A BOOTSTRAP STRATEGY FOR ENHANCING THE CONFIDENCE INTERVAL ESTIMATION OF SMALL SAMPLES

Chih-Chin Ho, Internal Revenue Service
1111 Constitution Avenue, Washington D.C 20224

Key Words: Realized coverage rate, Expected coverage rate, Calibration rate, Calibrated average range

This paper applies a bootstrap resampling method to enhance the estimation efficiency and coverage sufficiency of confidence intervals (CIs) for small samples from a skewed population. Normal distribution theory CIs for the original (full-size) samples are compared with bootstrap CIs for the reduced (smaller-size) samples. Such comparisons shed insights into how well a bootstrap strategy can compensate for the loss of efficiency and coverage sufficiency of the confidence interval due to a reduction in sample size.

1. BACKGROUND

The IRS has conducted a series of surveys in the Taxpayer Compliance Measurement Program (TCMP). In recent years, the IRS has become concerned about the cost associated with the surveys and has considered a reduction in sample size.

The classical statistical theory dictates that a reduction in sample size would increase the variance and result in the loss of efficiency in the confidence interval estimation, other things being equal. Furthermore, such a reduction in sample size may reduce the sufficiency in covering the population mean by confidence intervals.

It has been argued that given the rather large size of TCMP samples[1], a reasonable size reduction would not threaten the coverage sufficiency of confidence intervals. This view may be correct in the context of national aggregate. However, if we focus on selected segments of the national aggregate, we often encounter much smaller samples. A size reduction from these inherently small samples of selected market segments would further undermine the coverage sufficiency of the estimated confidence intervals[2].

This study was originated to find out to what extent a size reduction can be compensated by the enhanced estimation efficiency and coverage sufficiency of the bootstrap resampling method in estimating confidence intervals.

2. DATA

The IRS's Individual Return Transaction File (IRTF) was the source of the data from which two sets of 100 samples of different sizes were randomly drawn. The selected measure used in this study is the tax liability reported on the return.

2.1 Population

The population consists of 38,000 individual tax returns filed in the IRS San Francisco District with Schedule C (Business) total gross receipts (TGR) of more than $25,000 but less than $100,000 for tax year 1989.

2.2 Original Sample Set

This sample set consists of 100 independent random samples of size 52 from the population. These samples are labeled as A1, A2, ...A100. The choice of 52 is based on the sample size of the same stratum in the TCMP individual survey on tax year 1988 returns.

2.3 Reduced Sample Set

This sample set consists of 100 random subsamples of size 47, one from each of the 100 full-size samples. These samples are labeled as B1, B2,...B100. The reduced samples represent about 10 percent reduction in size from the original samples.

3. METHODS

A bootstrap resampling method and normal distribution theory were used to estimate CIs for the reported tax liability for both sample sets.

3.1 Bootstrap Replicates

The Mccarthy and Snowden (1985) "with-replacement bootstrap" method was used to create bootstrap replicates. Ignoring the finite population correction, this method used with replacement samples of size n-1 from the original sample for each bootstrap replicate[3].

One thousand bootstrap replicates were selected independently for each of the 100 samples in the selected sizes. For example, for sample A1, the 1,000 bootstrap replicates randomly selected are A1b1, A1b2, ... A1b1000, or for sample B100, the 1,000 bootstrap replicates are B100b1, B100b2, ...B100b1000.

3.2 Bootstrap Confidence Interval

The bootstrap confidence interval (BT) for a particular sample can be readily observed from the nonparametric distribution of the means of its 1,000 bootstrap replicates. For a two-sided 68% BT, the 160th and 841st bootstraps are the lower and upper bound, respectively. For a two-sided 95% BT, the 25th and 976th bootstraps are the lower and upper bound, respectively[4].

3.3 Normal Distribution Confidence Interval

The normal distribution confidence interval (ND) for a particular sample can be observed from the sample mean and standard deviation (sigma) estimates. For a two-sided 68% ND, the sample mean $\pm$ 1 sigma are the upper and lower bound, respectively. For a two-sided 95% ND, the sample mean $\pm$ 1.96 sigma are the upper and lower bound, respectively.

4. MEASURES

For a particular sample set and by a particular estimation method, the average range of estimated CIs was used as a measure of efficiency. Coverage properties of estimated CIs were calculated to provide a measure of sufficiency and to form a basis of calibration for efficiency.

The average range of a set of estimated CIs calibrated by coverage sufficiency was used to measure efficiency after the sufficiency adjustment.

