Journal of Official Statistics, Vol. 14, No. 2, 1998, pp. 181–188

Bias Correction in the Balanced-half-sample Method if the Number of Sampled Units in Some Strata Is Odd

Ger T. Slootbeek[1]

The balanced-half-sample method is a general method for estimating variances of statistics based on complex sample surveys. A simple extension of this method to the case of more than two sampled units in some strata is the grouping method. The sampled units in a stratum are randomly grouped into two equal or nearly equal groups and the balanced-half-sample method is applied to these groups. If all strata have an even number of sampled units, then for linear estimators this grouping method results in unbiased variance estimators, otherwise it results in biased variance estimators. For simple linear estimators an analytical formula for this bias is derived. It turns out that this bias can be reduced by an appropriate location transformation of the target and auxiliary variable observations.

Key words: BHS method; grouping method; location transformation; variance estimation; regression estimator.

1. Introduction

The balanced-half-sample (BHS) method is a general method for estimating variances of statistics based on complex sample surveys. It has been introduced for stratified sampling designs where in each stratum two primary sampling units ($n_h = 2$) are selected with replacement. There are many situations where some n_h are not 2. Therefore several extensions are described in the literature for $n_h > 2$. There are two approaches to extend the method to $n_h > 2$. The first approach concerns methods of simple practical use. As an example of this approach, Wolter (1985) describes the grouping method with all n_h even in which the sampled units within each stratum are randomly grouped into two equal groups and the BHS method is applied to these groups. The second approach makes use of orthogonal arrays. Wu (1991) extends the balanced-half-sample method to the case of unequal sample-strata sizes, using mixed orthogonal arrays of strength two. Sitter (1993) extends the orthogonal array method to the orthogonal multi-array method.

In this article we restrict our attention to the grouping method, where n_h is not necessarily even. We will follow the terminology and notation in Wolter (1985, Chapter 3). If the number of sampled units in some strata is odd, then the grouping method results in a biased variance estimator. In Section 2 we derive an analytical formula for this bias for simple linear estimators. A location transformation of the target variable can be used to reduce this bias or other techniques can be applied (e.g., Wu's or Sitter's method). In Section 3

[1] Statistics Netherlands, Department of Statistical Methods, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. The views expressed in this article are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

we extend the grouping method for the simple regression estimator as an example of a nonlinear estimator that can be linearized. In Section 4 we discuss when and how to use bias corrections by transformation.

2. An Extension of the BHS-method for Linear Estimators in the Case $n_h > 2$ and Odd for Some Strata

2.1. Introduction

Let L be the number of strata, N the population size and N_h the population size of stratum h, $h = 1, \ldots, L$. If in stratum h the number of sampled units is odd, then let $n_h = 2m_h + 1$, else $n_h = 2m_h$ ($m_h > 0$). The theoretical results of this section only apply to the textbook estimator of the population mean $\bar{Y}$, that is to

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{2.1}$$

where $\bar{y}_h$ is the sample mean of stratum h and $W_h = N_h/N$.

2.2. The grouping method

In the grouping method the sampled units in each stratum are randomly divided into two groups. Let p_h be the number of sampled units in the first group and q_h the number of sampled units in the second group and let $p_h \leq q_h$. The BHS method is applied to the groups thus formed. The set of balanced half samples can be specified by using a Hadamard matrix. An entry, $\delta_{\alpha h}$, of $+1$ in the (α, h)th cell signifies that the first group of stratum h is part of the αth half sample, and similarly $\delta_{\alpha h} = -1$ for the second group.

We define the αth BHS estimator of $\bar{Y}$ by

$$\bar{y}_{st,\alpha} = \sum_{h=1}^{L} W_h a_{hi(\alpha)} \bar{y}_{hi(\alpha)} \tag{2.2}$$

where $\bar{y}_{hi(\alpha)}$ is either $\bar{y}_{h1}$ (the sample mean of the p_h units of the first group) or $\bar{y}_{h2}$ (the sample mean of the q_h units of the second group) according to the value of $\delta_{\alpha h}$. For any stratum h, $a_{h1} = 2p_h/n_h$ and $a_{h2} = 2q_h/n_h$ are chosen proportional to p_h and q_h, so that the stratum sample mean can be written as

$$\bar{y}_h = (a_{h1}\bar{y}_{h1} + a_{h2}\bar{y}_{h2})/2 \tag{2.3}$$

(If $p_h = q_h$, then $a_{h1} = a_{h2}$.)

