of the estimated cumulative distribution function values with either a one-sided confidence bound (upper or lower) or a confidence interval for $F_a(x_k)$.

3 Conditions Under Which This Method Applies

- Probability sample with known inclusion probabilities
- Discrete resource
- Arbitrary subpopulation
- All units sampled from the subpopulation must be accounted for before applying this method
- When the indicator value is missing, exclude this missing value and the corresponding inclusion probability and size-weight; use the Horvitz-Thompson ratio estimator

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.

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.

4.3 Graphical Display Considerations

Two issues should be resolved before graphing the CDF: 1) how many points to use and 2) what are the first and last points on the plot. The following are guidelines for the three alternatives mentioned in Section 2. In all three approaches, the plotted points are connected by line segments. Confidence limits are not to exceed one or to drop below zero. The sample $\boldsymbol{y}$ is understood to be in ascending order for this discussion.

If the empirical CDF is chosen, the number of points plotted is at most n_a+2. The first plotted point is (0,0) when the indicator takes on only positive values. Otherwise, choose a point smaller than y_1 as the abscissa and assign zero as the ordinate. Choose a point larger than y_{n_a} and assign 1 as its ordinate. Where there is more than one occurrence of an indicator level in the data set, plot only one point using the largest cumulative distribution function value associated with this level as the ordinate.

If the midpoints of adjacent values in $\boldsymbol{y}$ are used for the levels $\boldsymbol{x}$, at most n_a+1 points are plotted. To determine the first plotted point, calculate the distance between y_1 and y_2. Take half this distance and subtract it from y_1 to obtain the abscissa. If this abscissa is a negative number and the indicator can never be negative, instead assign zero as the abscissa. Use zero as the ordinate. Similarly, to determine the last plotted point, calculate the distance between the largest $\boldsymbol{y}$ values, y_{n_a-1} and y_{n_a}. Halve this distance, add it to y_{n_a} and plot this abscissa using 1 as the ordinate.

The recommended approach uses equally spaced levels across the potential range of the indicator. The levels used should be potential real values that the indicator could attain. In this case of discrete data, integer values should be used. As mentioned previously, ideally this range is known *a priori* and extends beyond the range of any particular data set. If an informed guess cannot be made for this range, one suggested range would be to use the midpoint approach for obtaining the first and last plot points as explained in the previous paragraph. How many points to use is a subjective decision and should take into account the chosen range, the size of the data set, and sometimes the data distribution itself must be examined. The following suggestions are given to help decide how many points to use.

In most cases, using the same number of points as used in the empirical distribution, n_a+2 points, will be sufficient for plotting the CDF. Extreme outliers in a particular data set may have a great influence on the graph. In this case, more points may be needed to achieve greater resolution within the body of the data. In the case of large data sets, plotting less than n_a+2 points should be adequate. Begin by using 100 points for these larger data sets. The range of the indicator will have a part in determining if this is an adequate number of points. Trying the plots with differing numbers of points may be useful to see if the graph changes significantly.

The y-axis (CDF) should range in values from zero to 1. This method may result in confidence limits which drop below zero or exceed 1. These limits should not appear on the plot. Instead, truncate the plotted upper limit at 1 and the plotted lower limit at zero.

5 Formulas and Definitions

The estimated size-weighted CDF (proportion) for indicator value x_k in subpopulation a, $\hat{F}_a(x_k)$, with known subpopulation size, W_a; Horvitz-Thompson estimator is

$$\hat{F}_a(x_k) = \frac{\sum_{i=1}^{n_a} \frac{w_i}{\pi_i} I(y_i \le x_k)}{W_a} .$$

The estimated size-weighted CDF (proportion) for indicator value x_k in subpopulation a, $\hat{F}_a(x_k)$, with estimated subpopulation size, $\hat{W}_a$; Horvitz-Thompson ratio estimator is

$$\hat{F}_a(x_k) = \frac{\sum_{i=1}^{n_a} \frac{w_i}{\pi_i} I(y_i \le x_k)}{\hat{W}_a} ; \qquad \hat{W}_a = \sum_{i=1}^{n_a} \frac{w_i}{\pi_i} .$$

