CS-BIGS 5(2): 111-114

http://www.bentley.edu/csbigs

Random sampling issues in a federal court case, a case study

Kristin Kennedy
Bryant University, USA

James Bishop
Bryant University, USA

The authors of this paper, statistics professors at Bryant University, recently had the opportunity to act as expert witnesses in a case involving the Internal Revenue Service (IRS) as the plaintiff. The IRS used some sampling techniques in selecting a random sample that did not appropriately represent the population at hand. This action led the IRS to draw inferences about the population that were not likely conclusions on their part. The purpose of this paper is to highlight the fact that fundamental mistakes are made in the business and legal world regarding sampling. This legal case is a good case study to present to any statistics class, revealing both the pitfalls of inappropriate statistical sampling techniques, and incorrect inferences made based on an inappropriate sample. Basic random sampling techniques are developed in elementary and intermediate statistics classes, and the following paper highlights an example on random sampling that can be used in class, and is understandable by all students.

Introduction

Every introductory statistics class covers the topic of random sampling. For those who work with statistics, it is well known that it is not always easy to create or secure a good random sample for study. In fact, the topic of random sampling is often mentioned and reviewed many times throughout a semester in any statistics course, if for no other purpose than to stress to the students that finding a good random sample can be tricky. Certainly, students are made aware of the fact that there are many variations of valid sampling approaches such as stratified, systematic, or cluster sampling. Regardless of the variation used, the emphasis is to obtain the best random sample possible for the population or sub-population of interest. Without valid random sampling, statistical inferences will be limited or void of meaning at the end of the project.

Simple random sampling is a method in which each set of n items selected from a population has an equally likely probability of being selected as sample. This sampling technique is popular and is often the technique that non-statisticians would think of using in order to create a sample. Stratified sampling is used if the population can be divided into mutually exclusive subgroups, called strata. The strata should have population items that are as much alike as possible within them. Random sampling can then be applied to each stratum. Combining the data items that are selected from each stratum, the statistician has the hope that the sample reflects the whole population better for a given sample size. Other sampling approaches such as cluster sampling or systematic sampling are discussed in most statistical textbooks.

The authors of this paper, statistics professors, recently had the opportunity to act as expert witnesses in a case involving the Internal Revenue Service (IRS). The IRS was the plaintiff and the defendant was the owner of a tax preparation service. Some of the other co-workers were named as defendants too, but the owner was the person being prosecuted by the IRS. The main point of this paper is to highlight that the IRS in their statistical analysis improperly used their sample to make inferences about the population as a whole. That error damaged their case. The IRS used a selective method of sampling that resulted in a sample which was not representative of the whole population. This led to the fact that their inferential arguments regarding the data were unsubstantiated.

The purpose of this paper is to highlight the fact that fundamental mistakes are made in the business and legal world regarding sampling. This legal case is a good case study to present to any statistics class, revealing both the pitfalls of poor statistical sampling techniques, and, invalid inferences made based on an inappropriate sample. These mistakes can be avoided with some basic skills that are easily developed in an elementary statistics class.

Bankruptcy prediction has gained increasing attention since the 1960s (Altman, 1968), and not without reason. Predicting the financial distress of firms benefits the company leaders by identifying internal problems, but also assists auditors in their work for finding potentially troubled firms. Above all, bankruptcy prediction produces information for investors and banks so that they can make sounder lending and investing decisions (Wilson & Sharda, 1994; Atiya, 2001). At present, the applied methods range from well-known statistical methods to advanced soft computing techniques (Kumar & Ravi, 2007). Nevertheless, predicting the probability that a firm will fail is not sufficient, because it does not reveal the causes behind the event. This paper proposes to use a technique called ensemble of locally linear models combined with forward variable selection. It is able to assess the importance of the variables, thus providing more interpretability than "black box" models.

The stated case of the IRS

The entire details of the lawsuit brought by the IRS against the defendant will not be covered in this paper. However, parts of this case are statistically interesting. The defendant was the owner of a tax preparation firm with several locations, and he was directly or indirectly responsible for the preparation and filing of at least 24,399 federal income tax returns for the tax years 2003 through 2007. The IRS stated that they reviewed 345 returns of the 24,399 identified. Of the 345 which the IRS reviewed, 313 resulted in needing additional tax assessment. This means that 91% of the original sample had returns that owed additional tax to the IRS, and the additional tax was owed for a variety of reasons. The IRS calculated from these 345 returns that the actual tax loss directly due to these returns being improperly prepared by the defendant(s) was in excess of $1.1 million (United States v. Brier, et. al., pg. 3). The IRS further stated that if this rate loss were applied to all 24,399 returns, then the estimated loss to the United States government would be in excess of $85 million for the years 2003 through 2007 (United States v. Brier, et. al., pg. 5). Thus the IRS was looking for damages close to 85 million dollars.

