BOOTSTRAPPING POST-STRATIFICATION AND REGRESSION ESTIMATES FROM A HIGHLY SKEWED DISTRIBUTION

William Wong and Chih-Chin Ho, Internal Revenue Service

Key Words: Variance, Confidence Intervals, Modeling

Random samples from highly skewed distributions are apt to yield volatile results. The usual symmetric two sigma confidence intervals would not apply since the distribution of sample results is likely to be skewed. This paper illustrates the volatility and skewness of sample estimates from one such distribution. To reduce the volatility, where possible, a regression model for the population was calculated from the sample. Where the regression model did not apply, post-stratification was used. Bootstrap sample estimates of both totals and ratios were employed to analyze the distribution, quantify the results, and compute confidence intervals.

We begin with some background on the data set to be examined and a description of the original estimation approach. Next, some initial attempts at improvement are described. This is followed by a discussion of how bootstrap samples were generated. Selection of the regression model, creation of estimators, and treatment of outliers are also briefly explained; then, some results are presented. Finally, we conclude with some ideas for future research.

BACKGROUND

The IRS conducts a series of surveys in the Taxpayer Compliance Measurement Program (TCMP). The 1986 study of tax-exempt organizations raised an interesting problem.

The population consisted of a class of 28,500 tax returns filed in 1987 and 1988, covering tax periods that include December 1986. The population was stratified into seven classes of tax-exempt organizations based on the 1954 Internal Revenue Code, Section 501(c). Each strata was then stratified into two income classes. Within organization type by income stratum, a probability sample was selected. To select the probability samples, we employed a method of selecting intervals of transformed taxpayer identification numbers developed by Harte (1986). Using these intervals, instead of fixed sample sizes, caused variability in the resulting sampling rates and sample sizes. However, considering the small sample sizes, and in the absence of a more extensive analysis, this variability is considered to be relatively minor. The primary statistic of interest was the ratio of the total reported tax divided by the total corrected tax (the value as determined by Internal Revenue Service's Examination Office). In the original estimation approach, the usual sample-weighted combined ratio estimator was calculated along with two sigma confidence intervals, using Cochran's standard linearized variance estimator (Cochran, 1977). Since the population was known to be skewed, both the sample estimates and the confidence intervals were suspect.

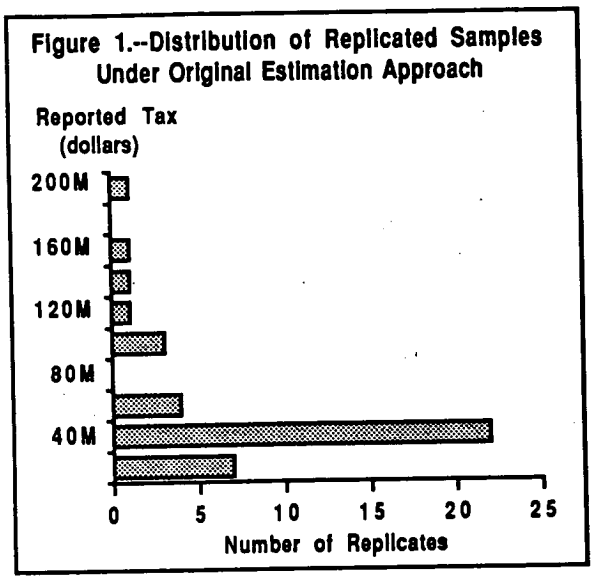
Approximately a year after the sample was selected, a population file containing the reported tax for all returns became available. We sought to improve the reliability of our estimates, measure their quality, and calculate confidence intervals. As a first step, we compared the sample-weighted total reported tax with the population file sum of the reported tax. For three of the seven tax-exempt organization strata, the sample-weighted estimate was only half of the population sum. Further investigation revealed that the primary causes of the discrepancies were (in descending order of severity):

- sampling variability;
- early sampling closeout;
- coding and processing differences, between the sample and the population, for both the tax-exempt organization codes and tax amounts; and
- out-of-scope tax returns and other differences.