The one-sided $100(1 - \alpha)\%$ upper confidence bound, $B_U(x_k)$ is

$$B_U(x_k) = \hat{F}_a(x_k) + z_\alpha \sqrt{\hat{V} [\hat{F}_a(x_k)]} .$$

The one-sided $100(1 - \alpha)\%$ lower confidence bound, $B_L(x_k)$ is

$$B_L(x_k) = \hat{F}_a(x_k) - z_\alpha \sqrt{\hat{V} [\hat{F}_a(x_k)]} .$$

The two-sided $100(1 - \alpha)\%$ confidence interval, $C(x_k)$ is

$$C(x_k) = \left(\hat{F}_a(x_k) - z_{\alpha/2} \sqrt{\hat{V} [\hat{F}_a(x_k)]} , \quad \hat{F}_a(x_k) + z_{\alpha/2} \sqrt{\hat{V} [\hat{F}_a(x_k)]} \right) .$$

For these equations:

$\hat{V} [\hat{F}_a(x_k)] =$ estimated variance of the 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.

n_a = number of units sampled from subpopulation a.

z_α = z-score from the standard Normal distribution.

α = level of significance.

6 Procedure

6.1 Enter Data

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

Calcium y_i	Inclusion Probability π_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 π_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 Obtain Subpopulation Size (Size-Weighted Total)

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 Horvitz-Thompson ratio estimator. Divide each w_i by the inclusion probability, π_i , for all units in the sample *a*. Sum each of these quantities to obtain $\hat{W}_a$.

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

6.4 Input Indicator Levels of Interest

Assign indicator levels of interest, ***x***, based on graphical display considerations. Choose one of the three methods previously discussed in Section 4.3.

6.4.1 The Recommended Approach — Levels of Interest

Form an expected range of the indicator before looking at the data. Next, examine the data set to see if the estimated range encompasses all ***y*** values. If not, increase the range to encompass the outlying $\boldsymbol{y}$ values. If there are large outliers, more points than n_a+2 may be needed to retain good resolution in the body of the plot. Determine evenly spaced $\boldsymbol{x}$ values across the chosen range.

For this example, the estimated range was .5 to 9.5 mg/L. The range does not have to be adjusted because it includes the observed y_i values. The point spacing interval for $\boldsymbol{x}$, $x_{int} = (x_{max} - x_{min})/(n_a - 1) = (9.5 - .5)/(10 - 1) = 9/9 = 1.0$. The 10 $\boldsymbol{x}$ values = (x_{min} , x_{min}+1.0, x_{min}+2(1.0), ...) = (.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5). Try obtaining the cumulative distribution function first with these $\boldsymbol{x}$ values and then again with an increased number of $\boldsymbol{x}$ values spaced closer together. More points across the range may be needed because all but one of the y_i values are less than 3.0.

6.4.2 The Empirical CDF — Levels of Interest

For the empirical CDF, $\boldsymbol{x}$ values = (.7395, 1.2204, 1.5992, 2, 2.3707, 2.8196, 2.9399, 7). Duplicate values in the data set, 1.5992, do not have to be repeated when forming $\boldsymbol{x}$.

6.4.3 The Midpoint Approach — Levels of Interest

Calculate the midpoints of each pair of y_i values to form $\boldsymbol{x}$. The first $\boldsymbol{x}$ value is (.7395+1.2204)/2 = .9800. In this particular data set, there are three occurrences of 1.5992. As a result, there are two midpoints of 1.5992. Regardless of how many times a midpoint is repeated, include it only once in $\boldsymbol{x}$. The $\boldsymbol{x}$ values = (.9800, 1.4098, 1.5992, 1.7996, 2.1854, 2.5952, 2.8798, 4.9700).

6.5 Compute Cumulative Distribution Function Values

Calculate $\hat{F}_a(x_k)$ for each element in $\boldsymbol{x}$ using the formulas from Section 5.

To calculate $\hat{F}_a(x_1)$, compare each y_i to x_1. If y_i is less than or equal to x_1, then w_i/π_i is added to the computation of $\hat{F}_a(x_1)$ until y_i exceeds x_1 (when using sorted data). Divide the cumulative sum of these $w_i/\pi_i's$ by $\hat{W}_a$ or W_a (depending on the estimator used) to obtain $\hat{F}_a(x_1)$.

