A Simple Method for Approximating the Variance of a Complicated Estimate

Author(s): Ralph S. Woodruff

Source: *Journal of the American Statistical Association*, Vol. 66, No. 334 (Jun., 1971), pp. 411-414
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2283947
Accessed: 06-02-2016 22:44 UTC

© Journal of the American Statistical Association
June 1971, Volume 66, Number 334
Theory and Methods Section

A Simple Method for Approximating the Variance of a Complicated Estimate

RALPH S. WOODRUFF*

A method often used for computing the variance of a complicated sample estimate is to first apply the Taylor approximation to reduce non-linear forms of the variables to linear form. This article shows the useful results which can be obtained by merely reversing the order between selection units and component variables in this linear expression. The method is completely general (assuming that the samples are large enough to justify using the Taylor approximation) involving no restrictions on (a) the form of the estimate, (b) the number of random variables involved in the estimate, (c) the type, complexity or number of the sample designs involved in the estimate.

1. THE PROBLEM

The complicated estimate for which a variance is desired will be called F. F is a function of M estimated totals $(X_{1\cdot}, X_{2\cdot}, \cdots, X_{i\cdot}, \cdots, X_{M\cdot})$ which are derived from a sample survey (or surveys).

The usual so-called large sample approximation to the variance of F is

$$E\left(\sum_{i=1}^{M} \frac{\partial F}{\partial X_{i\cdot}} (X_{i\cdot} - EX_{i\cdot}) \right)^2 \qquad (1.1)$$

where the partial derivatives $(\partial F/\partial X_{i\cdot})$ are evaluated at expected values. The estimate in the brackets is the first-order Taylor approximation of the deviation of F from its expected value.[1] It is assumed that we are dealing with sample sizes large enough that the remaining terms of the Taylor approximation can be neglected.[2]

The usual method of evaluating the variance approximation in (1.1) is to expand the squared sum into M squares (variances) and $M(M-1)$ cross-products (covariances). This paper will show that the evaluation of the variance can be greatly simplified by reordering the components of this sum before taking squares and cross-products.

To explain the method, it is necessary to introduce notation which provides for identifying both the type of variate and the random draw (selection unit). Let

- x_{ij} (referred to as the unweighted draw-variate) indicate the (unweighted) value of the ith type of variate for the jth random draw,[3]
- P_j indicate the probability of selection for the jth random draw,[4]
- X_{ij} indicate the value of the weighted random-draw-variate (x_{ij}/P_j),
- $X_{i\cdot}$, $\overline{X}_{i\cdot}$ indicate estimated totals and means respectively (for the ith type of variate) derived from sample observations.

When taking expected values, we will need to refer to members of the population (i.e., selection units of the same form as the random draws) and means and totals derived from the population. For this purpose, the symbol Y will be used. Thus Y_{ij} indicates the value of the ith type of variate for the jth selection unit in the population.

An example is provided to illustrate the procedure for approximating the variance of a complicated estimate. The example used involves a stratified sample of business establishments (a simple random sample of n_h out of N_h establishments is drawn without replacement from each R strata). Note that the number of random draws (n in the general notation) becomes $\sum_{h=1}^{R} n_h$ in this example, and that the generalized symbol j for a given random draw is defined by the combination of the symbols h and k (for draw within stratum).

The probability of each random draw (P_j in the general notation) becomes n_h/N_h in this example. Each business establishment drawn into the sample is canvassed for its sales (x_1) and its inventory (x_2) where the subscripts 1 and 2 represent the subscript i of the general notation which denotes the type of variate. The complicated estimate F for which the variance is to be approximated is the sales-inventory ratio computed over all strata so that:

$$F = \frac{\sum_{h=1}^{R} \sum_{k=1}^{n_h} \left(\frac{N_h}{n_h} x_{1hk} \right)}{\sum_{h=1}^{R} \sum_{k=1}^{n_h} \left(\frac{N_h}{n_h} x_{2hk} \right)}$$

where parentheses are used to identify the weighted draw-variates.

---

* Ralph S. Woodruff is assistant chief for research and methodology, Business Division, Bureau of the Census, Washington, D.C.

[1] Equation (1.1) is a well-known and widely-used result in the sampling theory. For example, see [9, p. 585].