The BHS-variance estimator is defined by

$$v_k(\bar{y}_{st}) = \frac{1}{k}\sum_{\alpha=1}^{k}(\bar{y}_{st,\alpha} - \bar{y}_{st})^2 \tag{2.4}$$

where k is the number of half samples used.

Theorem 1. The expectation of the BHS-variance estimator (Equation 2.4), applied to

the linear half-sample estimators given by (2.2), can be written as

$$E(v_k(\bar{y}_{st})) = VAR(\bar{y}_{st}) + \frac{1}{k}\sum_{\alpha=1}^{k}(E(\bar{y}_{st,\alpha}) - E(\bar{y}_{st}))^2 \tag{2.5}$$

Proof: see Appendix.

The second term of the righthand side of (2.5) is the bias of the BHS-variance estimator and can have the same magnitude as the variance (the first term) if there are many small and odd n_h's, see example below. For even n_h in a given stratum one can take $p_h = q_h$ (p_h, q_h integers) and the corresponding bias contribution becomes zero. If the number of sampled units in some strata is odd, then in these strata p_h and q_h have to be taken nearly equal to minimize the bias of the BHS-variance estimator, as the following theorem shows.

Theorem 2. If the number of sampled units in some strata is odd, then the BHS-variance estimator is biased. This bias can be minimized by taking $p_h = m_h$ and $q_h = m_h + 1$ in the strata in which the number of sampled units is odd.

Proof: see Appendix.

Corollary. The bias of the BHS-variance estimator of $\bar{y}_{st}$, that is the second term of the righthand side of (2.5), can be written as

$$\sum_{h=1}^{L} W_h^2 (a_{h1} - a_{h2})^2 \bar{Y}_h^2/4 \tag{2.6}$$

where $\bar{Y}_h$ is the stratum h mean.

Example: A population of $N = 341$ units is divided into seven strata. Seven stratified samples are drawn with replacement. Since in this example the information about the whole population is known we can calculate the stratum variances σ_h^2 and the stratum means $\bar{Y}_h$. The population information is given in Table 1. Besides it is possible to calculate VAR($\bar{y}_{st}$) and BIAS (Equation 2.6). In Table 2 the samples and the results are given. Sample 1 shows that the bias (second term of 2.5) can be serious. Sample 4 is a stratified sample proportional to the stratum size. In Sample 5 the values of the target variable are reduced by 1,500 (a location transformation). It shows that a location transformation of the target variable can be useful. If we change the sample sizes in Stratum 3 and 4 in Sample 4 from 3 to 2 and 4 respectively we get Sample 6. Sample 7 is Sample 4 modified so as to comprise even sample sizes in all the strata.

Table 1. Population information

h	N_h	σ_h^2	$\bar{Y}_h$
1	60	53,903	356
2	38	46,878	324
3	23	26,699	344
4	26	330,215	3,534
5	73	264,557	1,607
6	49	250,184	1,356
7	72	319,125	1,440

Table 2. Seven samples with results

Sample No.	n_1	n_2	n_3	n_4	n_5	n_6	n_7	VAR ($\bar{y}_{st}$)	BIAS	REL*
1	3	3	3	3	3	3	3	11,936	36,348	3.05
2	7	7	7	7	7	7	7	5,116	6,676	1.31
3	15	15	15	15	15	15	15	2,387	1,454	0.61
4	7	5	3	3	9	6	9	4,824	10,861	2.25
5	7	5	3	3	9	6	9	4,824	4,868	1.01
6	7	5	2	4	9	6	9	4,684	2,734	0.58
7	6	6	2	4	10	6	8	4,767	0	0.00

*REL = BIAS/VAR ($\bar{y}_{st}$)

2.3. Bias correction by a location transformation

It turns out that the bias of the BHS-variance estimator depends heavily on the values of the stratum means $\bar{Y}_h$ for strata with odd n_h. This suggests that the bias can be reduced by subtracting a constant A_h from each observation of the target variable, i.e., by the location transformation $z_{hj} = y_{hj} - A_h,\ h = 1, \ldots, L$ and $j = 1, \ldots, n_h$. For each stratum A_h has to be a predetermined constant. This transformation implies that $\bar{z}_h = \bar{y}_h - A_h$ and

$$\bar{z}_{st} = \sum_{h=1}^{L} W_h \bar{z}_h = \sum_{h=1}^{L} W_h (\bar{y}_h - A_h) \tag{2.7}$$