Similarly, to calculate $\hat{F}_a(x_2)$, compare each y_i to x_2, add the $w_i/\pi_i's$ until y_i exceeds x_2, and then divide this sum by $\hat{W}_a$ or W_a.

Do this for every value in $\boldsymbol{x}$.

Below is an example for obtaining the cumulative sum for each $\hat{F}_a(x_k)$. Complete results for the example data are in Section 6.7.

Calcium y_i	Inclusion Probability π_i	Lake Area w_i	Indicator Level of Interest x_k	Cumulative Sum for $\hat{F}_a(x_k)$
.7395	.00375	1.081	.7395	1.081/.00375
1.2204	.01406	7.758	1.2204	1.081/.00375+7.758/.01406
1.5992	.07734	24.249	1.5992	1.081/.00375+7.758/.01406+24.249/.07734+ 28.018/.75000+29.702/.03750
1.5992	.75000	28.018		
1.5992	.03750	29.702		
2.0000	.75000	52.953	2.0000	1.081/.00375+7.758/.01406+24.249/.07734+ 28.018/.75000+29.702/.03750+52.953/.75

6.6 Compute Confidence Limits

Calculate the confidence bound (upper or lower) or confidence interval for each $F_a(x_k)$ using the formulas from Section 5.

Estimate the variance of $\hat{F}_a(x_k)$ using an applicable method listed in Section 7. Next, take the square root of the variance and multiply this square root by the z-score from the standard Normal distribution corresponding to the desired confidence level.
Add this quantity to $\hat{F}_a(x_k)$ to obtain the upper bound, $B_U(x_k)$. Subtract this quantity from $\hat{F}_a(x_k)$ to obtain the lower bound, $B_L(x_k)$. For the confidence interval, obtain both $B_L(x_k)$ and $B_U(x_k)$. For example, 1.645 would be the z_α for a one-sided 95% upper or lower confidence bound, and the $z_{\alpha/2}$ for a two-sided 90% confidence interval. A two-sided 95% confidence interval would use 1.96 for $z_{\alpha/2}$.

6.7 Output Results

Output the indicator levels of interest, the associated size-weighted CDF value, and either a confidence bound (upper or lower) or a confidence interval for $F_a(x_k)$. If the output generated will be used for graphing the CDF, append the first and last graph points to this output as directed for the three methods below. The tables in Section 6.7.1 – 6.7.3 contain results for the ratio estimator applied to the example data. A hypothetical variance is used in confidence bound and interval calculations.

When upper bounds exceed 1, they are set equal to 1. Lower bounds less than zero are set equal to zero.

6.7.1 The Recommended Approach — Results

Append the point (0,0) to the output file for graphing purposes. Since x_{max} , 9.5, exceeds the maximum y_i , 7, no other points are appended.

Calcium x_k	Size-Weighted CDF, Ratio Estimator $\hat{F}_a(x_k)$	Hypothetical Variance $\hat{V}\ [\hat{F}_a(x_k)]$	One-sided 95% Lower Conf. Bound $B_L(x_k)$	One-sided 95% Upper Conf. Bound $B_U(x_k)$	Two-sided 90% Conf. Interval $C(x_k)$
0*	0*	0*	0*	0*	(0,0)*
0.5	0	0	0	0	(0,0)
1.5	.0054	.000032	0**	.0147	(0,.0147)
2.5	.1719	.032777	0**	.4697	(0,.4697)
3.5	.3769	.084169	0**	.8542	(0,.8542)
4.5	.3769	.084169	0**	.8542	(0,.8542)
5.5	.3769	.084169	0**	.8542	(0,.8542)
6.5	.3769	.084169	0**	.8542	(0,.8542)
7.5	1	0	1	1	(1,1)
8.5	1	0	1	1	(1,1)
9.5	1	0	1	1	(1,1)

*appended **set to 0

6.7.2 The Empirical CDF — Results

Append the point (0,0) to the output file for graphing purposes. Append also a point slightly larger than the largest $\boldsymbol{x}$ value and assign an ordinate of 1. For this example, the point (7.5,1) is appended.