To determine the effect of sampling variability on the estimator, we replicated 40 samples from the population file, using the original stratification and sampling intervals of transformed taxpayer identification numbers. We defined the 40 transformed taxpayer identification number intervals in a cyclical manner, in an attempt to evenly cover the entire population. This resulted in the reported tax distribution given in Figure 1 and the following reported tax statistics:

Population sum	58.9 million
Original weighted sample	39.6 million
Mean of the weighted replicated samples	62.8 million
Median of the weighted replicated samples	45.4 million.

Obviously, the distribution of the replicated sample estimates is skewed. It follows that using symmetric two sigma confidence intervals for the



total reported tax would be inappropriate. Using symmetric two sigma confidence intervals for the ratio of the total reported tax over the total corrected tax is, therefore, questionable. Distributions for individual classes of tax-exempt organizations were more variable but similar.

The second cause of the discrepancies between the population values and the weighted sample estimates was the early sample cutoff. Due to the detailed and expensive nature of calculating corrected tax, sample selection closed out a little over a year after all the tax returns were required to be filed. Based on the population file, five percent of the returns had not been filed yet and another five percent were still being processed and unavailable for sampling. Consequently, up to ten percent of the discrepancy between the original sample-weighted estimates and the population sums can be attributed to the early cutoff. The two remaining causes: out-of-scope returns and other processing and coding differences, accounted for less than five percent of the discrepancy. Further details of the purpose of the study and the results are given in Cox (1991) and Nunns (1991).

EXPLORATORY ATTEMPTS AT IMPROVEMENT

To improve the estimates we first tried using post-stratification. We ordered the population by reported tax within tax-exempt organization strata and defined our tax class post-strata within each stratum. Each sample unit was then either given a new weight equal to the stratum population count divided by its sample count; or had its original weight adjusted by a factor equal to the stratum population count divided by its weighted sample count. Both methods yielded similar results and showed no significant improvement over the original estimator. Once we recognized the severity of the skewness, we tried using regression to estimate the corrected tax for the largest 100 returns in each organization class stratum. We decided to use 100 returns because, in most strata, more than 80 % of the tax was paid by them. Using both reported tax and regression-estimated corrected tax values for the largest 100 population returns, we could form certainty strata and calculate improved estimates of the ratio of the reported tax divided by the corrected tax. The variability of the ratio estimates would be sharply reduced, since the new estimator makes skewness an advantage instead of a disadvantage. We had two remaining problems: how good was the new estimator and, having rejected using two sigma confidence intervals, how do we construct confidence intervals around the new estimates?

ANALYSIS METHODOLOGY

Creating Bootstrap Samples

We wanted a simple method to calculate confidence intervals. Most replication methods are simple to implement. The question now is which method is likely to yield the best results when regression modeling is used? After reviewing the literature (Efron and Gong, 1983; Rao and Wu, 1988; Sitter, 1990a, 1990b), we decided to use bootstrapping. Of the bootstrapping methods, Sitter's "mirror matching" approach, in theory, seemed to yield the best estimates. Implementation of McCarthy and Snowden's (1985) "with-replacement bootstrap" appeared to be simpler. They suggested using bootstrap stratum sample sizes of

$(n-1)/(1-f)$,

where n is the original sample size, and

$f = n/N$ is the the finite population correction factor.

Ignoring the finite population correction, this reduces to using bootstrap stratum sample sizes of $n(h)-1$ for stratum (h).

Thus, we selected 400 bootstrap samples independently for each tax-exempt organization stratum, as follows:

- For each of the two original income sampling strata (h), we obtained the original sample and selected from it a with-replacement sample of size $n(h)-1$. If a particular sample return was selected m times, then we made m duplicate copies of that return.
- Each bootstrap now consisted of $n(1)-1$

sample returns from stratum (1) and n(2)-1 sample returns from stratum (2). We repeated this procedure 400 times.

In addition to the 400 bootstrap samples, we created an "all data" sample replicate, to provide a reference point for our modeling and analysis.