The variance of $\bar{z}_{st}$ can be written as

$$VAR(\bar{z}_{st}) = \sum_{h=1}^{L} W_h^2 VAR(\bar{y}_h - A_h) = \sum_{h=1}^{L} W_h^2 VAR(\bar{y}_h) = VAR(\bar{y}_{st}) \tag{2.8}$$

and the expectation of the BHS-variance estimator can be written as

$$\begin{aligned} E(v_k(\bar{z}_{st})) &= VAR(\bar{z}_{st}) + \sum_{h=1}^{L} W_h^2 (a_{h1} - a_{h2})^2 \bar{z}_h^2/4 \\ &= VAR(\bar{y}_{st}) + \sum_{h=1}^{L} W_h^2 (a_{h1} - a_{h2})^2 (\bar{Y}_h - A_h)^2/4 \end{aligned} \tag{2.9}$$

in which $\bar{z}_h$ is the population mean in stratum h of the transformed variable. Equation (2.9) shows that the closer A_h lies to $\bar{Y}_h$ in each stratum, the more the bias is reduced. The question arises how A_h can be determined. We assume, as is usually the case with recurring surveys, that it can be calculated from an earlier survey. There are two options for A_h. The first option is that $A_h = A$ for each stratum and the second option is that A_h differs from stratum to stratum. An example of option one: if y is the income of a person in a country, then A_h can be chosen as the modal income of a country calculated from an earlier year. An example of option two: if the region is the stratification variable, then A_h can be chosen as the modal income of region h calculated from an earlier year. Conclusion: if the number of sampled units is odd in a stratum, then one may try a location transformation of the target variable and apply the BHS method to the transformed variable.

3. An Extension for Simple Regression Estimators

3.1. Introduction

In this article the textbook estimator is used as an illustration of a linear estimator for which the discussed method is perfectly valid. If an estimator $\hat{T}$ can be linearized, then we may derive an analytical formula of the appropriate bias term, so as to find a useful transformation for each variable. This idea is illustrated in an example about the simple regression estimator. It turns out that both a location transformation of the target variable and the auxiliary variable may reduce the bias.

3.2. Example: The simple regression estimator

We consider a population U divided into L strata. Let each stratum consist of N_h elements $u_{h1}, \ldots, u_{hN_h}$. Each element u_{hj} is associated with a value y_{hj} of the target variable y and a value x_{hj} of the auxiliary variable x. The population mean and the stratum mean of the auxiliary variable are assumed to be known and will be denoted by $\bar{X}$ and $\bar{X}_h$. From the population U in each stratum a simple random sample is selected with replacement.

Let the population regression coefficient be denoted by B, cf., Bethlehem and Keller (1987). The regression estimator for $\bar{Y}$, the population mean of a target variable y, based on the parent sample is defined by

$$\bar{y}_R = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st}) \tag{3.1}$$

in which b is the usual estimator for B that is based on the parent sample. The regression estimator based on half sample α is defined by

$$\bar{y}_{R,\alpha} = \bar{y}_{st,\alpha} + b_\alpha(\bar{X} - \bar{x}_{st,\alpha}) \tag{3.2}$$

in which b_α is the estimator for B that is based on half sample α. The BHS-variance estimator is defined by

$$v_k(\bar{y}_R) = \frac{1}{k}\sum_{\alpha=1}^{k}(\bar{y}_{R,\alpha} - \bar{y}_R)^2 \tag{3.3}$$

The regression estimator $\bar{y}_R$ is approximated through Taylor linearization by replacing b and b_α by B in Equations (3.1) and (3.2). By using the approximated regression estimators of (3.1) and (3.2), the BHS-variance estimator (Equation 3.3) can then be approximated by

$$v_k(\bar{y}_R) \doteq v_k(\bar{y}_{st}) + B^2 v_k(\bar{x}_{st}) - 2B cov_k(\bar{x}_{st}, \bar{y}_{st}) \tag{3.4}$$

in which the BHS-covariance estimator is

$$cov_k(\bar{x}_{st}, \bar{y}_{st}) = \frac{1}{k}\sum_{\alpha=1}^{k}(\bar{x}_{st,\alpha} - \bar{x}_{st})(\bar{y}_{st,\alpha} - \bar{y}_{st}) \tag{3.5}$$

It can be easily shown (in the same way as for $v_k(\bar{y}_{st})$, see Wolter 1985) that for the linear estimators $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ and $\bar{y}_{st}$ in the two-per-stratum situation, $cov_k(\bar{x}_{st}, \bar{y}_{st})$ is an unbiased estimator of $COV(\bar{x}_{st}, \bar{y}_{st})$.