[2] Hsu [5] proves the validity of these approximations when the samples become indefinitely large. There remains the question as to how "large" a sample must be in practice to justify the use of these approximations. For a discussion of this point as applied to ratio and regression estimates, see [2, pp. 159–60, 196–8].

[3] In practice, the random draws may be the results of a complicated operation involving several sets of notation. For example, a given random draw might be the result of the lth area sample segment draw from the kth primary sampling unit in the hth stratum of the gth sample design. However, for the purpose of simple presentation we will assume that all the random draws involved in F are reordered into a single numbering system $(1, 2, \cdots, j, \cdots, n)$.

[4] This probability should be computed through all stages of sampling.

2. PROCEDURE FOR APPROXIMATING THE VARIANCE OF THE COMPLICATED ESTIMATE

The suggested procedure is a generalization of the principle used by Keyfitz [6] to obtain variances for specific types of estimates derived from specific sample designs.[5] It is explained as a series of steps.

Step 1. Express the estimate F in terms of the weighted draw-variates.[6]

In the generalized notation:

Express F as a function of $X_{11}, \cdots, X_{ij}, \cdots, X_{Mn}$.

In the example:

$$F = \frac{\sum_{h=1}^{R} \sum_{k=1}^{n_h} X_{1hk}}{\sum_{h=1}^{R} \sum_{k=1}^{n_h} X_{2hk}}$$

where

$$X_{1hk} = \frac{N_h}{n_h} x_{1hk}.$$

Step 2. Take the partial derivative of F with respect to each weighted draw-variate. Evaluate this partial derivative at expected values and then multiply by the weighted draw-variate. Sum over all weighted draw-variates. Call the result of this operation F'.

In the generalized notation compute:

$$F' = \sum_{i=1}^{M} \sum_{j=1}^{n} \frac{\partial F}{\partial X_{ij}} X_{ij} \tag{2.1}$$

where $\partial F/\partial X_{ij}$ is evaluated at expected values.

In the example:

$$F = \sum_{h=1}^{R} \sum_{k=1}^{n_h} \frac{1}{Y_2} X_{1hk} - \sum_{h=1}^{R} \sum_{k=1}^{n_h} \frac{Y_{1.}}{Y^2_{2.}} X_{2hk}. \tag{2.1E}$$

Step 3. Reverse the order of summation in (2.1) so that subtotals for each draw are obtained.

In the generalized notation rearrange F' so that

$F' = \sum_{j=1}^{n} U_j$ where U_j is the weighted draw subtotal, i.e.,

$$U_j = \sum_{i=1}^{M} \frac{\partial F}{\partial X_{ij}} X_{ij}. \tag{2.2}$$

In the example:

$$F' = \sum_{h=1}^{R} \sum_{k=1}^{n_h} U_{hk} \tag{2.2E}$$

where

$$U_{hk} = \frac{N_h}{n_h}\left(\frac{x_{1hk}}{Y_{2.}} - \frac{Y_{1.}}{Y^2_{2.}} x_{2hk}\right).$$

Step 4. Find the variance of F' as expressed in (2.2).

In the example, it was specified that the sample design consisted of a simple random sample of n_h business establishments drawn without replacement from a population of N_h in each of R strata. The variance of the sum

$$F' = \sum_{h=1}^{R} \sum_{k=1}^{n_h} U_{hk}$$

with such a sampling system is

$$E(F' - EF')^2 = \sum_{h=1}^{R} \left(\frac{N_h - n_h}{N_h}\right) n_h \frac{\sum_{k=1}^{N_h} (V_{hk} - \overline{V}_h)^2}{N_h - 1}. \tag{2.3E}$$

(The symbol V is used to denote population values corresponding to the random variates U.)

Step 5. Substitute for the V values in the variance developed in Step 4.