Selection of the Regression Model

Before bootstrapping the regression-modeled estimators, we needed to determine the general regression model. First, we plotted the variable to be modeled -- corrected tax -- against each of the variables on the population file. The plots against both of the two best regressors showed a linear relationship except for a spike at zero. All stepwise regressions yielded dismal R-squares of less than .5 because of the spike. Since our primary interest was to model the corrected tax for the high tax returns, those returns with zero reported tax (i.e., the spike) were removed prior to the regression. This modification resulted in R-squares of around .9 or better. R-squares of .9 were achieved by using reported tax as the sole regressor. Including other (correlated) regressors made little additional improvement. Thus, we decided to use reported tax as the sole regressor. A further analysis showed that the relationship was basically linear, so higher order terms were not needed. Our final model was:

$y = a + bx + e$,

where: y = predicted value of the corrected tax,
x = reported tax value on the population file,
a,b = regression coefficients, and
e = random noise added back (explained below).

Technically, the model is $y(r,x) = a(r) + b(r)\,x + e(r,x)$, since the regression coefficients vary by bootstrap replicate, r, and the random noise, e, varies randomly with each bootstrap and tax return.

More specifically, to create the 400 bootstrap samples, the procedure was:

- Using the organization stratum-by-stratum plots of all the sample returns with positive reported tax, we predetermined, across all the bootstraps, which sample units, if hit, would be considered an outlier. Thus, we ensured that the regression R-squares and models were properly specified. In doing so, we ignored the original sample weights.
- Next, we determined which sample units (with duplication) were in the bootstrap, had positive reported tax, and were not outliers.
- With these units, we used unweighted ordinary least squares to calculate bootstrap regression coefficients, R-squares, and Root Mean Square Errors.
- For each stratum, we then analyzed the R-squares, RMSE's, and plot the predictors and the residuals against the regressor for the first five bootstraps. Where necessary, we redefined the outliers and ran another set of five bootstraps, before processing the full set of 400 bootstraps.

For each of the 400 bootstrap samples, we then generated a separate regression model and calculated the predicted corrected tax by applying the bootstrap regression estimated coefficients to the population file reported tax and adding back the random noise. The random noise was a normal variate distributed Normal(0,sigma(r)), where sigma(r) is the r-th bootstrap residual Root Mean Square Error from the regression. Having defined the general regression model, we could then calculate bootstrapped modeled estimates to determine the variability of our estimators, including the part due to regression modeling.

Selection of the Estimators

Before calculating bootstrapped-modeled estimates, the exact form of the estimator had to be determined. We experimented with developing different models from various tax class definitions and applying the resulting regression coefficients to a variety of tax classes. This was an attempt to build in some cross-validation. In many of the trials, the models fit poorly, as indicated by bootstrap R-squares of less than 0.5. In the final analysis, we abandoned our attempt at cross-validation and decided to generate the bootstrap regression models using all positive reported tax sample returns. For most of the tax-exempt organization strata the regression models fit quite well -- their R-squares were around 0.95.

After generating these models, we decided to apply the resulting regression coefficients to all population file returns that had positive reported tax. Thus, we obtained corrected tax estimates for the entire positive reported tax strata. For the zero reported tax strata, we decided on using a post-stratified estimator.

For the two smallest tax-exempt organization strata, Civic Associations and Fraternal Societies, only 25 and 18 positive reported tax sample returns were available to perform the regression. In neither case was the linear trend clear.

For Civic Associations, we decided to reject the regression model because it did not appear to improve the estimates. Instead, we post-stratified both the zero and positive reported tax strata and then ratio adjusted the estimates of the population-

reported tax totals. This adjustment involved multiplying the post-stratified estimates by the ratio of the total population-reported tax divided by the weighted sample-reported tax. Thus, for example, the bootstrap reported tax estimates always equaled the total population-reported tax.

For Fraternal Societies, we also tried hot-deck imputation of the corrected tax to reported tax ratios to the entire positive reported tax population. Since it had only 18 positive reported tax sample returns, the method proved too volatile and was rejected.