4.1 Unadjusted Average Range

The unadjusted average range of a set of estimated CIs based on a particular method provides a measure of estimation efficiency. For example, if the average BT range is **narrower** than the average ND range for a particular sample set given a prescribed level of confidence, it would indicate that the bootstrap method is more **efficient** than the normal distribution method in the CI estimation, holding other things constant[5].

4.2 Realized Coverage Rate

The realized coverage rate (RCR) is defined as a percentage by which a set of estimated CIs actually cover the population mean given a prescribed level of confidence. For example, a 70% RCR means that in 70 out of 100 samples, the estimated CI actually cover the population mean.

The RCR provides a measure of coverage sufficiency. For example, if the ND's RCR is **greater** than the BT's RCR for a particular sample set given a prescribed level of confidence, it would indicate that the normal distribution method is more **sufficient** than the bootstrap method in the CI estimation, holding other things constant.

4.3 Expected Coverage Rate

The expected coverage rate (ECR) is defined as a percentage by which a set of estimated CIs should have covered the population mean given a prescribed level of confidence. For example, the ECR is equal to 68% for a set of estimated 68% CIs, which means that these CIs should

have covered the population mean 68 times in 100 cases.

The ECR provides a benchmark to assess coverage sufficiency. For example, if the RCR is **equal** to the ECR given a prescribed level of confidence, it would reflect that the CIs actually covered the population mean as **often** as they should have.

4.4 Calibration Rate

The calibration rate is a factor by which the estimated CIs need to be adjusted to equalize the realized coverage rate with the expected coverage rate. The factor adjustment is based on the standard normal distribution function. The calibration rate (CR) is defined as:

$$CR = Z^{-1}([1+ECR]/2)/Z^{-1}([1+RCR]/2)$$

where Z^{-1} is the inverse of the standard normal distribution function.

The CR forms a basis of calibration for estimation efficiency upon coverage sufficiency. For example, when a set of 95% CIs actually covered the population mean **exactly** 95 times out of 100 cases (i.e., CR=1), it would indicate that **no adjustment** is necessary for these CIs' estimation efficiency based on their coverage sufficiency.

4.5 Calibrated Average Range

The calibrated average range is equal to the unadjusted average range multiplied by the calibration rate.

The calibrated average range provides a measure of estimation efficiency after the adjustment of coverage sufficiency. For example, if only 64 out of 100 samples had the population mean covered by its estimated 68% CI, then the unadjusted average range needs a **upward** adjustment to reflect the **less than sufficient** coverage (i.e., CR>1)[6]. In this instance, the calibrated average range would be **wider** than its unadjusted counterpart.

Similarly, when a set of 95% CIs actually covered the population mean 96 times out of 100 samples, the unadjusted average range needs a **downward** adjustment to reflect the **more than sufficient** coverage (i.e., CR<1)[7]. In this instance, the calibrated average range would be **narrower** than its unadjusted counterpart.

5. RESULTS

Tables 1-4 present the unadjusted average ranges, the realized coverage rates, the calibration rates, and calibrated average CI ranges, respectively, at both 68% and 95% confidence levels, by sample size and estimation method.

Table 1
Unadjusted Average Range
at 68% & 95% Confidence Level

Sample Size	ND	BT
68% CI		
52	$1,701	$1,671
47	$1,788	$1,739
95% CI		
52	$3,334	$3,284
47	$3,504	$3,362

Table 2
Realized Coverage Rate
at 68% & 95% Confidence Level

Sample Size	ND	BT
68% CI		
52	64%	72%
47	62%	68%
95% CI		
52	94%	96%
47	93%	95%

Table 3
Calibration Rate
at 68% & 95% Confidence Level

Sample Size	ND	BT
68% CI		
52	1.087	0.926
47	1.136	1.000
95% CI		
52	1.037	0.951
47	1.077	1.000

Table 4
Calibrated Average Range
at 68% & 95% Confidence Level

Sample Size	ND	BT
68% CI		
52	$1,849	$1,547
47	$2,031	$1,739
95% CI		
52	$3,457	$3,123
47	$3,774	$3,362

6. FINDINGS

6.1 Comparison Within Sample Size And Estimation Method

o Both the unadjusted and calibrated average ranges are **narrower** for the bootstrap CIs within the same sample size.

o The realized coverage rate is **higher** and the calibration rate is **lower** for the bootstrap CIs within the same sample size.

o Both the unadjusted and calibrated average ranges are **wider** for the reduced samples based on the same estimation method.

o The realized coverage rate is **lower** and the calibration rate is **higher** for the reduced samples based on the same estimation method.

