NERA
Economic Consulting

5 August 2009

Complex Sampling for Litigation

By **Dr. Eugene P. Ericksen***

Introduction

Parties in complex lawsuits often require extensive data collection to provide the necessary evidence to prove their cases. The data collection can involve personal interviews, examination of transactions, reviews of personal files, or observing actual behavior. In many cases, the populations of interest are large, and collecting information on all cases is prohibitive in terms of cost and time. Statistical sampling can provide solutions that are relevant, cost-effective, and precise.

Such procedures can be illustrated by the case of *Mobil Oil v. Husky Corporation*.[1] In this case, Husky had manufactured nozzles to be used by customers at Mobil gas stations. They were intended to trap the gasoline vapors to prevent their escape into the atmosphere. Many of the nozzles unfortunately also trapped small amounts of gasoline after use, so that later customers sometimes spilled gas on themselves. I was asked to select a sample of about 100 nozzles from a set of 8,200 nozzles gathered at a warehouse and stored in 24 crates.

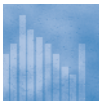
To select the sample, I randomly chose 8 of the 24 crates, grouped the nozzles from each selected crate into 10 piles of approximately equal size and randomly selected one of the piles into the sample. This created a sample of about 270 nozzles from which I selected a final sample of 106. I sent these to an engineer who tested them and found 50, or 47.2%, to have the gas-trapping flaw.

It is unlikely that a sample estimate will be exactly equal to the population percentage being estimated. The statistician takes this uncertainty into account by calculating a confidence interval. In this case, the 95% confidence interval stretched from 38.9% to 56.6%. This was the most likely range of population values, and it was sufficiently narrow to be relied upon at trial. The time and cost of selecting the sample and testing the selected nozzles, about two weeks, was a fraction of what it would have taken to test all 8,200 nozzles.

* Eugene P. Ericksen is a Special Consultant with NERA Economic Consulting. The author thanks Melissa Pittaoulis and Healey Whitsett for their valuable contributions to the paper.

Statistical sampling follows a codified set of procedures that are scientifically well established. There are two basic principles. The first is that probability sampling must be used. Probability sampling is defined by the statement that every population element must have a known, non-zero chance of selection. In the example just described, this probability was 1 in 77 for each of the 8,200 nozzles. The second is that the sample estimates and confidence intervals must be estimated in a manner consistent with the method used to select the sample. For example, if different sampling rates are used for different parts of the population, the variations must be accounted for by appropriate weighting. If weights are used, or if population elements are selected in clusters, the confidence intervals must be calculated by methods that are more complex than the standard formulas given in basic textbooks.[2]

The *Manual for Complex Litigation* sets forth seven principles by which survey evidence may be evaluated by the Courts.[3] The second of these seven principles states that the selected sample must be representative of the population. This is guaranteed by the use of probability sampling methods. The fourth principle states that the data are to be analyzed in accordance with accepted statistical principles. This includes the correct calculation of confidence intervals.



In the remainder of this article, I review the sampling methods and give several examples about how they have been used and misused in litigation. The intent is to inform readers about how samples used in litigation can be evaluated for *Daubert* motions, especially in situations where more complex methods need to be used.

Simple Random Samples

The most basic sampling method is called "simple random sampling" and this is the method that is assumed by most of the statistical procedures described in textbooks. The first step in selecting a simple random sample is to assign a unique number to every element of the population. For example, if there were 3,000 members of a plaintiffs' class, then each of these members would be assigned a unique number between 1 and 3,000. If I had selected the sample of nozzles by simple random sampling, I would have numbered them from 1 to 8,200.

The second step is to select randomly from among the numbers that have just been assigned. If the desired sample size is 300, I would use a random number generator to select 300 numbers. Because each selection is independent of every other selection, it is possible that some elements could be selected more than once. If this happens, then such multiple selections would be included in the sample as many times as they were selected. If the sample of 300 included 280 elements selected once and 10 elements selected twice, I would count each of the 10 doubly selected elements two times when calculating sample averages or other statistics.

