CALIBRATION WEIGHTING AND ITS EFFECT ON MEASURES OF VARIATION

Richard A. Moore[1], U.S. Census Bureau
Richard A. Moore, U.S. Census Bureau, Room G176- FB3, Washington, DC 20233

Key Words: Calibration, Random Group Replication, General Variance Function

1 Introduction

The 2002 Survey of Business Owners (SBO) produces the most comprehensive set of available economic statistics for U.S. non-farm businesses by race, Hispanic, and gender of business owners. Ownership statistics for each race, Hispanic, and gender class include (1) the number of firms, (2) the aggregate receipts, (3) the aggregate number of persons employed, and (4) the aggregate payroll. Although a list of businesses with the corresponding sales and receipts, employment, and payroll is available, the race, gender, and Hispanic ownership of each is missing. (There is administrative information available which gives some indication of the likelihoods of ownership in the various racial, Hispanic, and gender classes.) The SBO collects this information from a sample of businesses. Each selected business is originally assigned a weight equal to the inverse of its probability of selection. Ownership class firm count estimates can then be produced by summing the weights of all businesses whose response indicates that they belong to that class. Similarly, aggregate estimates for the auxiliary variables (receipts, employment, and payroll) within these classes can be produced by inflating each variable by the sampling weight and summing the data for the businesses in the sample. Variance estimates are produced for each estimate using the random group replication method.

Several problems can arise. First, the distributions of receipts, employment, and payroll are often highly skewed. Consequently, the weighted sum of the data for the selected sample units may differ significantly from the sum of the characteristics for all units in the same class. Second, we are using one sample to produce estimates by race, Hispanic ethnicity, and gender. Although measures are taken to ensure that the sample is representative, administrative information often indicates that the sample is not representative for at least one of these three types. Calibration weighting can be used to correct these problems. This paper discusses the Survey of Business Owners and the development of a calibration weight for the 2002 SBO. It then discusses variance estimation in the context of calibration weighting. Finally, the paper compares random group estimation alone and in combination with a calibration adjustment to the design-based variances.

2 The Survey of Business Owners

The SBO produces the most comprehensive set of minority- and women-owned business statistics for U.S. companies operating in the private non-farm sector. The survey is currently conducted once every five years, in the years ending in 2 or 7 (e.g., 1992, 1997, 2002). Statistics are based on the responses to questions on the race, gender, and Hispanic origin of firm owners, which are asked of a probabilistic sample of businesses from a universe that was constructed from administrative tax records. The 2002 sample consists of approximately 2.6 million of the over 20 million businesses which filed tax returns in 2002. Of these approximately half are businesses with paid employees. The sample is designed to produce accurate estimates of firm counts, aggregate receipts, aggregate employment, and aggregate payroll by minority and gender of owner, for each state (including DC) and each three-digit NAICS (North American Industry Classification System) code. The minority and gender groups include four race groups (Black, Asian, Hawaiian/Pacific Islander, and Native American), one ethnicity group (Hispanic), and two gender classes (Female, and 50 percent Female). For example, the 1997 survey estimated that were 264 Hispanic-owned Health Service operations (doctors, dentists, hospitals, nursing homes, etc.) in Connecticut. These businesses had $31.2 million in sales and receipts, employed 171 people, and had payroll of $7.4 million.

To produce the estimates, we need a list of businesses which operated in 2002 and certain information for each of these businesses. This information includes the (1) state of operation, (2) 3-digit NAICS code, (3) sales and receipts, (4) number of paid employees, (5) payroll, (6)

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed here are those of the author and not necessarily those of the U.S. Census Bureau.

racial ownership composition, (7) gender ownership composition, and (8) Hispanic ownership composition. The U.S. Internal Revenue Service (IRS) supplies the U.S. Census Bureau with a file of all businesses which file a tax return in 2002. This is used as the universe for the survey. The first five types of information listed above are obtained from tax records supplied by the IRS.

