Is Your Agency Too Conservative?

Deriving More Reliable Confidence Intervals

Alan H Kvanli and Robert Schauer

Introduction

With the advent of the computer, many theoretical advances in statistical analysis have become practically feasible. These developments, when viewed in the backdrop of statistical study, are emerging at an amazing pace. However, in our opinion, thus far few government agencies have utilized these new approaches in evaluating random samples in such areas as projecting total tax underpayments or projecting monetary recovery amounts. This article is intended to shed some light on these new approaches that are more efficient and could potentially raise tax revenue, for instance.

Government agencies tend to be conservative like many other organizations and are oftentimes hesitant to take advantage of anything new. This conservatism can also be seen in how some government agencies evaluate random samples. These agencies take a "***lower limit***" approach in evaluating random samples (e.g., for projecting total tax underpayment or deficiency from a sample to the population).

We will offer the ***Bootstrap*** and the ***Empirical Likelihood*** (EL) as viable alternatives in computations of lower limits. In fact, we will assert that these methods are better than the traditional statistical approaches because they provide more reliable confidence intervals (e.g., 90% confidence intervals are indeed 90% confidence intervals) and the lower confidence limits are considerably larger. On the other hand, these methodologies are so new that few agencies have had a chance to take a serious look at them.

Some Background and a Call for Help

When examining tax error amounts, the error populations are unique in that they usually contain a large number of zero values due to the fact that most of the numeric-value sample items contain no error. Traditional methods of dealing with such samples are quite inadequate but government auditors have continued to use the traditional procedures over the years simply because there was nothing else available. In 1988, a report by the National Research Council exhorted statistical and auditing academicians to work together to come up with a procedure for deriving more reliable confidence intervals when dealing with audit populations[1]. On page 58 of this report, they state:

> It is important that intensive research be carried out for the purpose of developing more reliable procedures for determining lower confidence bounds. The financial benefits to the government from such research should be significant. ··· (The previous discussion) reveals that the statistics profession as a whole has not been heavily involved with the important statistical problems that arise in auditing. This may be due in part to the fact that there has not been adequate nor regular interaction between researchers from the accounting and the statistics professions.

In an earlier article in the *Journal of Government Financial Management* [6] we had suggested using the Bootstrap methodology. The Bootstrap offers a great improvement over the traditional procedure for audit populations containing a very large percentage of zero values. On the negative side, it is computer-intensive and requires some effort when an auditor wants to replicate the results using the same audit data. In 2003, an article in the *Canadian Journal of Statistics* by Chen, Chen, and Rao [3] suggested an improvement over a methodology proposed

in 1998 [3] for deriving more reliable confidence intervals when dealing with audit data. In our judgment, the bootstrap procedure offers a dramatic improvement over the traditional methodology. Nevertheless, the procedure explained by Chen, Chen, and Rao provides a solution that finally answers the plea put forth 16 years ago in the National Research Council publication.

Lower Limit Calculations

Some government agencies, rather than consuming extensive government and taxpayer resources required in a detailed audit, will use lower limits in projecting total tax underpayment (or deficiency) from a sample. Since projections are estimates, it is always possible to underestimate or overestimate the tax underpayment. The greater the chance of over-assessment, the less desirable the audit result is. A lower limit is conservative in that it offers protection against over-assessment. In a "lower limit" approach, the ***risk*** of over-assessment is calculated mathematically and considered in making an estimate of any unpaid taxes.

The first component of this risk is ***sampling risk***. This is the risk of overstating the underpayment amount because the auditor examined a sample of items rather than the entire population of items. Up until recently, government agencies have relied on "traditional" mathematical formulas to compute this sampling risk.

The other type of risk is ***non-sampling risk***, which is the risk that the total error amount was overstated generally due to unsound sample selection procedures. The only way to deal with non-sampling risk is to adhere to what we call "quality control" measures in the adopted sampling procedure. Although a very important topic in itself, we will not deal with non-sampling risk within this article.

For example, a government agency may perform a random sample of accounting records suggesting/showing a substantial underpayment of taxes, even though most of the tax had been paid correctly. Suppose the underpayment of taxes is estimated to be $1,400,000. Using an 80% ***confidence interval***, sampling risk, using the traditional methods, is computed to be ± 200,000. Here, we can say that there is an 80% chance that the unknown amount is somewhere between $1,200,000, the ***lower limit***, and $1,600,000, the ***upper limit***. Therefore, the conservative auditor will make a lower limit assessment of $1,200,000. Further, the auditor, can be 90% confident that the actual total tax underpayment will be at least the lower limit of $1,200,000.

