IMPROVING THE STRATIFICATION OF THE BANK AND CORPORATION SAMPLE

Lawrence Gilbert, Franchise Tax Board
Research Bureau, P.O. Box 2229, Sacramento, California 95810-2229

KEY WORDS: Stratified Sampling, Optimum Allocation, Income Distribution

INTRODUCTION

This paper will describe the sampling design initially employed for the Franchise Tax Board's Bank & Corporation tax return sample. Following this will be a description of some modifications made to the basic scheme in order to improve the sample's reliability, and changes we are in the process of evaluating.

Nearly 450,000 corporations do business in California. When FTB's document processing bureau receives a corporation's tax return, certain key data items on the return are placed on the Bank & Corporation master file, or BCM. These items include (1) State Net Income, which is the proportion of the corporation's total income attributable to its presence in California, (2) a two-digit industry code indicating the type of industry the corporation's primary activity falls into, and (3) several other fields, primarily descriptive in nature but also including a few monetary items.

Because of the desirability of doing more in-depth study of corporate tax behavior and characteristics than the limited information on the BCM allows, it is necessary to select a sample and code a much larger number of data items (up to 250 per return). Because of resource limits, the sample size cannot exceed approximately 10,000-12,000.

DESCRIPTION

Originally, the B&C sample consisted of essentially two strata: large corporations, which were defined as those with state net income greater than + or - $5,000,000, and everyone else. These strata were defined by statute. For the sample, we selected all large corporations and 2% of the remaining ones. Other small groups of taxpayers were sampled at the 100% rate -- those who paid large amounts of business license or personal property tax (these were required for the computation of a now-expired bank tax and are no longer part of the sample) and those multi-national corporations that elected to file on a water's edge basis; i.e., using only their domestic income for the denominator of the formula which determines their SNI. This constitutes the 3-strata design (large corporations, small corporations, and Water's Edge corporations).

It became apparent from our analysis needs that we needed more accurate statistical estimates for industry subgroups and data items that were industry-related. Thus, for the 1989 income year, we post-stratified the sample by industry type and SNI level. For that sample, we used 8 industry groups and 18 SNI categories, for a total of 144 strata. Because of limitations at that time in the availability of universe counts for expansion weights, all corporations with an income loss of up to -$5,000,000 (which accounted for over 35% of all corporations) were placed into one stratum. As a result, we had 16 narrow positive income ranges and two wide negative income ranges. For the 1990 sample, we did not have that restriction, so we reduced the number of positive-income categories and constructed an equal number of negative-income groups.

For stratified random sampling, the formula for the variance of a total is:

$$V(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

Under poststratification, the variance formula becomes

$$V(\hat{y}_{post}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^{L} (1-W_h) S_h^2$$

where $W_h = \frac{N_h}{N}$

Poststratification involves forming strata after the sample has been selected. The probability of selection for each stratum is not determined beforehand so that the ultimate composition of the stratum is not known until after the data have been collected. In effect, the strata sample sizes are themselves random variables. Because of this additional component to the poststratification formula (2nd part of the formula), the standard error is slightly higher than one would get under proportional stratification (1st part), even though the sample is distributed almost proportionately to the sizes of the strata populations.

We needed to go further, however, than poststratifying. In order to improve the precision of sample estimates even more, we defined the strata in such a way that the bulk of the sample's diversity fell between strata, not within them. We did this in three ways: (1) by using stratifying variables that are most likely to be related to the estimates we consider most important, (2) by selecting appropriate strata cut-off points, and (3) by allocating more of the sample to those strata that are likely to have a large, diverse population.

For the new strata, we used the Dalenius-Hodges method to determine better income boundaries, or cut-off points. This method minimizes the sum of the products of the relative stratum size and the within-stratum standard deviation. The process consists of constructing a scale of the cumulative square root of the frequency distribution of the stratifying variable, y. One then finds the y-values that correspond to the endpoints of equal-sized intervals on the scale; these become the strata boundaries. It should be noted that different variables have different ideal boundaries. The boundaries we selected are recommended for variables highly correlated with SNI or taxable income.

As a result of this method we derived 19 income groups, which when combined with the eight industry types give us 152 strata. Because of the unique characteristics of banks and savings & loans, we have recently decided to separate them from the remaining financial corporations, creating ninth and tenth industry strata, leading to a total of 190 strata.