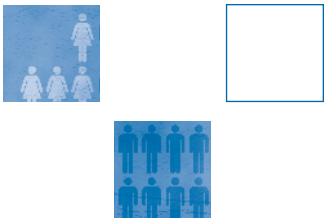
Standard errors and confidence intervals can be calculated by straightforward and well-known methods. Sample results are typically presented in two parts. The first is the "best estimate" obtained by a simple tally of the sample data. For example, if 130 of the 300 selections from the plaintiffs' class described above claimed to have suffered a particular injury, the "best estimate" would be 130/300 = 43.3%. I would also present a confidence interval, which is the best estimate plus or minus a "margin of error." This in turn is the product of the standard error multiplied by an appropriate "t-score." For our example, the standard error is .0286, the t-score is 1.97, and the 95% margin of error is .0286*1.97 = 5.6%. I would present the 95% confidence interval as ranging from 37.7% to 48.9%.[4]

Simple random samples are probably the most commonly used samples in litigation, and if the investigator is careful, these can be selected and analyzed by experts with only modest training in statistics. Problems are more likely to occur when experts take shortcuts or when more complex sampling methods are needed.

Why Probability Sampling is Important

Probability sampling is important because populations vary. Interviewers are usually not able to identify "typical" people when they set out to find them. When they are sent to designated blocks and told to find people to interview there, interviewers tend to select higher income people living in nicer houses.[5] Such methods created biases in favor of Republican candidates in political polls taking place before 1950.[6] The biased selections were often made subconsciously, but interviewers tended to favor people who seemed more likeable, had higher incomes, and were therefore more likely to vote Republican.

In litigation surveys, the use of probability sampling is especially important because of the need for trustworthiness. This is especially true when the person taking the interviews is also the person analyzing the data and testifying to the results. The recent case *Floorgraphics Inc. v. News America Marketing In-Store Services, Inc*.[7] provides a good example of the point. In this case, the plaintiffs' expert witness conducted interviews with 28 individuals whom the expert knew personally or was referred to by one of the individuals already interviewed. His goal was to establish whether certain statements made by the defendants' employees had damaged the plaintiffs' business. Because the sample was selected in such a biased manner, the Court deemed the expert's conclusions to be suspect and excluded his testimony.[8]



The use of a non-probability sample created problems in a second recent case that did not involve personal interviews. In the case of *Tiffany Inc. and Tiffany and Company v. eBay Inc*.,[9] the plaintiffs alleged that counterfeit Tiffany goods were being sold in large numbers on the eBay website. To demonstrate that the problem was serious, plaintiffs' expert attempted to select a probability sample of Tiffany silver jewelry up for sale on eBay. He selected daily samples for two two-week periods in 2004 and 2005. The statistical problem occurred because the population of Tiffany goods sold on eBay was constantly changing and different silver jewelry items were up for sale for different numbers of days. Without knowing how long each sample item had been up for sale, plaintiffs' expert had no way of calculating the probabilities of selection and establishing the proper weights needed for data analysis. The court did not exclude the expert's testimony, but accorded it limited weight, in large part because the sample was not a probability sample, making it impossible to calculate confidence intervals or otherwise generalize to the larger population.[10]

Mall Intercept Surveys

Mall intercept surveys are one of the most frequently used litigation survey methods, especially in intellectual property cases. They are not probability samples, but their frequent use, along with the lack of plausible alternatives, creates a "gray area" in the consideration of scientific standards. Mall intercept surveys are efficient because shoppers congregate in malls and they are necessary because these surveys typically involve showing products or pictures to respondents. Shari Diamond, writing on survey standards in the *Reference Manual on Scientific Evidence*, points out that mall intercept surveys are based on non-probability samples and states that "[s]pecial precautions are required to reduce the likelihood of biased samples."[11] She also points out that such surveys are frequently accepted by the courts, noting that one court accepted such a survey because "results of these studies are used by major American companies in making decisions of considerable consequence."[12,13]

Mall intercept surveys in trademark confusion or false advertising cases frequently have randomized designs. Half the respondents are randomly assigned to a treatment group, where they might see the allegedly infringing product, and the other half are assigned to a control group, where they would see a different product. The difference in the percentages of respondents identifying the product as having a certain brand is a frequent statistic of interest. Critics evaluating such studies do not frequently argue for their exclusion because of the lack of a probability sample of mall shoppers. It is difficult to construct a good argument about how the non-probability sample design affects the comparison between treatment and control groups.