The problem in these situations is that the "traditional approach" tends to overstate the risk. In our example, the actual risk will likely be much smaller than 10%. Further, if a better (more reliable) measurement of sampling risk were made, say $150,000, then the auditor could have increased the assessment while still maintaining a position with the same degree of conservatism.

As emphasized in the 1988 National Research Council report, government agencies have long been aware that the traditional approach generally overstates the sampling risk. On page 4 of this report they state that "There is a serious tendency for the use of standard statistical techniques that are based upon the approximate normality of the estimator of total monetary error to provide erroneous results. Specifically, as will be reviewed in the following chapter, both confidence limits tend to be too small." As noted before, agencies adhered to this method anyway because, one, it was prudent to make some allowance for sampling error, and two, there were no other reliable methods that could produce a reliable measurement of sampling risk.

Measuring Risk with the Traditional Approach

We stated that the traditional approach overestimates sampling risk. But how exactly does this happen? The answer cannot easily be explained in a non-technical article such as this, but it has to do with the nature of the populations sampled in tax audits.

Auditors will begin their audit focusing in on certain business transactions that have a possible tax impact, a ***population of interest***. These transactions usually have some value attached to them that we call ***examined amounts***. These are the monetary amounts of transactions made by the taxpayer. Examined amounts can be invoice amounts, book entries, detail amounts reported on a tax return or any other transaction that is applicable to the audit. Some people may refer to the examined amounts as ***book values***, ***recorded values*** or ***reported values***. These terms can be interchangeable, but they usually refer to specific types of transactions.

The population of interest, or the focus of the tax auditor is not really on the examined amounts, it is almost always the ***tax errors*** associated with the examined amounts – a point easily overlooked since the examined amounts are more visible and are available prior to carrying out an audit. It is true that there often is a correlation between the examined amounts and the tax errors. Note that the more compliant the taxpayer, the smaller the correlation that subsists between these two sets of numbers. In other words, the tax errors are a function of both the taxpayer behavior and the examined amounts.

If a population is described as ***normal***, the "bell curve" characterization of the population values will be symmetrical around the population mean. If they are asymmetrical, we say the values are ***skewed***. There is considerable empirical evidence (as detailed in the 1988 National Research

Council report) that suggests the nature of examined-amount populations, or accounting populations in general, are highly skewed, that is, non-normal. Such populations generally contain a few very large monetary values and mostly small monetary values.

The nature of the tax errors for any population of transactions is even more complex because of a dichotomy in clustering -- the tax errors consist of two very different types of values. Non-errors, represented as zeros, are very ***homogeneous*** since they are all identical. The proportion of zero values in the sample can be quite large because, as noted before, accounting populations, by their nature, tend to be mostly correct. The remaining values, the population units with tax error, or those represented by non-zeros, are much less frequent, are usually non-normal (skewed), ***heterogeneous*** (dissimilarly distributed patterns) in nature, and frequently resemble the distribution of values found in the population of examined amounts. So, unlike the population of examined amounts, the error population is a mixture of two populations; the zero values and nonzero values. A typical error population is shown in Figure 1, where the error values consist of (1) a "spike" at zero and (2) nonzero values which are highly skewed.

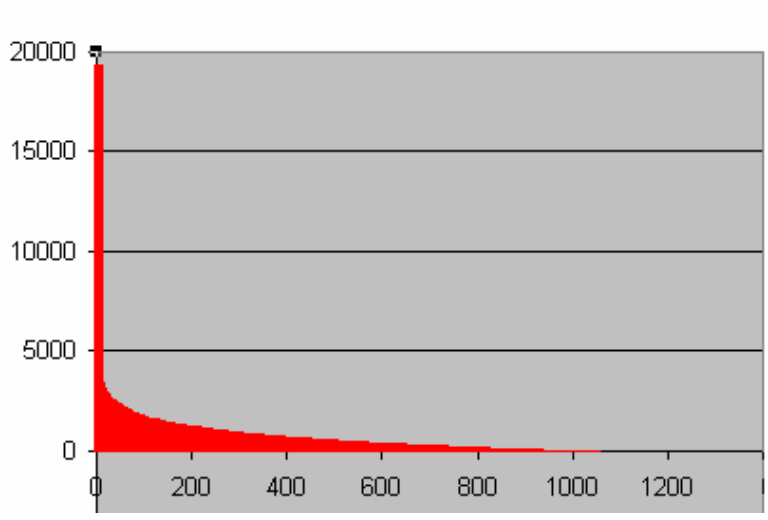

Figure 1

So why is all of this important? This is because the traditional approach makes certain assumptions. The biggest assumption is that the underlying population being evaluated has a normal distribution. As explained above, accounting populations are usually very far from normal. Traditional statistical inference can be applied to skewed populations, provided a sufficiently ***large sample*** is drawn. But then, how large is "large" and how do you know if you have a large sample? There is no single number that represents a large sample – this differs from population to population.[2] Oftentimes, what represents a large sample requirement may simply be too large to be realistic.

