BNA, INC.

DAILY LABOR REPORT

BNA Employee-Owned Since 1947

CLASS ACTIONS

When used correctly, sophisticated sampling techniques can provide accurate information at a reasonable cost for use in class action employment litigation, Navigant Economics director and principal William J. Carrington writes in this BNA Insights article. He discusses sample designs that make use of stratification, clustering, and different probabilities of selection. The purpose of this article, he writes, is to familiarize attorneys with the mechanics of standard sampling methods in order to reduce both the misuse of sampling techniques in court and to reduce some attorneys' skepticism of the methods discussed.

Sampling Techniques Used in Class Action Employment Litigation

BY WILLIAM J. CARRINGTON

There are many instances in the course of class action employment litigation where gathering all the relevant information is very expensive. In wage and hour matters, an accurate assessment of each employee's off-the-clock time might require the examination of daily timekeeping records kept only in paper format. In discrimination cases, the calculation of economic losses associated with a layoff often requires detailed analyses of each laid-off employee's personal circumstances, even when a detailed analysis for each laid-off employee is very costly.

William J. Carrington, Ph.D., is a director and principal in the Washington, D.C., office of Navigant Economics, which provides economic and financial analysis of legal and business issues to law firms, corporations, and government agencies.

In these and other circumstances, there is understandable interest from all parties in using sampling techniques to reduce costs while still producing a reasonably accurate answer to the questions driving the case and, indeed, sampling techniques are widely used. Though common, sampling techniques are often misused in litigation and, even when used correctly, are often viewed with skepticism by courts and by opposing counsel. The purpose of this article is to familiarize attorneys with the mechanics of standard sampling methods in the hope that a bit more knowledge will reduce the incidence of both problems.

When most nonstatisticians think about sampling, they have in mind what statisticians refer to as an urn model, whereby all elements of the population are thrown in an urn and then a subset is picked out at random to comprise the sample. This method is known as simple random sampling (SRS), but most sampling methods used in litigation and elsewhere depart from SRS in important ways. The divergence from SRS sampling is driven by a recognition that collecting information is costly and that the goal of sampling should be to get the most accurate answers for any given cost. In most cases, SRS sampling does not achieve this goal and, as a result, statisticians use a more sophisticated set of tools to improve the accuracy of their samples. The remainder of this article describes these tools and discusses their application in the litigation environment.

Stratification.

The first dimension on which sampling often diverges from SRS is in the use of *stratification*. The idea behind stratified random sampling is simple. If we divide the population into strata such that every element (e.g., a person) in the population belongs to one and only one stratum, then stratification entails taking a separate random sample from each stratum where the number of elements sampled from each stratum is fixed in advance. For example, if we were sampling employees from a firm with half its workforce in California, we might stratify our sample to guarantee that half of our sample comes from California as well. The virtue of stratification is that, absent this mechanism, there will be sampling variability in each stratum's share of the sample, e.g., for purely random reasons, we might oversample (i.e. disproportionately sample) or undersample California. If the strata are very different from one another, then stratification provides a benefit by precluding the inadvertent sampling of strata at different rates.

There may be instances where we are unable to use stratification in advance of drawing the sample because we do not know which stratum individuals belong to until they have been sampled. For example, we may know in advance that 30 percent of an employer's workforce took a particular training course, but we may not know which particular employees took the course until we dig into their paper records. In some cases, it may make sense to take an SRS and to then, after the fact, reweight each element in the sample so that the 30 percent of the weighted sample has taken the course. Such *post*-stratification solves part of the problem that standard or pre-stratification seeks to address—it ensures that the different strata are given the appropriate weight in any sample-based calculations. Unlike standard stratification, post-stratification does not guarantee each stratum's representation in the sample and, as a result, overall sample calculations are less accurate than had pre-stratification been possible.

Clustering.