There is little reason or evidence to believe that mall shoppers differ in important ways from consumers who do not go to shopping malls. I recently completed a telephone survey, based on a probability sample, in which I asked respondents how many times that they had gone to a shopping mall in the past year. I found that 93% had gone at least once and that 87% had gone at least twice. I calculated similar percentages for subgroups defined by demographic variables such as age, race, gender, and income, and observed that the only variable related to the frequency of mall visits was income. Unemployed and low-income people were slightly less likely to visit shopping malls but substantial majorities of each of these groups still visited shopping malls regularly. For most surveys, weighting the data by the frequency of recent shopping mall trips has a negligible impact on the results.

I recently conducted a mall intercept survey to ascertain the impact of the content ratings assigned by the Entertainment Software Rating Board on decisions to purchase a video game. I also interviewed a random sample of telephone respondents regarding the stores at which they had bought the game. This was a potential source of bias because some shoppers bought the game at electronics stores, frequently located in malls, but many others had bought the game in "big box" stores such as Wal-Mart that are typically not located in malls. In the mall-intercept survey, I asked the respondents where they had bought the game, and I weighted the results so that the distribution of purchase locations of the shopping mall survey respondents was the same as that of the telephone survey respondents. This minimized the potential bias of the shopping mall survey. The weighted and unweighted results were similar, suggesting that the potential bias was not large.

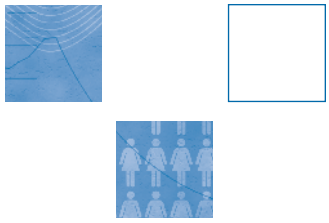
Web Surveys

Survey researchers have increasingly been relying on Web surveys, which are self-administered questionnaires that are completed online by respondents. In recent years, I have seen an increase in the use of Web surveys in litigation. Web surveys offer several advantages to researchers, mostly related to the time and cost of conducting the survey. Web surveys can often be completed in a short period of time since invitations to participate in the survey can be sent out to many people at the same time. They are also typically cheaper than other modes of survey administration such as telephone and face-to-face interviewing, making them appealing, cost-effective alternatives to more traditional telephone or mall-intercept surveys.

Because of the ease with which they can be conducted, the quality of web surveys can vary dramatically. Few web surveys use probability samples. Designing a probability sample for use with a web survey requires certain conditions, such as the availability of a list of email addresses for population members. For most populations, including consumer populations, there is no such list available and survey researchers rely upon volunteer panels of Internet users. Panel members are recruited by invitations on popular websites, banner ads, and Internet portals, and, after joining the panel, are periodically invited to participate in surveys. The major drawback of surveys completed by panel members is that they may not be representative of the population.[14] Web surveys using online panels must overcome two hurdles—for the variable of interest, they must demonstrate that Internet users and non-Internet users are similar and that panel members are not different than Internet users who do not belong to the panel. This can be hard to show. Studies have repeatedly shown that access to the Internet differs on characteristics such as age, race, education, income, urban/rural location, and region of the country.[15] For example, higher income families and college graduates are more likely than average to have Internet access. Blacks and Hispanics are less likely than average to have home Internet access. In 2003, the Census Bureau surveyed the American population to learn about the extent of Internet usage at home. The Bureau estimated that 64% of adults 25 to 44 and 64% of adults 45 to 64 were Internet users, but only 34% of adults 65 and over used the Internet.[16] Researchers employing a Web panel design must consider the extent to

which a non-representative panel biases the results of the study and if possible calculate adjusted results that account for the bias.

A second issue with Web surveys is the frequency with which panel members respond to surveys. It is accepted industry practice to disqualify prospective respondents who have participated in a survey during the preceding three-month period in order to avoid people whose experience with surveys creates points of view making them different than the average consumer. For example, such respondents may be more sensitive to ads they see, they may be more sensitive to variations in packaging they encounter, and they may be more sensitive to the goals of the survey, making them susceptible to demand effects.[17] Having a sample of respondents more sensitive to these factors than the actual population of consumers renders the sample unrepresentative of the population. Any differences that might exist between consumers with Internet access and consumers in general would be magnified if the consumers with Internet access were dominated by frequent survey takers who are sometimes known as "professional respondents." While research on frequent survey respondents is limited, available results are pertinent. For example, a telephone survey of people with Internet access found that 65% did not belong to any online panels, 11% belonged to just one such panel, and the remainder belonged to more than one panel. A separate survey of online panel members found that half of them also belonged to a second panel. A third study found that frequent survey takers belonged to an average of seven panels, and took an average of approximately one survey per day.[18] While I do not suggest that Web surveys are unreliable, we should be cautious in their use. An understanding of these possible biases and the way in which they may affect the results of the study at hand is crucial to the successful application of a Web survey methodology. The potential biases are greater than they are for a mall intercept survey.