Further, a random sample taken from a population of tax errors will have all the problems found with a population of examined amounts and more. In Figure 1, the injection of numerous zero values (low error rate populations) will make the population even more skewed than the examined amounts and cause the traditional methods to produce even less reliable lower confidence limits.

The Alternatives to the Traditional Approach

Today, we have several valid statistical approaches available to evaluate a random sample.[3] Tax enforcement agencies have relied on several sources for the traditional approach such as Cochran [see reference 1], whose primary focus was survey sampling. Another oft-cited source is from an AICPA publication intended for financial auditors by Roberts [2]. Regardless of the auditor's source, the traditional methodology will tend to overestimate lower limits. Besides this problem, the traditional approach also makes certain statistical assumptions which at times can be very tenuous.[4]

Bootstrap Approach. Although the bootstrap procedure itself has been around for some time, its full development has not been realized until just recently. It has been detailed in several different publications [4 & 5] and a recent article [6] offered the Bootstrap as an alternative to the traditional approach. This approach is more reliable in most instances concerning tax estimates - the Bootstrap provides an improved estimate of the risk. Further, the statistical assumptions required by the traditional approach are not necessary for this approach. However, there are some practical limitations to this approach that must be taken into account.

Empirical Likelihood (EL) Approach. The EL approach, introduced by Owen [7] and offered as a viable alternative by Chen, Chen, and Rao [8], is the most attractive approach for three basic reasons: (1) it provides for reliable measurements of risk (i.e., a 90% lower limit is indeed a 90% lower limit), (2) it does not require that some problematic statistical assumptions be made (i.e., normality of the distribution, see endnote 4), and (3) it does not suffer from some of the practical problems when compared to the bootstrap procedure (see Arguments For and Against Changing: Bootstrap Approach). It seems the only argument against its use is the fact that it is new!

Some Actual Samples

As stated previously, we now have several additional methods upon which to evaluate samples. We want to demonstrate that these alternative methods are worthy of serious consideration. We used the same 12 data sets from the recent bootstrap article published in this Journal in 2002 [11]. These data sets are random samples from recent sales and use tax audits carried out by the Washington State Department of Revenue. The table contained the lower limits of 90% confidence intervals using the traditional and bootstrap procedures. These same data sets were

used with the EL confidence interval procedure and the results are summarized in the attached table.

Observations. In all cases, bootstrap and EL lower limits are considerably less conservative than the lower limit produced by the traditional procedure. This is especially true for samples containing less than 10 nonzero errors. For the two samples containing the fewest errors (Samples 1 and 2), the EL lower limits are much less conservative than the bootstrap lower limit. For samples 1 through 4, those with the smallest error rates, the bootstrap and EL lower limits are much less conservative than the traditional limits. Note that in sample 1, the extremely large margin of error resulted in a negative lower limit when the margin of error was subtracted from the point estimate using the traditional procedure.

Both the bootstrap procedure and the EL procedure produce a minimal gain in the lower limit over the traditional procedure when the *sampling fraction* (sample size divided by the population size) is over 10%. Finally, the value of the lower limit when using the bootstrap or the EL procedure is only slightly larger than the traditional lower limit whenever the sample contains 30 or more nonzero errors.