The expectation of the BHS-variance estimator can be approximated by

$$E(v_k(\bar{y}_R)) \doteq VAR(\bar{y}_{st}) + \sum_{h=1}^{L} W_h^2(\bar{Y}_h)^2(a_{h1} - a_{h2})^2/4 + B^2 VAR(\bar{x}_{st}) + \sum_{h=1}^{L} W_h^2(\bar{X}_h)^2(a_{h1} - a_{h2})^2/4 - 2BCOV(\bar{x}_{st}, \bar{y}_{st}) - 2B\sum_{h=1}^{L} W_h^2(\bar{X}_h)(\bar{Y}_h)(a_{h1} - a_{h2})^2/4 \quad (3.6)$$

in which (in the same way as Equation 2.5) the expectation of the BHS-covariance estimator based on the grouping method, can be written as

$$E(cov_k(\bar{x}_{st}, \bar{y}_{st})) = COV(\bar{x}_{st}, \bar{y}_{st}) + \sum_{h=1}^{L} W_h^2(a_{h1} - a_{h2})^2(\bar{Y}_h)(\bar{X}_h)/4 \quad (3.7)$$

Again, we make use of a location transformation A_h of the target variable y and C_h of the auxiliary variable x. The approximation of the expectation of the BHS-variance estimator $\bar{z}_R$ (transformed regression estimator) can be done in the same way as in Section 2.2. For each stratum A_h and C_h have to be predetermined constants. Again, the closer A_h lies to $\bar{Y}_h$ and C_h lies to $\bar{X}_h$ in each stratum, the more the bias is reduced. Since for each stratum the population mean of the auxiliary variable x is assumed to be known we can take C_h equal to $\bar{X}_h$. The bias of both $v_k(\bar{x}_{st})$ and $\text{cov}_k(\bar{x}_{st}, \bar{y}_{st})$ is zero. So, the expectation of the transformed regression estimator can be approximated by

$$E(v_k(\bar{z}_R)) \doteq VAR(\bar{y}_R) + \sum_{h=1}^{L} W_h^2(\bar{Y}_h - A_h)^2(a_{h1} - a_{h2})^2/4 \quad (3.8)$$

So, the expectation of the BHS-variance estimator of the regression estimator can be approximated by the same equation as the expectation of the BHS-variance estimator of the linear estimator in Section 2 (Equation 2.9).

4. Discussion

To make the BHS method applicable to the case $n_h > 2$, we have chosen a practical solution that is easy to implement. In this article we have extended the grouping method to the case that some of the strata have an odd number of sampled units. In this situation the BHS-variance estimator is biased even for linear estimators. Of course, strata with an even number of sampled units do not contribute to this bias.

If the number of sampled units in each stratum with an odd number of sampled units is sufficiently large, then it is not worthwhile to pay much attention to the bias of the BHS-variance estimator. However, when the number of sampled units is small we can reduce the amount of bias by an appropriate location transformation.

The constants A and C of a location transformation of the observations of the target and the auxiliary variable have to be predetermined. A good choice of these constants results in a substantial reduction of the bias. It is not necessary, however, to know the optimal value of these constants exactly. The function of the bias appears to be relatively

flat in the neighbourhood of the optimum. If the stratum mean of the target variable $\bar{Y}_h$ does not differ much between the odd-sized strata, then the location transformation with constants A and C for the whole population will already give a substantial bias reduction. In other cases it may be worthwhile for each stratum to specify stratum dependent constants A_h and C_h to improve upon the bias reduction.

For the linear estimators the discussed method is perfectly valid. If a nonlinear estimator can be linearized, then we may derive an analytical equation of the approximate bias term, which can be used to find a useful transformation for each variable. As an illustration, the discussed method is shown to work for the simple regression estimator, and it will work similarly for multiple regression. In other cases it is not generally possible to demonstrate the feasibility of the method discussed.