Although the three ownership composition items are missing, there is other information available for many businesses that can provide some evidence about their ownership composition. For example, if the business has been in operation more than five years, it may have been selected in previous SBO and provided this information for a previous period. Assuming most businesses do not drastically change ownership composition over time, the previous ownership composition information can be used as a good indication for the current ownership composition. For some businesses, the IRS provides the Census Bureau with a list of the Social Security Numbers (SSNs) of the owner(s). These are sent to the Social Security Administration (SSA), which then supplies the Census Bureau with information such as each SSN owner's (1)gender, (2) race, (3) country of birth, (4) last name, and (5) the surnames of the parents. This provides indications on the possible race, Hispanic origin and gender of all known owners. Indicators for each known owner can be consolidated into one record for each business.

Other "weaker" evidence can be used to provide further indicators of ownership composition. The IRS also provides us with the mailing address of each business. This includes the 5-digit ZIP Code. Racial population distributions from the 2000 Census of Population and Housing can be used to further identify possible minority-owned businesses. (For example, if a large percentage of the population in a particular ZIP Code is Hispanic, this may indicate that the businesses located in that ZIP code may be more likely to be Hispanic-owned.) Finally, if you assume that the ownership composition of the businesses in a particular state and industry may not change significantly every five years, the distribution of businesses by race, ethnicity and gender estimated by state and industry from the previous SBO can be used as indicators of the potential race, ethnicity and gender composition of a business in a particular state and industry.

All of the above indicators are used as input for a logistic regression model. The model produces ownership likelihoods for each of seven "race" categories (White, Black, Asian, Hawaiian/Pacific Islander, Native American, Other, and Publicly-Held Corporation), a likelihood the business is Hispanic-owned, and a likelihood the business is female-owned. Based on the model, the record for each business in the frame is augmented with (1) the ownership's most likely minority race, (2) the likelihood of the most likely minority race, (3) the likelihood of Hispanic ownership, and (4) the likelihood of female-ownership.

Each business record in the frame now contains the following information: (1) state of operation, (2) 3-digit NAICS code, (3) sales and receipts, (4) employment, (5) payroll, (6) ownership's most likely minority race, (7) the likelihood of the most likely minority race, (8) the likelihood of Hispanic ownership, and (9) the likelihood of female-ownership. The three likelihoods are used to assign each business to one of nine ownership classes: (1) Potentially Asian-owned, (2) Potentially Hawaiian/Pacific Islander-owned, (3) Potentially Native American-owned, (4) Potentially White/Black Hispanic-owned, (5) Potentially Other-owned (6) Potentially Black-owned, (7) Potentially White Female-owned, (8) Probably Publicly Held , and (9) Probably White Male-owned. This ownership class is added to each frame record and used in the sample design.

The 2002 SBO uses a stratified random sample. Businesses are first stratified by four variables: (1) their ownership class, (2) state of operation, (3) 3-digit NAICS Code, and (4) employer/nonemployer status (i.e, whether the business has non-zero payroll). After stratification, businesses with very large (relative to the other records in the stratum) receipts, payroll, or employment are selected with certainty. Certainty cutoffs vary by stratum. The remaining (non-certainty) businesses are then subjected to sampling. Sampling rates vary by strata and are determined by the distribution of the likelihoods of the non-certainty businesses in the stratum.

All selected cases are mailed a questionnaire asking (among other things) their race, Hispanic, and gender ownership composition. Selected cases are assigned a sampling weight equal to the inverse of their probability of selection. Weight adjustments are made for non-response. Tabulations are based on the actual responses to the race, Hispanic, and gender ownership items and not on the likelihoods or the ownership class stratification variable. For example, suppose a business was classified as Potentially Asian-owned. Based on the administrative data available at the time of frame construction, it was assigned a likelihood of Asian-owned of 0.95, a likelihood of being Hispanic-owned of 0.02, and likelihood of being female-owned of 0.10. It was selected in the sample with weight of 10.0 and responds as White-Hispanic-Male-owned. It will be tabulated with a weight of 10.0 as a White-Hispanic-Male-owned business.