Sample (Smp)	Smp Size – Smp Fraction Population Size	Nonzero Errors Error Rate	Traditional Lower Limit (LL) in $	Bootstrap LL in $ *Increase over Traditional in %*	EL LL in $ *Increase over Traditional in %*
1	**365** - *0.1%* 509,985	3 .8%	-1,273,609	104,791 *108.2**	372,000 *129.2**
2	**300** - *1.5%* 19,712	6 2%	180,935	277,856 *53.6*	305,485 *68.8*
3	**365** - *0.2%* 195,767	8 2%	1,611,222	6,178,412 *283.5*	6,516,597 *304.5*
4	**365** - *0.1%* 301,410	10 3%	1,085,248	7,007,027 *547.7*	6,466,385 *495.8*
5	**247** - *20.0%* 1,234	10 4%	16,923	19,976 *18.0*	20,391 *20.5*
6	**300** - *0.3%* 91,706	11 4%	725,080	875,401 *20.7*	860,708 *18.7*
7	**120** - *6.3%* 1,891	12 10%	11,687	13,359 *14.3*	13,306 *13.9*
8	**250** - *1.8%* 13,796	13 5%	59,614	75,371 *26.4*	75,817 *27.2*
9	**300** - *7.4%* 4,048	13 4%	167,383	182,820 *8.9*	185,734 *11.0*
10	**300** - *1.5%* 19,536	15 5%	187,831	208,231 *10.9*	207,609 *10.5*
11	**300** - *0.1%* 287,027	18 6%	664,749	774,963 *16.6*	786,021 *18.2*
12	**300** - *0.5%* 63,976	30 10%	411,956	425,947 *3.4*	434,872 *5.6*

Table: Sample Results

(*) These two values are computed using the absolute value of:
X / Traditional Lower Limit * 100, where X = New Lower Limit – Traditional Lower Limit

Arguments For and Against Changing

Bootstrap Approach

PROS: 1) The Bootstrap confidence intervals are more reliable (based on the simulation results in [6]) and have larger lower limits than those derived using the traditional procedure. They clearly represent an improvement over the traditional intervals.

2) Although the bootstrap procedure is very computer intensive, there are existing software packages available to derive Bootstrap confidence intervals. The authors also constructed an Excel template (available upon request) that will derive such an interval.

CONS: 1) The Bootstrap procedure is difficult to replicate. Since this method involves randomly drawing samples from the original sample, the random number generator seed values must be captured in order to replicate the results. The template developed by the authors does capture these seed values and consequently allows the auditor to duplicate the results (obtain exactly the same confidence limits).

2) The formulas for a stratified sampling design are extremely complex. In larger tax audits, stratified sampling is used more often than simple random sampling.

3) Each time the Bootstrap procedure is run (assuming different seed values in the random process), a slightly different interval will result. The lack of a "fixed" lower limit calculation could present problems in tax audit applications.

Empirical Likelihood (EL) Approach

PROS: 1) No assumptions regarding the shape of the error population are made and so there are no assumptions to challenge during an audit appeal.

2) As with the Bootstrap technique, the EL methodology produces equally good or even more reliable confidence intervals (based on the simulation results in [8]) with less conservative lower limits.

3) Can be easily adapted to a stratified sampling design.

4) Unlike the bootstrap procedure, a set of error values will always provide a "fixed" lower limit calculation. Hence, there is more consistency in the results.

CONS: 1) There is not a closed-form expression for this confidence interval and to determine a lower limit, one must search a particular interval for the value satisfying a certain equation.[5] However, this procedure is very easy to program and within the computing abilities of a standard personal computer.

2) This is a very new methodology and most statisticians and auditors are unaware of it.

Conclusion

Accountants, tax audit agencies, and other related individuals or groups tend to be conservative. This is generally for good reason. Nevertheless, excessive conservatism can be problematic. As we have demonstrated, being overly conservative can mean that tax revenues are needlessly reduced. The newer methods offer more realistic lower limits and higher tax revenues. Any agency relying on lower limits should take a serious look at the Bootstrap procedure [6], and even more so, the Empirical Likelihood (EL) procedure introduced in this article.

Tax audit populations, that is low error rate populations, are highly skewed underlining the need for improved evaluation techniques that until recently were just not available. The more skewed the population, the more the traditional approach suffers. On the other hand, the Bootstrap and

the EL procedures are not saddled with these same often-overlooked assumptions. This is a distinct advantage in the event that the agency has a contested audit result.

Admittedly, the traditional approach has an advantage when compared to the newer alternatives in that they are more straightforward and easier to compute. However, the advent of sophisticated computer processing virtually wipes away this advantage. All of these new methods can be (and have been) programmed and tested for accuracy and should not present problems to the government auditor.

Indeed, the development of the computer has allowed statisticians to come up with new evaluation procedures. In our view, there remains no doubt about their improvement over the traditional methods currently in wide-spread use.[6] All that remains to be done is for agencies to implement the new approaches. We feel this can be done and this article provides the justification for doing so.

References Cited in Text

[1] William G. Cochran, *Sampling Techniques*, John Wiley & Sons, New York, 1977, chapters 2, 5, 5A, 6 and 7.