Appendix

Proof of Theorem 1: The expectation of the BHS-variance estimator can be written as

$$E(v_k(\bar{y}_{st})) = \frac{1}{k}\sum_{\alpha=1}^{k} E((\bar{y}_{st,\alpha} - \bar{y}_{st})^2)$$

$$= \frac{1}{k}\sum_{\alpha=1}^{k} VAR(\bar{y}_{st,\alpha} - \bar{y}_{st}) + \frac{1}{k}\sum_{\alpha=1}^{k}(E(\bar{y}_{st,\alpha} - \bar{y}_{st}))^2 \qquad \text{(A.1)}$$

Because of (2.3), for half sample α it holds that

$$VAR(\bar{y}_{st,\alpha} - \bar{y}_{st}) = VAR\left(\sum_{h=1}^{L} W_h a_{hi(\alpha)} \bar{y}_{hi(\alpha)} - \sum_{h=1}^{L} W_h \bar{y}_h\right)$$

$$= VAR\left(\sum_{h=1}^{L} W_h \delta_{\alpha h}(a_{h1}\bar{y}_{h1} - a_{h2}\bar{y}_{h2})/2\right)$$

$$= \sum_{h=1}^{L} W_h^2 \delta_{\alpha h}^2 VAR((a_{h1}\bar{y}_{h1} - a_{h2}\bar{y}_{h2})/2)$$

since sampling in one stratum is independent of that within another. Because of the random selection with replacement of the sampled units within a stratum and the random formation of the groups within a stratum, $\bar{y}_{h1}$ and $\bar{y}_{h2}$ are independent, yielding $\mathrm{COV}(a_{h1}\bar{y}_{h1}, a_{h2}\bar{y}_{h2}) = 0$. Hence

$$VAR(\bar{y}_{st,\alpha} - \bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 VAR((a_{h1}\bar{y}_{h1} + a_{h2}\bar{y}_{h2})/2)$$

$$= \sum_{h=1}^{L} W_h^2 VAR(\bar{y}_h) = VAR(\bar{y}_{st}) \qquad \text{(A.2)}$$

Equation A.2 holds for each half sample. Hence, a combination of A.1 and A.2 yields 2.5. Theorem 1 holds without using the balancing assumption.

Q.E.D.

Proof of Theorem 2: The second term of the righthand side in Equation 2.5 can be written as

$$\frac{1}{k}\sum_{\alpha=1}^{k}[E(\bar{y}_{st,\alpha}-\bar{y}_{st})]^2=\frac{1}{k}\sum_{\alpha=1}^{k}\sum_{h=1}^{L}W_h^2\left(E\left(\frac{a_{h1}\bar{y}_{h1}-a_{h2}\bar{y}_{h2}}{2}\right)\right)^2$$
$$+\frac{1}{k}\sum_{\alpha=1}^{k}2\sum_{h=1}^{L}\sum_{h'>h}^{L}W_hW_{h'}\delta_{\alpha h}\delta_{\alpha h'}\,\frac{E(a_{h1}\bar{y}_{h1}-a_{h2}\bar{y}_{h2})}{2}\,\frac{E(a_{h'1}\bar{y}_{h'1}-a_{h'2}\bar{y}_{h'2})}{2} \tag{A.3}$$

Since the set of half samples is balanced one has

$$\frac{1}{k}\sum_{\alpha=1}^{k}\delta_{\alpha h}\delta_{\alpha h'}=0 \tag{A.4}$$

Hence, the second term of the righthand side in (A.3) is zero. Since $E(\bar{y}_{h1})=E(\bar{y}_{h2})=\bar{Y}_h$ ($\bar{Y}_h$ is the stratum h mean), the first term of the righthand side of (A.3) can be written as

$$\frac{1}{k}\sum_{\alpha=1}^{k}\sum_{h=1}^{L}W_h^2\left(E\left(\frac{a_{h1}\bar{y}_{h1}-a_{h2}\bar{y}_{h2}}{2}\right)\right)^2$$
$$=\frac{1}{k}\sum_{\alpha=1}^{k}\sum_{h=1}^{L}W_h^2\bar{Y}_h^2(a_{h1}-a_{h2})^2/4 \tag{A.5}$$

Since $a_{h1}+a_{h2}=2$, Equation A.5 becomes 0 for $a_{h1}=a_{h2}=1$. So, if n_h is even then we form groups of equal size. If n_h is odd, (m_h, m_h+1) minimizes the righthand side of (A.5).

Q.E.D.

5. References

Bethlehem, J.G. and Keller, W.J. (1987). Linear Weighting of Sample Survey Data. Journal of Official Statistics, 3, 141–153.

Sitter, R.R. (1993). Balanced Repeated Replications Based on Orthogonal Multi-arrays. Biometrika, 80, 211–221.

Wolter, K. (1985). Introduction to Variance Estimation. Springer-Verlag, New York.

Wu, C.F.J. (1991). Balanced Repeated Replications Based on Orthogonal Arrays. Biometrika, 78, 181–188.

Received November 1995
Revised June 1997