Treatment of Outliers

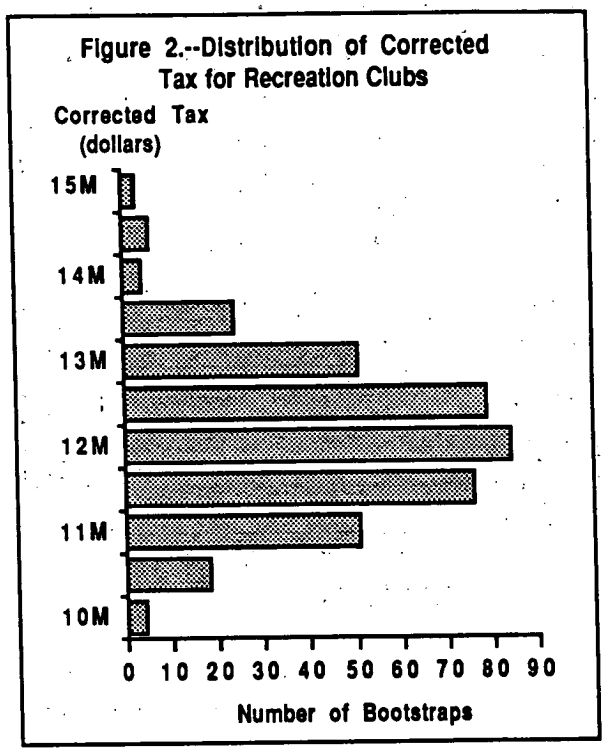
Since outliers were removed prior to calculating the regression coefficients, to avoid a bias, they had to be added back into the estimates. This was done as follows: If an outlier was hit m times in a bootstrap, then it was given a weight of m times its original sampling weight. In terms of corrected tax, the amount to be added back is m times its original sampling weight times the outlier's original corrected tax value. However, since the outlier's predicted corrected tax value is already imputed when the regression coefficients are applied to the population file of reported tax values, a value equal to m times the original outlier's sampling weight times the outlier's weighted predicted corrected tax must be subtracted back out. The reason for incorporating the outlier's original sample weight into this calculation is to reflect the notion that the outlier would sufficiently represent its presence in the population, based on its sampling weight.

RESULTS

Even with the extremely skewed distributions, the bootstraps were very well behaved. The corrected tax values and reported tax to corrected tax ratios appeared normally distributed for most of the tax-exempt organization strata. Figures 2 and 3 illustrate this typical behavior using Recreation Clubs as an example.

As you can see, for both corrected tax and reported tax / corrected tax, the skewness has been eliminated; the resulting distributions appear basically normal.

As mentioned earlier, the two exceptions are tax-exempt Civic Associations and Fraternal Societies. Once again, these two strata were difficult to model because they had very few positive reported tax sample returns. Though Fraternal Societies was the more stable of the two, it still had large bootstrap-to-bootstrap variation in its models, as exhibited by the large standard deviations of both the R-squares and regression slopes in Table 1 below. Large differences between the means and medians also confirm this.