Survey Non-Response

Survey non-response can be a sampling problem if respondents and non-respondents differ in important ways. This limits the ability of researchers to generalize from the sample of respondents to the larger population from which they were selected. In other words, the survey is limited to people who would be willing to respond to the survey if asked. As a practical matter, non-response is not commonly an important source of bias. Election surveys have high non-response rates because they are taken so quickly, yet they are usually very accurate.[19]

Non-response occurs for various reasons, including failure to contact the proper respondent, the respondent refusing to participate in the survey, or the respondent being unable to participate. In part because of the increasingly sophisticated technology used by potential respondents to screen their telephone calls,[20] levels of non-response to surveys have been rising in recent years. The lack of information about non-respondents concerns researchers worried about possible differences between respondents and non-respondents.

Shari Diamond has advised the courts to use strict criteria for survey non-response when they consider the weight to be accorded to survey evidence.[21] Based on guidelines from the former US Office of Statistical Standards, she writes that while response rates ranging between 75% and 90% generally yield reliable and unbiased results, "[p]otential bias should receive greater scrutiny when the response rate drops below 75%. If the response rate drops below 50%, the survey should be regarded with significant caution…."[22] While Diamond does qualify that "[d]etermining whether the level of non-response in a survey is critical generally requires an analysis of the determinants of non-response,"[23] she implicitly assumes that high non-response yields biased statistics, and contradicts the consensus of scientific opinion. Non-response has been found to affect survey statistics only in certain instances,[24] and the response rate itself is better understood as "…only an indicator of the *risk* of non-response error."[25]

The current scientific literature shows that non-response is only a problem in cases "when the causes of non-response are linked to the survey statistics measured."[26] One such example occurred at the onset of the HIV epidemic, as public health officials attempted to administer a survey to determine the prevalence of HIV in the population. HIV-positive persons were less likely to respond due to the stigma associated with the disease. As such, the survey's key statistic—the proportion of the population that is HIV positive—was a biased underestimate.[27] In a follow-up study of non-response to a more recent Canadian survey measuring Canadians' vehicle driving habits, researchers found that several statistics of interest were upwardly biased since respondents tended to drive more frequently than non-respondents.[28]

When the cause of non-response is not linked to the statistics measured, however, negligible differences may exist between respondents and non-respondents, and non-response bias will likely be low or nonexistent. Researchers have used several techniques to demonstrate that there may be little relation between non-response and non-response bias. For instance, in a secondary analysis of the Survey of Consumer Attitudes, researchers identified various sub-groups of respondents that were difficult to interview (i.e., multiple callbacks and refusal conversions). These sub-groups were excluded from the analysis, and the results were compared to the estimates derived from the total sample of respondents. The exclusions made little difference in the value of the estimates, suggesting that in this case, a lower response rate would not have produced biased estimates.[29] Similarly, Keeter et al. found negligible differences when comparing the results of two national telephone surveys that used the same survey instrument but obtained very different response rates—36.0% and 60.6%—due to variations in effort put forth to recruit respondents.[30,31] When this study was replicated in 2006, the researchers drew similar conclusions.[32] Recently, the director of polling at *ABC News* noted that polling prior to the 2008 presidential election was not substantively affected by non-response by cell phone-only segments of the population.[33]

The different effects of non-response in these examples illustrate the reason why, as the American Association of Public Opinion Research (AAPOR) has noted, "[t]here is currently no consensus about the factors that produce the disjuncture between response rates and survey quality."[34] Sometimes lower response rates introduce bias to survey estimates, but sometimes they do not; overall, there is no clear relationship between response rates and the quality of a sample survey. Because there is no one clear interpretation of a survey's response rate, researchers must evaluate response rates and potential non-response bias on a case-by-case basis and after considering a variety of factors about the respondents, the non-respondents, and the survey instrument itself.