[2] Donald M. Roberts, *Statistical Auditing*, American Institute of Certified Public Accountants, New York, 1978, chapters 5 and 6 and appendix 2.

[3] Alan H. Kvanli, Y.K. Shen, and L.Y. Deng, "A Construction of Confidence Intervals for the Mean of a Population Containing a Large Number of Zero Values," *Journal of Business & Economics*, July, 1998, pp. 362 - 368.

[4] Bradley Efron and Robert J Tibshiani, *An Introduction to the Bootstrap*, 1998.

[5] A.C. Davison and D.V. Hinkley, *Bootstrap Methods and their Application*, 1998.

[6] Alan H. Kvanli, and Robert Schauer, "The Bootstrap: What the Government Auditor Should Know," *Journal of Government Financial Management*, Fall, 2002, pp. 24 - 33.

[7] Art B. Owen, *Empirical Likelihood*, 2001.

[8] J. Chen, S. Y. Chen, and J. N. K. Rao, "Empirical Likelihood Confidence Intervals for the Mean of a Population Containing Many Zero Values", *The Canadian Journal of Statistics,* March, 2003, Vol. 31, No. 1, pp. 53 - 67. This article also discusses the application of the empirical likelihood (EL) methodology to stratified sample designs.

[9] See pages 39-44 in Cochran's *Sampling Techniques*.

[10] See page 83-84 and Appendix 4 in Roberts' *Statistical Auditing*.

[11] See footnote 8 in "The Bootstrap: What the Government Auditor Should Know" by Kvanli and Schauer.

ENDNOTES

1.. Panel on Nonstandard Mixture of Populations, *Statistical Models and Analysis in Auditing: A Study of Statistical Models and Methods for Analyzing Nonstandard Mixtures of Distributions in Auditing*, Washington, D.C. National Academy press, 1988.

2. Cochran at Reference [9] provides a rough means for determining what constitutes a large sample. This is much greater than the often-cited sample size of thirty, which applies to populations that are fairly homogeneous. Using Cochran's guide, tax populations often require sample sizes in the hundreds or thousands to allow for a reliable confidence interval to be computed.

3. In addition to the Bootstrap and the EL, there is another possible approach, the Likelihood Ratio (LR) Approach, first introduced in 1998 by Kvanli, Shen and Deng [3]. The LR approach is not only more mathematically sophisticated but it provides for a more reliable estimate of risk when compared to the traditional method. Although

the approach is valid, it depends on a key assumption regarding the shape of the nonzero error population. Consequently its authors did not actively encourage government agencies to adopt this approach.

4. Here are two troublesome assumptions that are made when using the traditional approach:

- Several of the projection methods often use the ***t distribution*** to compute the confidence interval. Anytime a t distribution is used to derive a confidence interval for a mean (or total), a key assumption is that you are sampling from a normal population. As illustrated in Figure 1, error populations generally contain a mixture of many zero values and highly skewed nonzero values; that is, populations which are extremely non-normal. Up until now, this method was used because the procedure did give an approximate result, and it was the only method available.
- Using the normal distribution to derive a confidence interval assumes that the sample mean follows an approximate normal distribution (based on the Central Limit Theorem). However, there is much evidence to suggest that when dealing with audit populations containing a large number of zero values, extremely large samples are required for the Central Limit Theorem to hold.

5. Define $f(\mu) = 2\sum_{i=1}^{n} \log[1 + A(Y_i - \mu)]$ where, for each value of μ, A satisfies the equation $\sum_{i=1}^{n} \frac{Y_i - \mu}{1 + A(Y_i - \mu)} = 0$. The lower limit of the two-sided 90% confidence interval for the difference population total is $N\mu_L$ where N is the size of the difference population and μ_L is the smallest value of μ for which f(μ) is equal to 2.7055. Similarly, the upper limit is $N\mu_U$ where μ_U is the largest value of μ for which f(μ) is equal to 2.7055. The value 2.7055 is the χ^2 (chi-square) value having 1degree of freedom and a right-tail area of .10. This right-tail area would be .20 for an 80% confidence interval and .05 for a 95% confidence interval.

6. Recently, the Federation of Tax Administrators has published a report on sampling in sales and use tax audits, *Sampling for Sales and Use Tax Compliance*, December 2002, which is available on their web site (www.taxadmin.org/fta/ftapub.html). Within the report is a summary of sampling procedures used by the states (reference Appendix A).