6.2 Comparison Across Sample Size And Estimation Method

o The unadjusted average range of the reduced samples based on the bootstrap method is **wider** than the corresponding measure of the original samples based on normal distribution theory.

o The realized coverage rate of the reduced samples based on the bootstrap method is **higher** than the corresponding measure of the original samples based on normal distribution theory.

o The calibration rate of the reduced samples based on the bootstrap method is **lower** than the corresponding measure of the original samples based on normal distribution theory.

o The calibrated average range of the reduced samples based on the bootstrap method is **narrower** than the corresponding measure of the original samples based on normal distribution theory.

7. CONCLUSIONS

The bootstrap method performs better in both estimation efficiency and coverage sufficiency in comparison with the normal distribution method within the same sample size. The bootstrap strategy has enhanced efficiency and sufficiency in the CI estimation.

The original samples perform better in both estimation efficiency and coverage sufficiency in comparison to the reduced samples based on the same estimation method. The 10% reduction in sample size has diminished efficiency and sufficiency in the CI estimation.

Although the bootstrap method has not improved estimation efficiency with the reduced samples enough to compensate for the loss of efficiency due to the 10% size reduction, its enhancement for coverage sufficiency has been

sufficient enough to gain a higher calibrated efficiency than what has been obtained with the original samples based on the normal distribution method.

In conclusion, the bootstrap procedure enhanced the estimation efficiency of confidence intervals for the reduced samples, after calibrated by coverage sufficiency. For both 68% and 95% confidence levels, the loss of estimation efficiency and coverage sufficiency due to the 10 percent reduction in sample size (from 52 to 47) can be compensated by using the bootstrap procedure in estimating confidence intervals.

8. FUTURE RESEARCH

In the future, we may apply the bootstrap strategy to different populations and measure the impacts of selected population characteristics on the effectiveness of bootstrapping. In addition, we may apply the strategy to larger sample sizes and measure the impacts of the size of the original sample on the effectiveness of bootstrapping.

Finally, we may also want to repeat this experiment to determine the variability of the calibration rates.

9. ACKNOWLEDGEMENTS

The author would like to thank Don Wilt for his review and valuable comments, and Dennis Cox and Shien Perng for their stimulating discussions. My special thanks go to William Wong for his valuable suggestions on methodology. I also thank Mary-Helen Risler for her assistance in preparing the publication.

NOTES

[1] For example, the most recent (1988) TCMP survey of individual returns is a national random sample of approximately 50,000 returns drawn from the population of filed tax returns.

[2] Since the underlying population in a particular market segment may not be normally distributed and the sample may be too small for the central limit theory to be useful, the resulting CI estimates based on normal distribution theory may not be sufficient in covering the population mean.

[3] This method was used in Wong and Ho (1991). For more detailed discussions, see Rao and Wu (1988) and Sitter (1990).

[4] The intervals need to be adjusted negligibly upwards to account for "small sample variability of the ranks." See Wong and Ho (1991).

[5] We assume, for example, the variance of the estimated CIs is small.

[6] In this case, ECR=0.68 and RCR=0.64; $Z^{-1}=([1+0.68]/2)=1$, $Z^{-1}=([1+0.64]/2)=0.92$. Therefore, CR=(1/0.92)=1.087.

[7] In this case, ECR=0.95 and RCR=0.96; $Z^{-1}=([1+0.95]/2)=1.96$, $Z^{-1}=([1+0.96]/2)=2.06$. Therefore, CR=(1.96/2.06)=0.951.

REFERENCES

McCarthy, P.J. and Snowden, C.B. (1985), "The Bootstrap and Finite Population Sampling," in Vital and Health Statistics, ser.2, no.95, Public Health Service Publication 85-1396, Washington, DC: U.S. Government Printing Office.

Rao, J.N.K. and Wu, C.F.J. (1988), "Resampling Inference With Complex Survey Data," Journal of the American Statistical Association, vol. 83, no 401, pp. 231-241.

Sitter, R.R. (1990), "Comparing Three Bootstrap Methods for Survey Data," Technical Report 152, Dept. of Mathematics and Statistics, Carleton University.

Wong, W. and Ho, C. (1991), "Bootstrapping Post-Stratification And Regression Estimates from A Highly Skewed Distribution," in American Statistical Association 1991 Proceedings of the Section on Survey Research Methods, pp. 608-613.