In the example:

$$\sigma^2_F \doteq \sigma^2_{F'} = \frac{1}{Y^2_{2.}} \sum_{h=1}^{R} \frac{(N_h - n_h)}{(N_h)} \frac{N_h^2}{n_h} \cdot \frac{\sum_{k=1}^{N_h}\left(Y_{1hk} - \frac{Y_{1.}}{Y_{2.}} Y_{2hk} - \overline{Y}_{1h} + \frac{Y_{1.}}{Y_{2.}} \overline{Y}_{2h}\right)^2}{N_h - 1}. \tag{2.4E}$$

3. PROOF THAT THE SUGGESTED PROCEDURE IS IDENTICAL WITH USUAL LARGE SAMPLE APPROXIMATION TO VARIANCE OF NON-LINEAR ESTIMATES

If we express the variance approximation from Section 2 as

$$E(F' - EF')^2 = E\left(\sum_{i=1}^{M} \sum_{j=1}^{n} \frac{\partial F}{\partial X_{ij}} X_{ij} - E \sum_{i=1}^{M} \sum_{j=1}^{n} \frac{\partial F}{\partial X_{ij}} X_{ij}\right)^2$$

and then sum over j within the bracketing we obtain (1.1) because $\partial F/\partial X_{ij}$ has the same value for all j and is equal to $\partial F/\partial X_i$.

4. ESTIMATION OF THE VARIANCE OF A COMPLICATED ESTIMATE FROM THE SAMPLE ITSELF

If the sampling system involves two or more draws per stratum, the variance of a complicated estimate can be easily estimated from the sample by using the values of U_{hk} and $\overline{U}_h$ in the usual estimating formulas. (See (4.1E) below for application of this principle to the example). If there is only one draw per stratum made, then the variance is approximated from the sample by collapsing strata (see [4, p. 400]). If multi-stage sampling is used, variance among weighted values at the first stage will produce the correct variance (but not allocated to stages) except for first-stage finite multipliers.[7]

A special problem arises in the estimation of the variance of a complicated estimate from the sample itself

[5] These principles were also used by Kish in [8] to find variances of indexes from complicated samples.

[6] In some complicated estimates the procedure would apply without weighting the draw-variates. However, since the weights are not always uniform and in some sample designs may be even random variables, they are included in the generalized procedure.

[7] If this approximation is used, the U_j values (draw subtotals) used in Section 2 should be summarized at the first stage. There is no theoretical or practical difficulty in obtaining values of U at the first stage since they are simple sums of the U_j values at the later stages.

because the formula calls for evaluating the partial derivatives at the expected values of the variables. Since these expected values are ordinarily not available, estimates of the variables available from the sample are substituted in the formulas for the expected values.

In the example given, the variance ($S^2_{F'}$) would be estimated from the sample in the following way:

$$S^2_{F'} = \sum_{h=1}^{R} \left(\frac{N_h - n_h}{N_h}\right) n_h \quad \frac{\sum_{k=1}^{nh} (U_{hk} - \overline{U}_h)^2}{n_h - 1},$$

or substituting for U_{hk} and $\overline{U}_h$ and using values derived from the sample to evaluate the partial derivatives, we obtain as an estimate from the sample of the variance shown in (2.4E):

$$S^2_{F'} = \frac{1}{(X_{2.})^2} \sum_{h=1}^{R} \frac{(N_h - n_h)}{(N_h)} \frac{N_h{}^2}{n_h} \frac{\sum_{k=1}^{n_h} \left(x_{1hk} - \frac{X_{1.}}{X_{2.}} x_{2hk} - \overline{X}_{1h} + \frac{X_{1.}}{X_{2.}} \overline{X}_{2h}\right)^2}{n_h - 1}. \tag{4.1E}$$

5. RELATIONSHIP OF THE PROPOSED SYSTEM OF VARIANCE COMPUTATION TO THE REPLICATION OR RANDOM-PART METHOD OF VARIANCE COMPUTATION

In the random-part or replication method of variance computation, the sampling process, regardless of how complicated, is simply repeated (usually without replacement) a certain number of times. The applicable variance then is that of simple random sampling without replacement (see [3]).

Let X_j be the weighted estimated total for the jth replication, n be the number of replications and f be the sampling fraction (no i subscript is used since we are referring to a single type of variate).

Then X (the estimated total) $= \sum_{j=1}^{n} X_j$.

The variance of X can be estimated from the sample from

$$(1 - f)\frac{n}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2, \tag{5.1}$$

where $\overline{X}$ is the average value per replication as estimated from the sample.