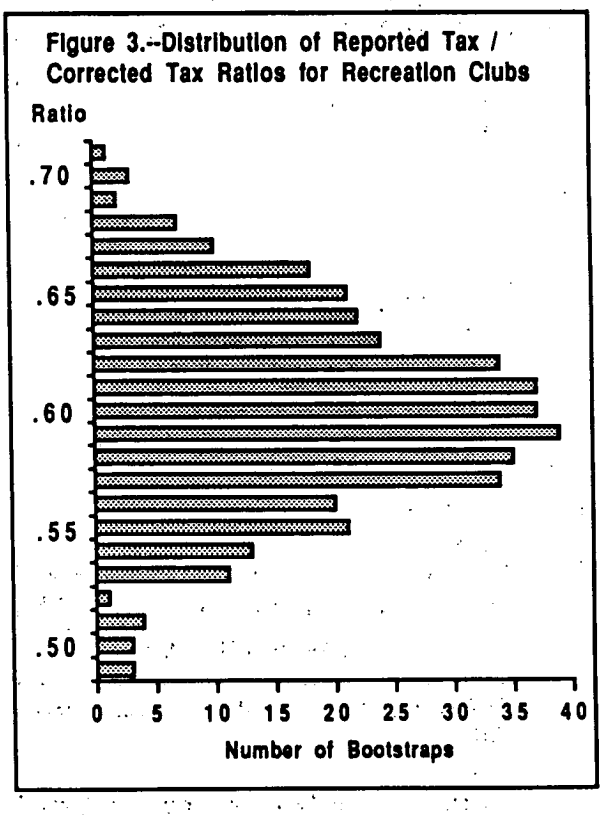


Table 1. Means, Medians, and Standard Deviations by Type of Organization

Type of Organization*	Statistic	Corrected Tax (Millions)	Reported Tax / Corrected Tax	R-Square	Regression Slope
Public Charities and Private Foundations	Mean	21.6	0.628	0.944	1.053
	Median	21.5	0.623	0.976	1.088
	Std Dev	2.5	0.071	0.060	0.060
Civic Associations**	Mean	21.3	0.656		
	Median	21.4	0.618		
	Std Dev	4.9	0.167		
Labor Unions	Mean	1.29	0.814	0.939	1.001
	Median	1.28	0.813	0.944	1.001
	Std Dev	0.11	0.067	0.031	0.018
Business Leagues	Mean	10.5	0.752	0.911	0.977
	Median	10.4	0.790	0.929	0.977
	Std Dev	1.1	0.084	0.059	0.047
Recreation Clubs	Mean	12.8	0.805	0.967	1.030
	Median	12.8	0.804	0.963	1.043
	Std Dev	0.9	0.041	0.035	0.025
Fraternal Societies***	Mean	0.48	0.973	0.755	0.676
	Median	0.45	0.908	0.901	0.818
	Std Dev	0.14	0.449	0.277	0.322
Other Tax-Exempt Organizations	Mean	35.1	0.987	0.999	1.003
	Median	35.2	0.984	1.000	1.002
	Std Dev	0.7	0.020	0.0004	0.004

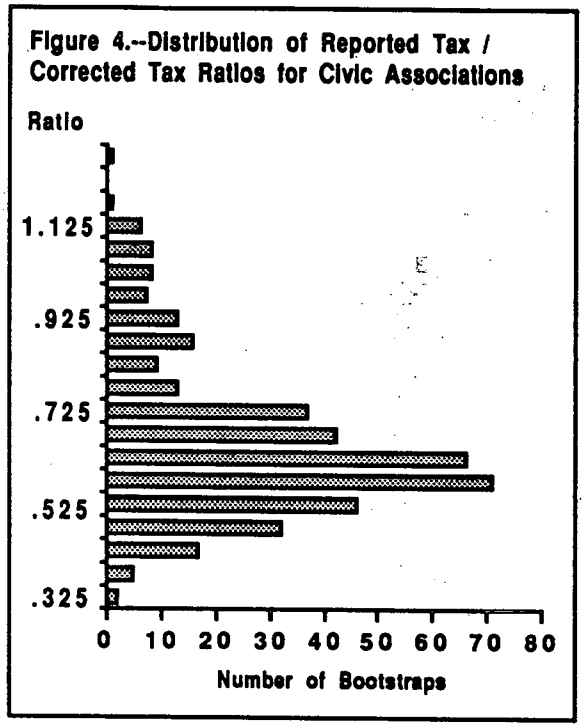
Notes:

* As defined under Section 501(c) of the 1954 Internal Revenue Code.

** Regression was not used for this stratum because of poor regression results; it consisted of only 25 positive reported tax sample returns.

*** Regression was weak in this stratum because the sample only contained 18 positive reported tax returns.

Figure 4 illustrates the distribution of bootstraps for one of these strata where regression is not used.



NOTE: Regression was not used for these tax- exempt organizations.

As you can see, like the behavior of the 40 replicate expansion estimates given in Figure 1, the distribution remained skewed.