For the various scenarios presented here, we allocated sample to the strata in two ways. The first, optimum stratification, is proportional to the stratum universe size and variance, and inversely proportional to the square root of the data collection cost (essentially the time spent coding the tax return). The second allocation, proportional allocation, is proportional to the size of the stratum only, and ignores within-stratum variance.

For this analysis, we selected four data items: tax liability before AMT (Alternative Minimum Tax), gross receipts, cost of goods sold, and inventory, the last three of which have some correlation with industry type.

In order to assure comparability with the sample as it now exists, we assumed in all scenarios 100% sampling in the SNI>= $5,000,000 category and we assumed a constant overall sample size of 10,500, the size of the 1990 sample.

RESULTS

In this portion of the paper, we use the coefficient of variation as a measure of sampling error. The coefficient of variation is the standard error of the estimate divided by the estimate itself, allowing us to compare sampling error across many different variables.

Moving from 3-strata stratification to 152-strata post-stratification improved the accuracy level slightly. The already low standard error for tax liability was halved, but we noted little change in sampling error for the other variables, even though they had a slight correlation to industry type. As we shall see later, the main impact on the error level came from the division of SNI into smaller segments and not on the inclusion of industry type as a stratifying variable.

Going from 3 strata to 152 generally reduces the within-stratum sums of squares but it also reduces the denominator in the variance formula. If the stratum is small enough but still diverse, the sum of squares may remain high while the number of observations becomes very low, leading to an increase in the within-cell variance. For a variable which occurs infrequently, 152 strata may be too many if they are not defined optimally for that variable.

Furthermore, strata defined after the fact are usually not allocated optimally. Instead, they tend to approximate proportional allocation. We see later that proportional allocation, though operationally simpler, leads to substantially higher sampling error than does optimum allocation, and post-stratification leads to even higher error.

A third problem, which existed with our early method of post-stratifying but which is not reflected in these data was the omission of any segmentation of corporations with losses. As a result, nearly 40% of California corporations fell into one stratum.

Upon application of the newly defined strata, the sampling error diminished, especially when the sample was allocated optimally. For instance, the coefficient of variation for tax liability fell from 0.84% under post-stratification to 0.76% under proportional stratification to only 0.1% under optimum stratification. Between post-stratification and proportional Dalenius-Hodges stratification, the standard error decreased 7% for gross receipts, cost of goods sold, and inventory. However, under optimum allocation of sample, standard errors fell another 65% to 85% to very low levels.

Now, we have compared optimum allocation to proportional allocation to post-stratification and observed that optimum allocation greatly reduces sampling error. What about the effect of including industry type as a stratifier?

If we sample proportionally to stratum size, adding industry type as a stratifier to SNI yields slightly lower standard errors. The standard error for tax liability fell 12.0%; the other three variables exhibited declines of between 8% and 10%.

If we sample optimally, however, industry type appears to have a more substantial effect, not on tax liability, which is unrelated to industry type and showed only a 6% reduction in standard error, but on the other data items: a 55% error reduction for gross receipts, 48% for cost of goods sold, and 43% for inventories.

Industry type by itself used as a stratifier yields poor standard errors. In many cases, even with the variables which are slightly related to industry, the error is greater than the estimate.

Some of these results are corroborated by SOI studies. (SOI is the Statistics of Income Division of the IRS). One study (Leszcz, Oh, Scheuren) notes that the addition of industry type as a stratifier reduced standard errors by as much as 17%, though it can increase the standard error for some variables. Another study (Clickner, Galfond, and Thibodeau) demonstrates that industry classification, especially when used in conjunction with Dalenius-Hodges strata cut-offs for total assets (a traditional SOI stratifier), can reduce coefficients of variation. In this latter study, SOI evaluated estimates of three variables: total assets, inventories, and tax liability. The more complex stratification schemes (industry as stratifier, with or without Dalenius-Hodges assets cut-offs, varying number of assets categories) improved estimates of assets the most, followed by inventories, but actually increased the coefficient of variation for tax liability.

As stated earlier, sample is allocated optimally in inverse proportion to the square root of the data collection cost, so we set out to estimate the cost. From productivity statistics we found that it took an average of 7.07 minutes to code a large return (i.e., one with State Net Income over $5,000,000) and 4.36 minutes to code the smaller returns, but we had no other production statistics. As proxies for cost or time spent coding, we developed two measures: (1) the average number of schedules per return per stratum, and (2) the average number of California subsidiaries per return per stratum. Combining these two statistics, matching against the productivity figures we did have, and indexing the results, we obtained each stratum's relative cost. The cost per stratum as determined by this measurement increases as SNI moves toward the extremes; it also varies by industry type, peaking for industrial companies, and reaching its lowest values for financial service, retail and wholesale trade, and construction firms. Although these variations existed, accounting for them in allocating sample affected strata sample sizes very little.