Why Appropriate Variance Estimation is an Issue

The major benefit of a probability sample is that the statistician can generalize the results from the sample to the larger population within a specified margin of error. This is known as the confidence interval, a range of values in which the true population value is likely to fall. The correct calculation of the confidence interval assures that the precision of the sample estimate is appropriately stated.

A confidence interval consists of a sample estimate plus or minus a margin of error. As described earlier, the margin of error is the product of the standard error multiplied by an appropriate "t-score." The appropriate t-score is determined by the sample size and the level of confidence desired by the statistician. In most cases, the 95% confidence level is used. For small samples, the appropriate t-score for a specified sample size can be found in a table known as "Student's t-distribution." For larger samples, generally samples of size 500 or more, the t-score for a 95% confidence interval is 1.96. Determining the appropriate t-score is straightforward, and anyone who has successfully taken a basic statistics class can do so.

The second component of the margin of error is the "standard error." The standard error is the square root of another statistic known as the "variance." The variance is the average of the squared deviations divided by the sample size. This is the formula taught in basic statistics courses, although

many textbooks and instructors neglect to tell their students that the formula is based on the assumption of simple random sampling.

The situation becomes more complex when data are weighted. Statisticians weight data for a variety of reasons, including the need to account for variations in sampling rates, the adjustment for non-response, or the fact that some observations are more important than others. The use of weighted data creates two problems: (1) the variances are increased, and (2) the formula appropriate for simple random sampling no longer works. More complex methods such as balanced repeated replications,[35] ratio estimation,[36] and the bootstrap[37] are needed.

For example, in the antitrust case *ConMed v. Johnson & Johnson*,[38] I selected a simple random sample of hospitals for a survey concerning their purchases of endoscopic surgery products. After an initial analysis, I wished to see if the results would be changed if I weighted the data by the numbers of surgeries performed at each hospital. In other words, the larger hospitals, which purchased more endoscopic surgery products, might have given different types of answers than the smaller hospitals did. I found that the weighted variances were about 10% larger than the unweighted variances. The ratio, 1.10, is known as the design effect and this is a useful indicator of the gain or loss in precision due to the use of weighting or some other sampling strategy different than simple random sampling.

Weighting is often necessary for a survey with multiple objectives. To illustrate, the statistician may be interested in obtaining the opinions of two groups of physicians, cardiologists and radiologists, and comparing them. The estimated comparisons achieve maximum precision when the two samples are equal in size, as this minimizes the standard errors of the average differences between the groups.

If the population includes twice as many radiologists as cardiologists, we must select the cardiologists at twice the rate of the radiologists to have equal sample sizes. If, however, we also wish to conduct some analyses of the combined population of physicians, we would use weights that were inversely proportional to the probabilities of selection. In other words, we would give each radiologist a weight of 2 and each cardiologist a weight of 1. Two-thirds of the total sum of weights would be assigned to the radiologists, reflecting the fact that they comprise two-thirds of the combined population. This use of weights provides unbiased estimates of the combined population, but the standard errors are larger than they would be without the weights.[39]

Using Cluster Sampling

Cluster sampling is a second important departure from the simple random sampling model. By selecting groups of respondents together, data collection costs are lower, but confidence intervals may be widened. The use of cluster sampling is cost- and time-efficient, but the elements in the same clusters may be similar to each other. If so, this will widen confidence intervals. For example, if someone wished to survey the citizens of Philadelphia to assess their economic situations, (s)he might first select a sample of 100 blocks and then interview five households in each sample block. The respondents in richer neighborhoods would consistently have above-average incomes and the respondents in poorer neighborhoods would consistently have below-average incomes. The grouping of similar cases by cluster reduces precision because the information obtained from one respondent is not independent of the information obtained from another respondent in the same cluster. Although there are 500 respondents in the survey the standard errors may be considerably larger than would be the case for a simple random sample with 500 respondents. The loss of precision due to clustering is proportional to the degree of similarity among respondents selected in the same cluster.

In many cases, this similarity is not substantial. In the example of the gasoline nozzles described earlier in this paper, the homogeneity of clusters was very low and there was little, if any, loss of precision due to the use of a cluster sample. In other cases, the similarity may be substantial, and the statistician always needs to make the proper calculations to find out.