The sampling selection error and the statistical error

Two serious errors were made by the IRS analysts when presenting their findings. One of these errors involved the method by which the 345 returns were selected from the overall population. The second error was the statistical inference made from the evaluation of this sample.

The first serious error was a fundamental sampling selection error. As we teach in any elementary statistics class, good statistical sampling is performed randomly from a population or sub-population. 345 returns could have been randomly selected from the entire population, and the analysis could then be performed by standard statistical methods.

As stated in the Plaintiff Motion for Preliminary Injunction (pg. 104 -105), the IRS selected their sample by choosing only returns that had a Schedule C attached. The IRS used selection criteria called a differential score that would only select certain returns (pg. 105). This in itself is not a problem. Representative statistical sampling can be done by stratifying the population, sampling randomly from sub-populations proportionally. However, one must know something about the overall numbers in each sub-population from which a sample is taken.

For example, suppose the population of 24,399 had 61.5% returns with a Schedule C attached. Then 15,005 of the returns have a Schedule C, and 38.5% of the returns do not, or 9,394 do not. When the stratified sample is taken, 61.5% of 345, or 212, should be selected with Schedule C and 38.5% of 345, or 133, should be selected without Schedule C. The court records indicate that all 345 returns were sampled from the Schedule C group. If

Schedule C returns are different from non Schedule C returns, then the sample that the IRS collected is not representative of the population. Therefore any estimates calculated based on this sample would risk being biased.

Since the 345 were selected with a particular criterion that was targeted to find discrepancies, any inferences made based on this sample could only be generalized to that particular sub-population. Basically, this sample was taken by the IRS with the intent of using a search criterion, which would find the maximum number of discrepancies. If a correct sample was taken, then almost certainly the total amount of projected discrepancies would be less than $85 million.

The other serious error that the IRS made is a simple calculation error that led to an improper conclusion. Unfortunately no one from the IRS checked the math before they entered federal court with their motion. 345 returns were selected and reviewed out of 24,399 in total. The discrepancies from this sample totaled to $ 1.1 million. The IRS then maintained that if the same error rate were followed for the rest of the population, then the discrepancies would calculate out to be about $85 million.

Most likely, the IRS computed this number using a simple ratio of 1.1/313 = x/24,399. This ratio calculates to x = 85.7 million, which similar to the figure they stated. Since 313 is the number of examined returns with discrepancies from the sample, the calculation assumes that the entire population has discrepancies or roughly the same amounts as in those sampled.

The IRS should have used the ratio of 1.1/345 = x/24,399, and this ratio calculates to x = 77.8 million. This number assumes that the amount of discrepancy for every 345 returns, the actual sample, continues for the entire population. There is almost an $8 million dollar difference in the two calculations.

Sampling options

To infer numbers about the entire population, sampling should be performed on any sub-group(s) of the population that may have heterogeneity of effects. In this case, the year in which the return was submitted, the office that prepared the return, the name of the preparer, and the income level of the client are all potential variables that could be proportionally sampled due to potential heterogeneity.

The only number that is of great concern to the case is the estimated $85 million dollars worth of discrepancies in returns as stated by the IRS. Let us compare three approaches to this problem demonstrating the value of good statistical sampling design. The discrepancies on a given return could fall in one of three categories: (1) the IRS owes the individual money, but those were found to be very small, in fact negligible, (2) there was no discrepancy in the return, or (3) a discrepancy was found and the individual owed the IRS more tax. For this discussion, we consider the first two categories to be as zero discrepancy. Thus there was either a discrepancy, in which more tax was owed or no discrepancy.

First consider our population of values with an unknown distribution with many of the discrepancies being zero (thus the distribution is skewed to the right). Clearly the data do not have the properties of the normal distribution. However, taking a large simple random sample, one would hope to invoke the central limit theorem in that the sampling distribution of the mean is normally distributed for a large enough sample. This is essentially what the IRS was doing except that they apparently only did this for Schedule C returns, and so their results would only apply to that population.