3 The Universe Used

For this study, we used the universe of firms with paid employees from the tax year 2001. (It is expected that this universe is very similar to the universe of businesses that filed a tax return in 2002. About 80 percent of the businesses filing tax returns in 2001 also filed returns in 2002.) The employer universe has over 5.4 million businesses. Administrative information from a variety of sources was evaluated to determine the likelihoods of gender, Hispanic, and racial ownership make-up of each business. Each business record contained the information described above, namely — (1) state of operation, (2) 3-digit NAICS code, (3) sales and receipts, (4) employment, (5) payroll, (6) owner(s)' most likely minority race, (7) the likelihood of the most likely minority race, (8) the likelihood of Hispanic ownership, and (9) the likelihood of female ownership. The three likelihoods are then used to assign each business to one of nine ownership classes: (1) Potentially Asian-owned, (2) Potentially Hawaiian/Pacific Islander-owned, (3) Potentially Native American-owned, (4) Potentially White/Black Hispanic-owned, (5) Potentially Other-owned (6) Potentially Black-owned, (7) Potentially White Female-owned, (8) Probably Publicly Held , and (9) Probably White Male-owned.

Stratification was performed as described above. There are 55,080 possible mutually exclusive strata. Cases with very large receipts, payroll, and/or employment in each were selected with certainty (i.e., their sampling weight is set to 1.00). A random sample (with sampling rates varying by strata) was selected from the remainder of the cases in each stratum to represent the stratum. The original sampling weights are set equal to the inverse of the probability of selection. About 350,000 businesses were selected with certainty. Almost 940,000 businesses were selected from the remaining 5 million.

4 The Need for Calibration Weighting

The 2002 SBO is a sample survey which attempts to measure many different characteristics (number of Black-owned businesses, aggregate receipts of Black-owned businesses, aggregate payroll for Black-owned businesses, and aggregate employment for Black-owned businesses, as well as the same characteristics for businesses owned by the other races, the Hispanics, and females). Usually, the sample (at the stratum level) is not representative of at least one of these characteristics. In fact, if the values of the sampling units within a stratum are skewed, it is possible that the sample will not give accurate estimates for the stratum.

The universe is stratified by four variables — the state of operation, the industry class, employer/nonemployer status, and one of the guesses. As mentioned before, cases with large employment or dollar volume of receipts or payroll (with respect to the other cases in the stratum) are selected with certainty. The other guesses and probabilities are used to sort the remaining cases. Suppose, for example, the stratum was defined to be probable Asian-owned restaurants in New York state. We would pull off all restaurants in New York state that look to be possibly Asian-owned. We would then sort so that the more likely Asian-Hispanic-owned operations would be grouped from the least likely Asian-owned operations. Using the female-owned likelihood as the next sort key, we would further modify the list so that the operations are listed in the following order: (1) likely Asian-Hispanic-female-owned operations, (2) likely Asian-Hispanic-male-owned operations, (3) likely Asian-non-Hispanic-female-owned operations, and (4) Asian-non-Hispanic-male-owned operations. Using the distribution of the probability of Asian-owned for the cases not selected with certainty, we can determine this stratum's optimal sample rate for a fixed sample. Finally, we take a systematic sample of the noncertainty cases at the optimal rate.

Notice that all of this work is based on the stratum's racial, gender, and Hispanic composition of each company. The design usually ensures a sample that is representative for the accurate estimation of each race, Hispanic, and gender group within the stratum. However, it is lacking in several ways. First, except for selecting a few cases with exorbitant receipts, payroll, or employment with certainty, very little work is done to ensure that the selected noncertainty units weight up to the stratum total. It is possible that the largest-valued noncertainty cases are selected in some strata, and the smallest-valued cases in others. Second, the selected noncertainty units may weight up to the stratum totals for the key variables, but they do not weight up to reasonable estimates for some of the published sub-domains (e.g., stratum receipts of Asian-owned, Hispanic-owned, or female-owned). Third, it is possible that (on occasion) some of the key sub-domains are under-sampled or over-sampled. For example, suppose the likelihoods estimate that there are 50 female-owned businesses in the noncertainty portion of a stratum. The sampling rate for the stratum is 1 in 5. Suppose one estimates that only 8 female-owned businesses were selected. If the weights of all selected cases remained at 5, then we would make the assumption that our sample would underestimate the number of female-owned businesses in this stratum by 20 percent. These are three good reasons for using weight calibration adjustment.