This is illustrated by a recent survey of drugstores. Here, the objective was to determine the extent to which a pharmaceutical product was being exported overseas, bought at a cheaper price, and reintroduced to the United States, thereby undercutting the profits of the manufacturer. To evaluate this, I selected a sample of drugstores, an average of 10 in each of 50 metropolitan areas. Because two drugstores in the same area usually bought their products from the same distributor, the answers they gave were similar. When I calculated the standard errors, I observed that the design effect was equal to 3. In other words, my clustered sample of 500 drugstores had the same level of precision as a simple random sample of 167 drugstores would have had. The cost of taking a survey based on such a simple random sample would have been more than three times as great as the actual costs, so the clustered design was the best strategy to use.

In summary, sample designs involving clustering or weighting require the more complex estimation of standard errors and confidence intervals. If instead of the variance estimation methods appropriate for use with cluster samples, a statistician used the SRS formula for calculating variances, (s)he would underestimate the variances and report confidence intervals that are too narrow, suggesting that the sample results are more precise than they actually are. Such a mistake violates the fourth principle in the *Manual for Complex Litigation*.[40]

Examples of the Use of Sampling in Actual Legal Cases

In this section, I give examples of sampling problems in actual cases and explain some of the key strategic decisions. First, in a survey of cardiologists and radiologists, the goal was to estimate the shares of physicians who were misled by certain marketing materials. I obtained a list of eligible physicians from the American Medical Association (AMA) and selected a simple random sample of these physicians to be interviewed by telephone. The critical decision was to determine the necessary level of precision.

For the key statistic, we expected that between 25% and 75% of the sample would say that they were influenced by the marketing materials. We desired a confidence interval of plus or minus 5 or 6 percentage points, and a sample of size 300 is sufficient to provide this level of precision. Any value in this range is sufficient to demonstrate the influential nature of the marketing materials at issue so there was no reason to bear the additional cost of a larger sample.

I therefore selected a simple random sample of 300 physicians. In the actual survey, 41% of the sample found a particular marketing brochure "influential on their decisions," and I concluded that between 35% and 46% of all physicians on the AMA list would have given such a response. This level of precision was sufficient to demonstrate the point and the size and nature of the sample was not a controversial issue in this case.

Using Unequal Probabilities of Selection

Simple random sampling is the best sampling technique when all population elements are equally accessible and equally important. This was true for the physicians on the AMA list in the case just discussed. There are other situations, however, where certain parts of the population are more important than other parts, and a different sampling procedure is needed.

This is illustrated by a case in which a plaintiff company sued the firm it had hired to manage its employee health plan. The major allegations were that the management firm had not monitored individual claims properly and the plaintiff company felt that it had overpaid on such claims. There

were millions of claims and to assess mismanagement a team of CPAs needed to look at them individually. Sampling was called for, but with a special twist.

Larger claims, more complex and often with several parts, were thought to have a higher rate of error than smaller, usually simpler claims. To take this into account, I selected claims with probabilities proportional to size (pps). For example, a $1,000 claim had 10 times the chance of selection as a $100 claim and a $5,000 claim had 50 times the chance of selection. For example, if 10% of the claims but 50% of the dollar values of the claims were in the $1,000 plus category, then 50% of the sample, rather than 10%, was selected from this category. The basic idea is that each dollar is equally important and therefore should have equal probability of selection. Accounting firms refer to this type of sampling as "dollar-based sampling." In this particular case, I selected a sample of about 1,000 claims from a population of six million. The estimated loss was $34 million with a margin of error of $8 million.[41]

Cluster Sampling

In the examples described so far, I obtained large savings by the use of sampling. There are other situations, however, where the geographic or physical arrangement of population elements makes them expensive or difficult to reach. For example, in an employment discrimination case against a large national corporation with multiple office locations, I needed to obtain information from individual employees and their records. Rather than selecting a simple random sample, I selected a clustered sample in two stages. In the first stage, I randomly selected a set of offices. In the second stage I randomly selected employees from each selected sample office. This allowed me to concentrate the search for records in a small subset of offices and to have employee interviews conducted in that same set of locations, saving substantial amounts of money and time. There were 120 interviews in 30 office locations, and they provided satisfactory estimates and confidence intervals for the variables of interest.