Statisticians also rely on the related concept of clustering. Clustering is similar to stratification in that it entails dividing the population into subgroups or "clusters" but, unlike stratification, some clusters are sampled while others are not. The motivation for clustering is related to the cost of collecting a sample. For example, if conducting a sample entailed visiting a manufacturing plant to examine onsite records or to interview employees, then trying to visit every site for a large firm might be extremely costly. Clustered sampling entails a first round of sampling where some clusters are sampled while others are not and then a second round of sampling *within* each sampled cluster, often using SRS.

The virtue of clustered sampling is that it is possible to get a larger sample per dollar spent than under SRS. Intuitively, if an information collector has gone to the trouble of getting to a remote site, then it makes sense to collect all the data available at that site. If an employer has many dispersed sites, then it will often make sense to sample the sites in a first round and then to sample employees within each of the sampled sites in a second round. A key feature of clustered sampling is that while it allows for a larger sample for any given cost, that sample is oftentimes less informative than a (more expensive) SRS sample of the same size. The reason for this is that individual elements within each cluster often bear more resemblance to one another than would randomly chosen elements not in the same clusters. For example, if we were doing clustered sampling of departments within a firm, it is likely that average earnings within the legal department are higher than those in the facilities department. In this context, clustered sampling would run the risk of sampling everyone from the legal department and nobody from the facilities department, or vice versa. The obvious consequence would be a mischaracterization of firmwide average earnings. For these reasons, clustering works best when the clusters are not too different from one another.

Sample Weights.

The last dimension in which statisticians and survey designers commonly depart from SRS is in sampling rates. In SRS, each population element has the same probability of being selected for the sample. For example, if the population has 1,000 people in it and the sample is 100, then each person has a 1 in 10 chance of being selected. Depending upon the question to be addressed, SRS can be inefficient.

To see why, suppose that we wanted to understand the source of wage differences between African Americans and whites and, further, that the firm employed 900 whites and 100 African Americans. An SRS sample of 100 would, on average, include 90 whites and 10 African Americans. This might be a reasonable sample if the goal was to understand the average earnings for the whole firm but, given that we want to understand group differences, this would not provide enough information about the earnings of the African Americans. Basically, we would have a precise measure of average income for whites but an imprecise measure for African Americans which, together, would result in an imprecise measure of the *difference* between the groups.

One solution to this problem is to sample the African Americans at a higher rate than the whites, what statisticians term oversampling, so that, for example, the final sample had 50 whites and 50 African Americans. Such a sample would have a somewhat less precise measure of average white earnings but would have a much more precise measure of earnings for African Americans. The net effect would be a more precise measure of the difference in averages between the two groups. For many of its demographic surveys, the federal government employs different sampling weights for individuals defined by race or location.

This discussion indicates that the best sampling schemes will often depart from simple random sampling. In particular, stratification, clustering, and differential sampling weights, either separately or in combination, can increase the likelihood of an accurate sur-

   ISSN 0418-2693

vey result, i.e., a sample calculation that closely matches the population it is designed to describe. It is important, however, that these properties of the sample be spelled out and understood *prior* to the execution of the survey. Surveys often depart from SRS in ways that are not well-understood or well-documented and, in such instances, there may simply be no way to use the surveys to arrive at statistically valid statements about the population as a whole. The next section of this article describes one mechanism by which such nonrandom samples are often developed.

Stratification, clustering, and differential sampling weights, either separately or in combination, can increase the likelihood of an accurate survey result, i.e., a sample calculation that closely matches the population it is designed to describe.

Stealth Samples.

Attorneys and experts often do exploratory analysis on the individuals for whom data is most readily available and then, later and without any explicit justification, interpret the resulting database as being a ''sample'' when, in fact, the database has none of the properties of a statistically designed random sample.