Suppose a simple random sample of n=345 is taken from the overall population, and 1.1 million dollars worth of discrepancies are found. The mean discrepancy for the sample is $3,188 (1.1 million divided by 345). If we compute the sample standard deviation, then a confidence interval can be obtained for the overall mean of the population discrepancy. Finally, an interval estimate for the sum of discrepancies for the entire population can then be computed.

The IRS did not make the actual data available for analysis and therefore we can only investigate potential results to this problem depending on various values of the sample variation. In addition, we can consider a variety of sample sizes.

We assume the Xs to be independent here. There are likely to be some correlated Xs (a friend recommends the tax firm to another friend), but the correlations among the 24,399 returns are assumed to be negligible.

If we then create a two standard deviation range for the n = 100 sample case using the smallest Var(X) from above, we obtain a discrepancy range of [308,000 - 328,000]. Creating the proper ratio for the entire population (24399/100), we obtain the range [$75.3 million – $80.2 million]. Remember that this analysis depends on the 100 samples are representatively taken from the overall population.

Two problems exist relative to this approach. One is that one or more sub-populations has significantly different discrepancy data than other sub-populations, thus making inferences from the sample inaccurate. The other problem is that even a somewhat large sample size is not enough to make the sample mean normally distributed due to the skewed nature of the population data, since we expect a large number of discrepancies to be zero.

A second approach to this problem would be to stratify the population. Suppose there are 4 different office locations, 4 different income levels of the clients, and 4 different years of returns. The number of sub-populations for stratifying is 4*4*4 = 64. Taking 30 or more returns from each sub-population (assuming there actually are that many returns in each sub-population) would give us a sense of potential differences between the sub-populations.

Stratifying the sampling as mentioned above will give us assurance that the entire population is appropriately represented. Unfortunately, this greatly increases our overall sample size. Rather than the 345 targeted samples chosen by the IRS, a proper sample would now consist of a couple of thousand samples. Due to the number of discrepancies with the value zero, an even larger sample may be needed from a number of the sub-populations in order to get accurate estimates.

As a third approach, we can separate out the returns with no or negative discrepancies and then consider only the positive discrepancies. Those would be more symmetrically distributed or closer to normally distributed. This method is often used in the insurance industry for cases when the values of claims are zero for all those people who never make a claim on their policy. The advantage in this case is that for reasonably large n (but smaller than would otherwise be required), we can consider the sampling distribution of the sum to be normally distributed.

This more advanced concept would require a variable N to represent the number of returns with positive discrepancies which would be distributed as a binomial B(24,399,p). The variability of the sample is then considered by combining the binomial variation with the sample variation of the positive discrepancies with the advantage of more assurance of a normal distribution of values from the positive discrepancies.

Combining this concept with the stratified sampling from 64 sub-populations, we could take 30 returns from each sub-population (total 1920 returns). Combining the standard deviations of the samples allows us to consider the final column of Table 1 (for 2000 returns) and thus compute an approximate confidence interval for the sum of all discrepancies.

Table 1. Sample standard deviation of the mean based on various values of the sample variance and n.

Var(X)	n:100	n:345	n:500	n:1000	n:2000
250000	5,000	9,287	11,180	15,811	22,361
1000000	10,000	18,574	22,361	31,623	44,721
4000000	20,000	37,148	44,721	63,246	89,443
16000000	40,000	74,297	89,443	126,491	178,885

Stratifying the sampling as mentioned above will increase the accuracy of the resulting analysis. Roughly speaking, a good statistical design would suggest close to 2,000 sample returns in order to approximate the sum of population discrepancies with any confidence.

Conclusion

The IRS made a fundamental mistake in sampling, and therefore arrived at inaccurate conclusions. In fact in court the IRS conceded the point immediately to the defense and stated that the 85 million dollar figure was in no way based on any valid statistical methodology. This error damaged the credibility of the IRS and the case against the defendant. With a small amount of careful planning, the IRS could have avoided this pitfall, and in fact built a strong case against the defendant.

This is a straightforward case that can easily be used in an elementary or even advanced statistics class to highlight the importance of well designed sampling techniques. This case also shows that anyone with some basic knowledge of statistics can catch errors that are being stated as fact from poor mathematical analysis and poor sampling techniques.

REFERENCES

United States of America v. Michael Brier, et. al. Complaint for Preliminary and Permanent Injunction, Case 1: 09-cv-00607-ML-DLM, Filed March 10, 2010.

United States of America v. Michael Brier, et. al. Plaintiff's Motion for Preliminary Injunction, CA NO. 09-607, Filed May 12, 2010.

Correspondence: kkennedy@bryant.edu