I selected a different type of cluster sample in a commercial litigation matter involving call recordings from a national call center. I designed a sample of the one million calls that had been made to the call center during the relevant time period. They had been saved onto 122 DVDs. Our job was to listen to and analyze the sample of recorded calls. Not only was the review of one million calls prohibitively expensive and time-consuming, but so too was a simple random sample of such calls. This is because there was no listing of the calls placed on each DVD. Creating such a list, sampling from it and then reviewing individual recordings of calls would have been very expensive and time-consuming.

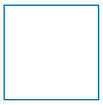
I instead selected a cluster sample. To do so, I first randomly selected 25 of the 122 DVDs. Next, I created a list of all the calls on these 25 DVDs, numbered them from 1 to the total of 206,969. I then selected 52 random numbers between 1 and 206,969. For each selected random number, I created a cluster of 50 call recordings by identifying the next 49 calls. For example, the first selected random number was 4,911 and the sample therefore included calls 4,911 through 4,960. I repeated this for each of the 52 random numbers. The total number of selected call recordings was therefore 50*52 = 2,600. When the people hired to review the calls did so, they saved substantial amounts of time and money by successively listening to the 50 calls in one cluster, usually doing so in less than one day.

This resulted in a manageable number of call recordings to listen to, analyze and summarize. The coding work for just this sample of 2,600 calls took approximately three months to complete. Had I not selected a statistically valid sample of calls, the analysis would have been inconclusive, time-consuming, and expensive.

With cluster sampling, the precision relative to simple random sampling is reduced. This is because, as described earlier, population elements in the same cluster are more likely to be similar than population elements in different clusters. There are special procedures designed by statisticians to estimate standard errors and confidence intervals for a clustered sample and these standard errors and confidence intervals are usually wider than they would be for a simple random sample of the same size. A skilled statistician is able to minimize the decrease in precision due to clustering, known as the "design effect," but it is mandatory that the clustering be taken into account for such calculations.

Continually Changing Populations

Simple random sampling is not the method of choice when the population of interest is continually changing. This occurs for populations such as Internet auction sites or video collections on websites. At any moment in time, population elements can be added or withdrawn and it may not be clear what the actual probabilities of selection are. With population elements spending different amounts of time on a website, the probabilities of selection are not likely to be equal without special sampling procedures being applied.



In the case of *Tiffany Inc. and Tiffany and Company v. eBay Inc.*,[42] I was asked to evaluate a sample presented by the plaintiff's expert witness. The expert selected a sample of Tiffany items that were for sale on eBay in an attempt to estimate the percentage of all Tiffany items for sale on eBay that were counterfeit merchandise. To do this, on each day during a 10-day period, the expert conducted a search for eBay items containing the phrase "Tiffany sterling," printed the search results, and selected a systematic random sample[43] from the items on this list. He then attempted to purchase each of the items that were selected for inclusion in his sample. One of the most severe problems with this methodology was that he did not know the probabilities of selection for individual items, and these are proportional to the numbers of days that items were listed on eBay. Had he not disrupted the eBay procedure by buying and then removing items from eBay, he would have known the number of days each item was available for purchase. For example, the items he selected and bought on Monday might otherwise have been available on Tuesday through Friday. He had no way of knowing how many additional days each item would have been available, so he was not able to calculate what the actual probabilities of selection were. Because he did not know the probabilities of selection, he could not properly weight the data in inverse proportion to these probabilities to calculate unbiased estimates or confidence intervals for those estimates.

However, there is a straightforward way that the expert could have selected an equal probability sample of Tiffany items on eBay—by restricting the chance of selection to the first day that an item appeared on eBay. Methods for doing so are described in the scientific literature.[44] For example, the expert could have selected his Tuesday sample as he did, and then checked each of his selections against the list of items that were available on Monday. If the selected Tuesday item was also available on Monday it would be deleted. Otherwise it would be included in the sample. He then would have selected a Wednesday sample, checked each of the sample selections against the lists of items that were available on Monday and Tuesday and deleted each sample selection that had already been listed. He would have continued the procedure on Thursday and Friday. This expert's limited statistical sampling expertise led to a flawed methodology that was accorded little weight by the court.

Summary and Conclusions

Statistical sampling can be used to create useful datasets representing large and diverse populations. In many situations, simple random sampling methods can be used to provide a valid sample quickly. I recommend such a strategy if all population elements are equally important or relevant to the issues at hand, if they are conveniently listed on hard-copy or in computer data files so that sampling is easy and straightforward, and if they can be contacted in a quick and cost-effective manner. One especially common application of simple random sampling methods is the selection of respondents for mail or telephone surveys.