5 The Weight Calibration

Not all cases had their weights calibrated. The following describes the criteria used to select cases for calibration.

1 Only cases which were selected, but not selected with certainty were used. Cases selected with certainty were tabulated with weight equal to 1.00 in all calculations.

2 If the sum of the unweighted likelihoods which correspond to the frame assignment (over all cases selected with non-certainty for a particular stratum) was greater than 3.00, then all selected cases, not selected with certainty, in that stratum were subjected to calibration adjustment. If the sum was less than 3.00, then NO cases were calibrated. For example, if the stratum was White-Hispanic-owned restaurants in Alabama $\sum$LIKE HISP < 3.00 for all cases selected, then NO calibration was performed.

The first criterion guaranteed that self-representing cases remained self-representing. The second criterion prevented calibration for strata in which the expected number of minority-owned businesses from the sampled cases are extremely small. This usually leads to extremely unstable adjustment factors. These criteria limited the calibration to about 840,000 (or 89 percent of the noncertainty units) selected cases in 9,179 noncertainty strata.

The SAS non-linear programming routine PROC NLPNMS, based on the Nelder Mead Simplex method, was used to determine the adjustment factors. Two adjustment factors ("a_{hk}" and "b_{hk}" means the adjustment factor for the k-th unit in the h-th stratum) were calculated. The "a_{hk}" attempted to calibrate firm count, while the "b_{hk}" attempted to simultaneously calibrate receipts, payroll, and employment.

The "a_{hk}" are constructed from the following 6 constraints.

1 $a_{hk} \geq 1/ w_h$, so that no adjusted weight was less than 1.00.

2 $a_{hk} \leq 5.00$, so that no case dominated the adjustment.

3 Stratum counts balance:

$$N_h = \sum_{k=1}^{n_h} a_{hk} * w_h$$

4 Stratum Race-owned estimated counts balance:

$$\sum_{i=1}^{N_h} LIKERACE_{hi} = \sum_{k=1}^{n_h} a_{hk} * w_h * LIKERACE_{hk},$$

where race is the predominant race in the stratum.

5 Stratum Hispanic-owned estimated counts balance:

$$\sum_{i=1}^{N_h} LIKEHISP_{hi} = \sum_{k=1}^{n_h} a_{hk} * w_h * LIKEHISP_{hk}.$$

6 Stratum Female-owned estimated counts balance:

$$\sum_{i=1}^{N_h} LIKEFEM_{hi} = \sum_{k=1}^{n_h} a_{hk} * w_h * LIKEFEM_{hk}.$$

The "b_{hk}" attempt to satisfy Constraints 1 and 2 above as well as the following 12 constraints below.

7 Stratum aggregate receipts balance:

$$\sum_{i=1}^{N_h} RCTS_{hi} = \sum_{k=1}^{n_h} b_{hk} * w_h * RCTS_{hk}$$

8 Stratum Race-owned aggregate receipts balance.

9 Stratum Hispanic-owned aggregate receipts balance.

10 Stratum Female-owned aggregate receipts balance.

11 Stratum aggregate payroll balances.

12 Stratum Race-owned aggregate payroll balances.

13 Stratum Hispanic-owned aggregate payroll balances.

14 Stratum Female-owned aggregate payroll balance.

15 Stratum aggregate employment balances.

16 Stratum Race-owned aggregate employment balances.

17 Stratum Hispanic-owned aggregate employment balances.

18 Stratum Female-owned aggregate employment balance.

Since the sample was designed so that firm counts usually are balanced, the firm count calibration adjustment, "a_{hk}", is usually not very interesting. The remainder of the paper will concentrate on the effect of the auxiliary variable calibration adjustment, "b_{hk}".