As an example, in one recent case attorneys were interested in understanding the economic loss for a large population of employees laid off by a single employer. Attorneys for plaintiffs sent out questionnaires to each class member and then commissioned a damages analysis for the first 10 questionnaires returned. There is nothing inappropriate about analyzing these particular 10 individuals, but there are significant problems with interpreting the average damages for this group as representative of the average damages of the entire population. In particular, the class members most engaged with the litigation may be both the most likely to respond quickly and the most likely to have experienced difficulties following their layoff. It is difficult to interpret this process as generating a genuinely random sample and, as a result, there may be no statistically valid way to extrapolate from the questionnaires received to the broader population of laid-off employees.

The attorneys in this case did not seek to draw an unrepresentative sample. Indeed, they did not appear to have thought about their process as even representing a sample in any formal statistical sense. Rather, the attorneys collected those data that were most readily available and then later sought to extrapolate, using rudimentary statistical techniques, to the entire population. Unless the sample was well-designed and properly documented, however, there is no assurance that the court or the opposing side will accept the sampling method as scientifically valid.

Calculating ''Standard Deviations'' When the Entire Population Is Sampled.

There is an interesting philosophical issue associated with the notions of samples, population, and the reliance on standard deviations as a measure of a study's reliability per *Hazelwood School District v. United States*, 433 U.S. 299, 15 FEP Cases 1 (1977), and other decisions. To review, *Hazelwood* stated that an inference of discrimination may be drawn if the group differences in outcomes were equivalent to 2-3 standard deviations or more. In most contexts, the *standard deviation* of a result, say the difference in pay between African Americans and whites, is calculated as the calculated pay difference, say $10, divided by the *standard error* of that difference. If the standard error is less than $5, then the resulting standard deviation will be more than 2 and the result will be presumed to be statistically significant. The standard error reflects the fact that, were we to draw a different random sample, we might get somewhat different results. This notion obviously relies on the assumption that it is possible to draw a different random sample than the one that the statistician has access to.

But what do we make of this concept when we make calculations on *all* the employees at a firm at a particular point in time? The sample is then the entire population and, were we to draw the same sized sample again we would get precisely the same sample and, by extension, exactly the same result. In this case, which happens frequently in litigation, there is no sampling variability and, as a result, the standard error is zero and the standard deviation is not defined because one cannot divide by zero. From this perspective, which is commonly held among survey statisticians, the *Hazelwood* decision is problematic when applied to cases where the entire population is sampled. Essentially, the key result of the analysis, the standard deviation, is not defined.

The logic of the *Hazelwood* decision may be salvaged if we think of the current population of a firm as being a sample of a much broader population that the firm itself draws from using its particular employment policies. In this case, we could think of the current workforce as a sample and, in our statistical calculations, we are considering how much our results might vary in the event that the employer had resampled its current workforce using the same employment policies to draw from that broader population. This is not an unreasonable position and is consistent with the way many economists and statisticians think about measuring statistical uncertainty in other economic environments. However, it is worth noting that in these kind of cases, the standard deviations to which so much attention is paid are not related to sampling variability per se, but rather to an often implicit assumption that the firm itself has drawn a sample of employees from some potentially larger pool.

Conclusion.

With large class actions apparently here to stay in employment litigation, there is likely to be continued interest in the use of sampling techniques to get answers to important questions at a reasonable cost.

When nonstatisticians think of sampling, they typically have in mind simple random sampling that can be thought of as throwing all population elements into an urn and then picking out a sample at random. To improve accuracy of samples, however, economists and statisticians often implement more complicated sample designs that make use of stratification, clustering, and different probabilities of selection. While these methods are somewhat more difficult to understand, they can provide more accurate answers to litigation-driven questions than does SRS sampling.

This article has also noted that attorneys, experts, and other participants in the litigation process often try to extrapolate from ''samples'' using statistical techniques. This process is only statistically valid, however, if the sample is drawn using standard sampling techniques. Further, if one side wants the other to embrace its use of a sample, it is often best to discuss the mechanics and implementation of the sample design before the sample is actually executed. To do otherwise often invites mistrust that the sample may have been nonrandomly picked—''cherry-picked''—to generate a desired result.

  DLR ISSN 0418-2693