SUMMARY

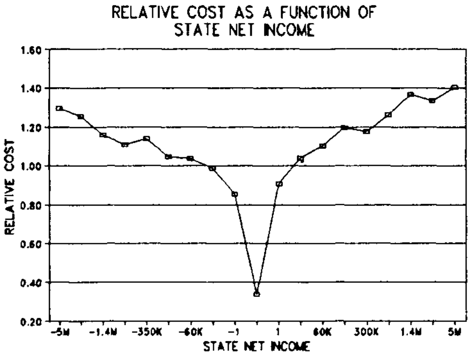
In conclusion, moving from 4 strata to 152 post-strata significantly decreased sampling error for some key items such as tax liability and SNI, but did not change the standard errors substantially for many items. Post-stratification is not the ideal way to stratify, especially with potentially small strata. First, sample is allocated proportionally, which leads to much higher standard errors than if the sample is allocated optimally. Second, our SNI strata definitions were not ideal. We used too many narrow categories, especially in the under $100,000 range. There is not enough heterogeneity between categories. The Dalenius-Hodges rule spreads out the categories, lumping the similar ones together.

Second, industry type itself is a poor stratifier and under proportional stratification added little to using SNI by itself.

However, if we allocate sample optimally, industry type becomes a more powerful stratifier. Under those conditions and by employing Dalenius-Hodges strata

boundaries, we push the coefficients of variation to below 5% for all the variables under consideration.

Since the optimum sample allocation and the strata boundaries depend on the variable we are trying to estimate, we should determine the degree to which these characteristics can vary for the B & C sample across many data items. The analysis of the four data items in this report indicates that the range may be reasonably constrained, so that we can find optimum values that work fairly well for most items.