There is no doubt that this scheme reduces the complexity of variance computation (sometimes, however, with a loss of efficiency over more complicated designs) and for some designs (for example, systematic sampling) it is the only way to obtain an unbiased estimate of the variance.

The proposed system of variance computation for complicated estimates is not a competitor with the replication system of variance computation but rather it is supplementary to the replication method.

Let F stand for a complicated estimate derived from a replicated sampling system. The variance form

$$(1 - f) \frac{(n)}{(n - 1)} \sum_{j=1}^{n} (F_j - \overline{F})^2$$

is often used as an approximation to the variance of F. However, this form is applicable only if

$$F = \sum_{j=1}^{n} F_j,$$

where F_j is an estimate for the jth replication of the same form as F. To illustrate this point with a simple example consider the ratio estimate between two variables $X_{1.}$ and $X_{2.}$ both of which are derived from the same replicated sampling system.

If the estimate T were derived from

$$\frac{\sum_{j=1}^{n} T_j}{n},$$

where T_j is the simple ratio

$$\frac{X_{1j}}{X_{2j}}$$

from a single random group, then the variance form

$$\frac{1}{n^2} (1 - f) \frac{(n)}{(n - 1)} \sum_{j=1}^{n} (T_j - \overline{T})^2$$

would be applicable. It is unlikely, however, that this estimate would be used because of the additional bias hazard involved,[8] and the probable form of estimate would be

$$T = \frac{X_{1.}}{X_{2.}} = \frac{\sum_{j=1}^{n} X_{1j}}{\sum_{j=1}^{n} X_{2j}}.$$

In this case the approximation

$$U_j = \left(\frac{X_{1j}}{X_{2.}} - \frac{TX_{2j}}{X_{2.}}\right)$$

should be substituted for X_j in the variance formula (5.1).

The variance of a complicated estimate derived from a replicated sampling system can be estimated from

$$S^2_{F'} = (1 - f) \left(\frac{n}{n - 1}\right) \sum_{j=1}^{n} (U_j - \overline{U})^2,$$

where

$$U_j = \sum_{i=1}^{M} \frac{\partial F}{\partial X_{ij}} X_{ij},$$

where the factor $\partial F/\partial X_{ij}$ is approximated (where necessary) by using estimates derived from all random groups. This combination of replicated sampling with the proposed method for approximating variances for complicated estimates provides the maximum simplicity in variance computation.

REFERENCES

[1] Brillinger, D. R. and Tukey, J. W., *Asymptotic Variances, Moments and Cumulants*, Multilith, Princeton, N. J.: Princeton University, 1964.

[8] See Section 4.4 of [9] for a further discussion of this point.

[2] Cochran, William G., *Sampling Techniques*, 2nd ed., New York: John Wiley & Sons, Inc., 1963.

[3] Deming, W. Edwards, *Sample Design in Business Research*, New York: John Wiley & Sons, Inc., 1960.

[4] Hansen, Morris H., Hurwitz, William N. and Madow, William G., *Sample Survey Methods and Theory*, Vol. I., New York: John Wiley & Sons, Inc., 1953.

[5] Hsu, P. L., "The Limiting Distribution of Functions of Sample Means and Application to Testing Hypotheses," *University of California, Berkeley Symposium on Mathematical Statistics and Probability. First Proceedings* (1945–1946), 359–402.

[6] Keyfitz, Nathan, "Estimates of Sampling Variance Where Two Units Are Selected From Each Stratum," *Journal of the American Statistical Association*, 52 (December 1957), 503–10.

[7] ———, "Sampling for Demographic Variables," in N. L. Johnson and H. Smith, Jr., eds., *New Developments in Survey Sampling*, New York: John Wiley and Sons, Inc., 1969, 562–77.

[8] Kish, Leslie, "Standard Errors for Indexes From Complex Samples," *Journal of the American Statistical Association*, 63 (June 1968), 512–29.

[9] ———, *Survey Sampling*, New York: John Wiley & Sons, Inc., 1965.

[10] Tepping, Benjamin, "The Estimation of Variance in Complex Surveys," *Proceedings of the Social Statistics Section of the American Statistical Association*, 1968, 11–18.