Calibration using a similar set of constraints was performed on the 1997 SBO response data (see Moore, 2002). In general, the calibration worked well. However, there were several situations where the data did not calibrate as well as expected. These situations were identified in the 2002 SBO universe and special precautions were taken where necessary. One of the major problems occurred when the number of constraints exceeded the number of observations. All calibration was performed at the stratum level. If a stratum contained 4 to 7 cases selected with non-certainty, only the 4 receipts constraints (in addition to Constraints 1 and 2) were used to find the calibration adjustments. These adjustments were then applied to estimate stratum totals for payroll and employment.

Unfortunately, as Table 1 below shows, this did not work well. Although receipts at the stratum level did calibrate quite well. The other variables, under calibrated weighting actually increased their difference from the true stratum total. Consequently, it was decided that four observations were insufficient for calibration of the three variables.

Table 1. Relative Absolute Differences Averaged Over 9,179 Strata (About 840,000 of the Cases Not Selected With Certainty)

4 or more Noncerts	Receipts	Payroll	Employment
Original Wt	8.5%	8.0%	11.6%
Calibrated	0.2%	13.1%	19.3%

Therefore, the minimum number of selected cases but not selected with certainty was increased to 8. This limited us to 5,802 strata. Although we lost 3,377 strata, we only lost about 17,000 cases in these strata. Consequently, the 5,802 strata contain about 87 percent of the cases not selected with certainty. If calibration showed measurable gains on this subset of cases, it would indicate gains for a large portion of the total receipts, payroll, and employment for the universe.

As with the constraint methodology used for four or more noncertainty observations, if a stratum had 8 to 11 noncertainty cases, then both receipts and payroll constraints were used (in addition to Constraints 1 and 2). The calibration adjustment was applied to estimate aggregate employment. (If a stratum had 12 or more observations, all 14 constraints were used.) Table 2 shows that we get 6 to 7 percentage point improvement in accuracy for all three variables, when we restrict to strata with at least 8 observations. We were fortunate, in that employment is well-correlated to payroll. Had this not been the case, we may have had to restrict calibration to strata with at least 12 cases not selected with certainty.

Table 2. Relative Absolute Differences Averaged Over 5,802 Strata (About 823,000 of the Cases Not Selected With Certainty)

8 or more Noncerts	Receipts	Payroll	# Paid Employees
Original Wt	6.3%	5.9%	8.7%
Calibrated	0.1%	0.0%	2.0%

6 Random Group Variance Estimator

Regardless of the guesses assigned to each selected case, each is tabulated according to the actual response of that case. For example, a selected company may have administrative information that indicates that it is Asian, non-Hispanic, Male-owned, so it would be stratified with other Asian-owned firms. Suppose this company were to reply that it was Black, Hispanic, Female-owned. The company would be tabulated as Black-owned. It would probably be tabulated with a substantial number of companies whose race was guessed correctly. These would be stratified with other likely-Black-owned companies. Since sampling rates vary by strata, the company mis-classified as Asian-owned would very likely have a different sampling rate, than the cases correctly classified as Black-owned. Since there are 9 possible race-Hispanic-gender stratum identifiers, it is conceivable that many estimates will have components with several different sampling weights.

Since the components of each estimate can come from several strata, it is operationally more feasible to use a replication method for variance, instead of the design-based formula. Random group is widely used in the programs of the Economic Directorate at the U.S. Census Bureau, so it was selected for the SBO.

Theoretically, every unit's weight in a calibrated adjustment may not be equal to another unit's weight. Any design-based formula would have several components — one for the equal weight within stratum design, a second for the variation of the calibration within that stratum. Things would be further complicated, when you factor in the cases sampled in one stratum but tabulated in another. Consequently, random group replication is a fast and simple way to get a variance estimate under calibration.