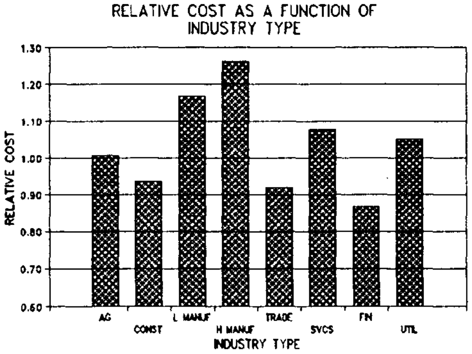


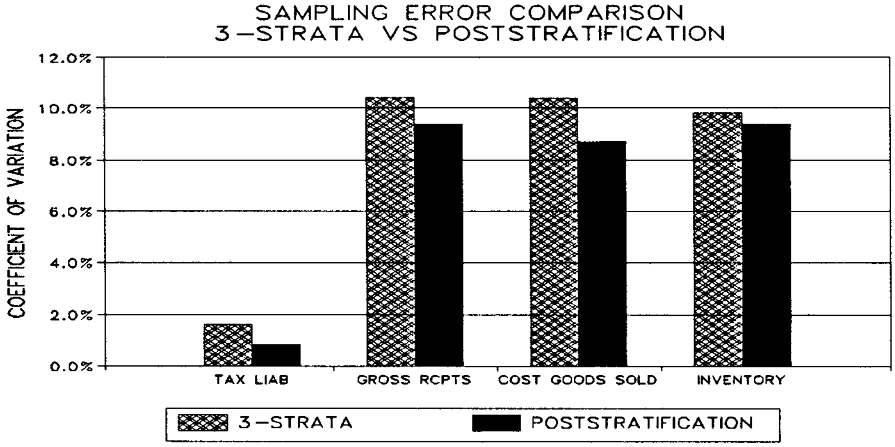
REFERENCES

Cochran, William, Sampling Techniques, John Wiley & Sons, 1977

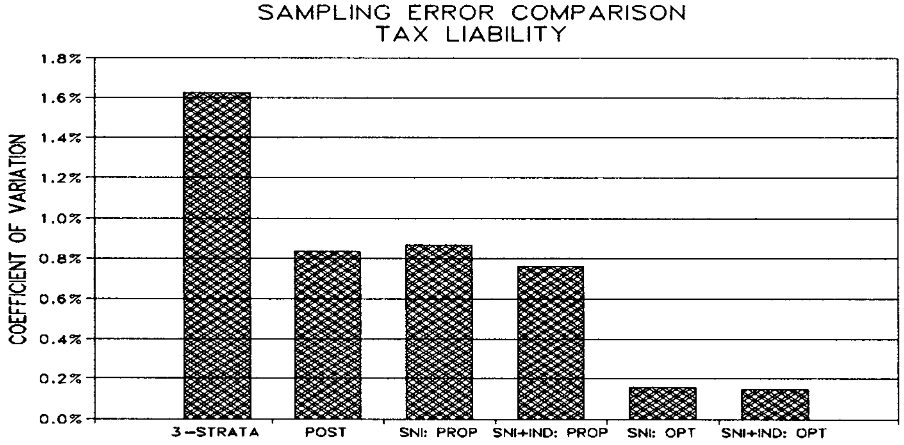
Leszcz, M. R., Oh, H. L., Scheuren, F. J., "Modified Raking Estimation in the Corporate SOI Program," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1983

Clickner, R. P., Galfond, G. J., Thibodeau, L. A., "Evaluation of the IRS Corporate SOI Sample," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1984

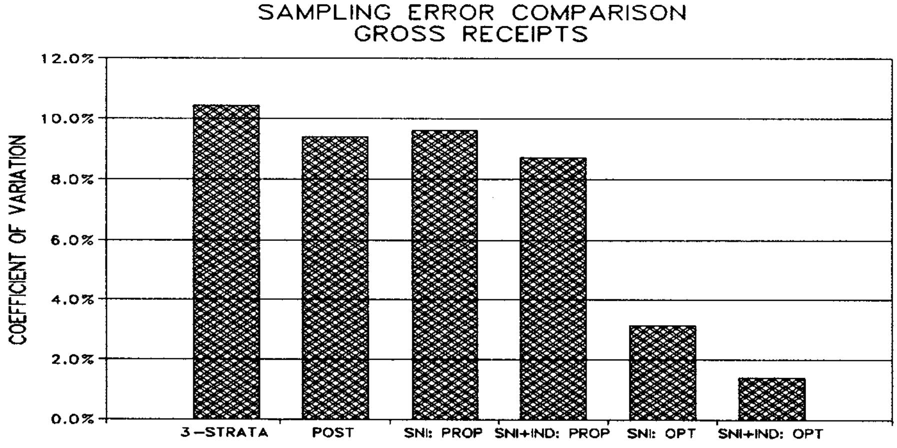




SAMPLING ERROR COMPARISON
3-STRATA VS POSTSTRATIFICATION



SAMPLING ERROR COMPARISON
TAX LIABILITY



SAMPLING ERROR COMPARISON
GROSS RECEIPTS

SAMPLING ERROR FOR STRATIFIED BANK AND CORPORATION SAMPLE UNDER VARIOUS STRATIFICATION SCENARIOS

STRATUM	VARIABLE	COEFF OF VARIATION
3 STRATA	TAX LIAB	1.6%
	GROSS RCPTS	10.4%
	COST GOODS SLD	10.4%
	INVENTORY	9.8%
POSTSTR 89 IND X SNI	TAX LIAB	0.9%
	GROSS RCPTS	9.2%
	COST GOODS SLD	8.4%
	INVENTORY	9.7%
POSTSTR 90 IND X SNI	TAX LIAB	0.8%
	GROSS RCPTS	9.4%
	COST GOODS SLD	8.7%
	INVENTORY	9.4%
NEW STRAT IND X SNI	TAX LIAB	
	OPT	0.1%
	PROP	0.8%
	GROSS RCPTS	
	OPT	1.4%
	PROP	8.7%
	COST GOODS SLD	
	OPT	1.5%
	PROP	8.1%
	INVENTORY	
	OPT	3.1%
	PROP	8.7%
NEW STRAT SNI	TAX LIAB	
	OPT	0.2%
	PROP	0.9%
	GROSS RCPTS	
	OPT	3.2%
	PROP	9.6%
	COST GOODS SLD	
	OPT	2.9%
	PROP	8.8%
	INVENTORY	
	OPT	5.4%
	PROP	9.5%
NEW STRAT IND	TAX LIAB	
	OPT	161.8%
	PROP	295.6%
	GROSS RCPTS	
	OPT	62.9%
	PROP	99.1%
	COST GOODS SLD	
	OPT	52.4%
	PROP	78.6%
	INVENTORY	
	OPT	171.5%
	PROP	209.6%