Calcium x_k	Size-Weighted CDF, Ratio Estimator $\hat{F}_a(x_k)$	Hypothetical Variance $\hat{V}\ [\hat{F}_a(x_k)]$	One-sided 95% Lower Conf. Bound $B_L(x_k)$	One-sided 95% Upper Conf. Bound $B_U(x_k)$	Two-sided 90% Conf. Interval $C(x_k)$
0*	0*	0*	0*	0*	(0,0)*
0.7395	.0019	.000006	0**	.0057	(0,.0057)
1.2204	.0054	.000032	0**	.0147	(0,.0147)
1.5992	.0128	.000129	0**	.0314	(0,.0314)
2.0000	.0132	.000134	0**	.0323	(0,.0323)
2.3707	.1719	.032777	0**	.4697	(0,.4697)
2.8196	.2127	.039204	0**	.5383	(0,.5383)
2.9399	.3769	.084169	0**	.8542	(0,.8542)
7.0000	1	0	1	1	(1,1)
7.5000*	1*	0*	1*	1*	(1,1)*

*appended **set to 0

6.7.3 The Midpoint Approach — Results

Determine the first plotted point by calculating the distance between the first two y_i values, .7395 and 1.2204. Take half this distance and subtract it from .7395 to obtain .7395 – [(1.2204 –.7395)/2] = .4991. Append to the output (.4991,0) as the first plotted point. If a negative number were obtained and the indicator can never be negative, append (0,0) as the first plotted point. Similarly, to determine the last plotted point, calculate the distance between the two largest y_i values, 2.9399 and 7. Take half this distance and add it to 7 to obtain 7 + [(7–2.9399)/2] = 9.0301. Because the distance between these last two y_i values is relatively large, choosing the last point slightly above 7 with an ordinate of 1 may be preferable over appending (9.0301,1) to the output. For this example, (7.5,1) was appended.

Calcium	Size-Weighted CDF, Ratio Estimator	Hypothetical Variance	One-sided 95% Lower Conf. Bound	One-sided 95% Upper Conf. Bound	Two-sided 90% Conf. Interval
x_k	$\hat{F}_a(x_k)$	$\hat{V}\ [\hat{F}_a(x_k)]$	$B_L(x_k)$	$B_U(x_k)$	$C(x_k)$
.4991*	0*	0*	0*	0*	(0,0)*
.9800	.0019	.000006	0**	.0057	(0,.0057)
1.4098	.0054	.000032	0**	.0147	(0,.0147)
1.5992	.0128	.000129	0**	.0314	(0,.0314)
1.7996	.0128	.000129	0**	.0314	(0,.0314)
2.1854	.0132	.000134	0**	.0323	(0,.0323)
2.5952	.1719	.032777	0**	.4697	(0,.4697)
2.8798	.2127	.039204	0**	.5383	(0,.5383)
4.9700	.3769	.084169	0**	.8542	(0,.8542)
7.5000*	1*	0*	1*	1*	(1,1)*

*appended **set to 0

7 Associated Methods

An appropriate variance estimator for this estimated size-weighted CDF for discrete resources may be found in Method 7 (Horvitz-Thompson Variance Estimator).

8 Validation Data

Actual data with results, EMAP Design and Statistics Dataset #3, are available for comparing results from other versions of these algorithms.

9 Notes

The method which uses the ratio estimator may perform better under certain conditions and may be used even if the subpopulation size is known. Sampling done with variable probability and variable sample size, n_a, are two of these conditions. The ratio estimator retains a stability under these cases which can be seen from comparing the two equations. The ratio estimator tends to have smaller variance than the other estimator because the numerator and denominator tend to be positively correlated. The estimator using the known subpopulation size does not compensate for variability in the numerator.

The ratio estimator should be used in the case of missing data. The estimated CDF applies only to the subpopulation for which data were obtained. Because the size of this subpopulation is not known, it must be estimated. Therefore, the ratio estimator is the only alternative for estimating the CDF. All graphs should be labeled as applying only to the population that was sampled and not to the original target population.

10 References

U.S. Environmental Protection Agency (EPA). 1993. *Surface waters 1991 pilot report.* EPA/620/R-93/003. Washington, D.C: U.S. Environmental Protection Agency.

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. 1987. *A sampling and analysis plan for streams in the National Surface Water Survey*. Technical Report 117. Corvallis, OR: Oregon State University, Department of Statistics.