Table 1 compares the results for each type of tax-exempt organization. Except for Civic Associations and Fraternal Societies, the standard deviations were small and the means were close to the medians. This further supports the distributional observation, above, that the highly skewed characteristics of the original population disappear when regression modeling is successful. R-squares were very high except for these two small strata. (In fact, R-squares for Civic Organizations were not included in the table because the initial set of 5 test bootstraps indicated the R-square values would be too low for regression to work in that stratum.) The regression slopes were all very close to 1.0 except, again, for Fraternal Societies.

Table 2 contains the upper and lower 21 bootstraps of the ratio for a typical stratum, Recreation Clubs. Confidence intervals can readily be obtained from it. For example, to obtain a 2-sided 90% confidence interval, we deleted the first 19 and last 19 bootstraps. (When using ranks to form confidence intervals, interpolation was necessary between adjacent ranks. A conservative alternative

Table 2. The Upper and Lower 21 Bootstrap Estimates of the Ratio for Recreation Clubs

Rank of Ratio	Reported Tax/ Corrected Tax (Ratio)	Rank of Ratio	Reported Tax/ Corrected Tax (Ratio)
1	0.493300	400	0.717576
2	0.494051	399	0.705756
3	0.497918	398	0.705500
4	0.504837	397	0.703166
5	0.506284	396	0.696012
6	0.509143	395	0.693306
7	0.510236	394	0.687880
8	0.511357	393	0.684161
9	0.513626	392	0.683777
10	0.515187	391	0.682952
11	0.520128	390	0.682867
12	0.530429	389	0.681948
13	0.531154	388	0.680935
14	0.533988	387	0.677732
15	0.535769	386	0.677004
16	0.536568	385	0.676304
17	0.537030	384	0.676019
18	0.537535	383	0.675293
19	0.537970	382	0.674822
20	0.538307	381	0.673437
21	0.538565	380	0.672385
196/197	0.602428	(All data)	

would have been to round to the next larger confidence interval. Also, the intervals need to be adjusted negligibly upwards to account for "small" sample variability of the ranks. They should be adjusted downwards, for "finite population correction.") For tax-exempt Recreation Clubs the "all-data" or entire sample replicate fell between rank 196 and 197. This was close to the median bootstrap estimate, as expected.

CONCLUSIONS AND AREAS OF FUTURE STUDY

In conclusion, the bootstrap procedure and regression modeling worked very well in improving the original estimates from highly skewed distributions. In the future, we may calculate a new set of bootstraps using sampling intervals of transformed taxpayer identification numbers and measure its effect on confidence intervals. Also, we may try to estimate the effect of the finite population correction on the bootstrapped-modeled confidence intervals. Finally, we would like to measure the basic model-to-basic model variation. One way this can be done is to do a simulation study by pretending the sample is the population, select repeated samples from it, calculate bootstrap confidence intervals from each, and examine their coverage properties. We expect to explore some of these options and hope to be able to report on them in the near future.

ACKNOWLEDGMENTS

The authors would like to thank Fritz Scheuren for suggesting the methodology, Wendy Alvey, Beth Kilss, and Karen O'Conor for their assistance in preparing the paper and its presentation, and Clementine Brittain and Nathan Shaifer for preparing the graphs and visuals.

REFERENCES

Cochran, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

Cox, Dennis (1991), "Preliminary Results From Statistics of Income Division and Research Division Joint Research on Unrelated Business Income Tax From Exempt Organizations," Working Paper, Internal Revenue Service, Washington, DC.

Efron, Bradley and Gong, Gail (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician,* vol. 37, no. 1, pp. 36-48.

Harte, James M. (1986), "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *American Statistical Association - 1986 Proceedings of the Section on Survey Research Methods,* pp. 603-608.

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 85-1369, Washington, DC: U.S. Government Printing Office.

Nunns, Jim (1991), "UBIT TCMP Study," internal memorandum to Ken Gideon on March 15, 1991, Treasury Department, Washington DC.

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. (1990a), "A Resampling Procedure For Complex Survey Data," Technical Report 149, Dept. of Mathematics and Statistics, Carleton University (preliminary publication).

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