7 Results

Table 2 has shown that our calibration method significantly corrects for the sampling error associated with selecting a sample that does not accurately estimate the auxiliary variables well. For the variables studied, the increase in "accuracy" is about 6 to 7 percentage points (for 5,802 strata encompassing about 87 percent of the cases selected, but not selected with certainty). However, is this increase in accuracy accompanied by a significant change in the estimate of the variance? Table 3 compares the coefficients of variation (cv's) of the random group estimates associated with equal weighting versus those associated with the calibration adjustment. As one can see, the cv's increase by about 1 percentage point.

Table 3. Coefficients of Variation Effects of Random Group Estimates When Combined With Calibration Average Over 5,802 Strata

CV's --- 8 or more Noncerts	Receipts	Payroll	# Paid Employees
Original Wt	15.3%	15.1%	18.6%
Calibrated	16.1%	15.6%	19.9%

8 Conclusion

When one attempts to estimate multiple characteristics with a single sample, the sample often incorrectly estimates several of the characteristics by a substantial amount. Used judiciously, calibration can correct estimates for a small number of strata. However, the stratum estimates that are correctable often represent a large percentage of the total estimate for each characteristic.

In our study of companies with paid employees, we were only able to accurately adjust the characteristic estimates for about 5,802 strata. However, this accurately reflected about 87 percent of the cases selected, but not selected with certainty. Also, no weight adjustments were performed on the self-representing (weight = 1.00) cases, which had large receipts, large payroll, or high employment. We were able to obtain more accurate stratum estimates for receipts, payroll, and employment using calibration than with an equal weighting within stratum scheme.

When restricted to this set of 5,802 strata, Table 2 showed that calibration reduced the relative absolute difference from the true stratum total by about 6 percentage points for each variable. When we combined calibration and random group replication on this set of strata, we increased the variance by about 1 percentage point. Consequently, calibration gives us a more accurate estimate of the true stratum mean without a significant increase in variance.

9 Future Research

It can be shown that the design-based variance within a stratum can be written in the form $V_{DB}(\hat{Y}_i)$, with

$$V_{DB}(\hat{Y}_i) \approx \sum_{k=1}^{n} w * (w - 1) * (y_{ik} - \overline{y_i})^2 \quad .$$

In these formulas,

N = the number of noncertainty units in a given stratum,
n = the number of sampled units from the stratum,
y_{ik} = the value of variable i for the k-th selected unit,
$\overline{y_i}$ = mean for variable i over the stratum,
w = sampling weight (N/n).

Since the calibration factors are weight adjustments, one could naturally ask, " When using calibrated weights is a good estimate of the variance $GVF(\hat{Y}_i)$,

$$GVF(\hat{Y}_i) = \sum_{k=1}^{n} (b_{ik}*w) * (b_{ik}*w - 1) * (y_{ik} - \overline{y_i})^2 \quad ?$$

We have done some very preliminary research into this. So far results do not look promising. This leads to several conjectures.

1 Are there any conditions that we can place on the calibration factors that would guarantee good results? For example, suppose the calibration factors are not well correlated to the actual data points (i.e, Corr (b_{ik}, y_{ik}) $\approx$ 0).

2 Do the calibration results look as promising, when you add the complexity of tabulating an estimate, where the components have different original sampling weights?

10 References

U.S. Office of Management and Budget (1998). North American Industry Classification System Manual. Bernan Press, Lanham, MD.

Cochran, W.G. (1977). Sampling Techniques. John Wiley and Sons, New York, NY.

Moore, Richard A. (2002). Using a SAS/IML Nonlinear Programming Procedure to Determine a Single Uniform Weighting Scheme For a Complex Survey Design. Proceedings of the Tenth Annual Conference of the South East SAS Users Group (Sept. 2002). Savannah, GA. Pages 338-343.

Sarndal, Carl-Eric and Victor Estevao (2000). A Functional form Approach to Calibration. Journal of Official Statistics, Vol 16, No. 4 (Dec. 2000). Stockholm, Sweden.

Sarndal, C. E., Swenson, B., and Wretman, J. (1992). Model-Assisted Survey Sampling. Springer-Verlag, New York, NY.

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