There are many situations where simple random sampling methods will not work and more complex methods are needed. This occurs when different population elements have different levels of importance. To illustrate, radio stations have different audience sizes, medical claims have different amounts, and customers buy different amounts of goods. In these situations, selection probabilities should be proportional to size. In a study of misleading advertising materials, a customer who bought twice as much of the product at issue than another customer would be twice as important and should have twice the selection probability. To interpret the survey findings, we should say that "X percent of the product was bought by customers who were misled" instead of saying that "X percent of customers were misled."

Cluster sampling is a useful procedure when population elements are costly to access. This can occur when data are to be collected by individuals, such as a survey of stores, interviews with employees, or collections of records in a discrimination case, or when records are densely packed on computer files. Simple random sampling methods may not provide the desired savings of time and money, so we would turn to cluster sampling methods. These methods can be time- and cost-efficient, but we must take care to use appropriate methods for estimating variances and confidence intervals.

More complex procedures are also needed for situations where populations are constantly changing. In order to generalize from the sample to the population, the statistician must be able to identify the probability of selection for each sample observation. Procedures for doing so exist, but they are likely to require a highly developed level of statistical expertise.

Selecting the best procedure and creating a valid sample design requires statistical expertise and experience. The expertise will direct the statistician to an optimal design that will provide the greatest precision for a given amount of money to be spent on the project. The experienced statistician will know that such an optimal design may not always be best. This is because many studies will have multiple objectives and the optimal design for one objective may not be the optimal design for another objective. Furthermore, goals may change during the life of the project, data problems may necessitate design revisions even after the project has begun, and new issues may arise during the life of the study. The experienced statistician will have dealt with such problems in past studies and is likely to realize that when it comes to complex sample design, "the perfect may be the enemy of the good." In complex situations such as these, the statistician will create a flexible design that will retain the two necessary traits: (1) each population element will have a known, non-zero probability of selection, and (2) the method of estimating the precision of the sample estimate will be consistent with the sample design.

End notes

1 *Mobil Oil v. Husky Corporation*, (E.D. Va. 1996).

2 The standard formula for the margin of error multiplies the 't-score' by $s/(n-1)^{.5}$, where $s^2 = \sum(x - \text{mean})^2/(n - 1)$. The 't-score' depends on the desired confidence interval. For a 95% confidence interval and a moderately large sample size, the 't-score' = 1.96.

3 Federal Judicial Center. 1995. *Manual for Complex Litigation*. 3rd ed., p. 102.

4 While there is no rule indicating what the level of confidence should be, the most commonly used such level is 95%.

5 Carter, Roy E. Jr., Verling C. Trodahl, R. Smith Schuneman. 1963. "Interviewer Bias in Selecting Households," *Journal of Marketing*, 27:27-34.

6 Gawiser, Sheldon R. and G. Evans Witt, 1994. *A Journalist's Guide to Public Opinion Polls*. Westport, CT: Praeger Publishers, p. 21.

7 *Floorgraphics, Inc. v. News America Marketing In-Store Services, Inc., et al*., No. 03 Civ. 3500 (AET) (D.N.J., 2008).

8 *Floorgraphics, Inc. v. News America Marketing In-Store Services, Inc., et al*., No. 03 Civ. 3500 (AET), 38-42, 546 F. Supp. 2d 155, 2008 US Dist. LEXIS 8263 (D.N.J. 2008).

9 *Tiffany (NJ) Inc. and Tiffany and Company v. eBay Inc*., 576 F. Supp. 2d 463 (S.D.N.Y, 2008).

10 *Tiffany (NJ) Inc. and Tiffany and Company v. eBay Inc*., No. 04 Civ. 4607 (RJS), 19-21, 2008 U.S. Dist. LEXIS 53359; 2008-1 Trade Cas. (CCH) P76, 219 (S.D.N.Y., 2008).

11 Diamond, Shari Seidman. 2000,"Reference Guide on Survey Research" *Reference Manual on Scientific Evidence*, 2nd ed. Washington, DC: Federal Judicial Center, p. 244.

12 *National Football League Properties, Inc. v. New Jersey Giants, Inc*., 637 F. Supp. 507, 515 (D.N.J. 1986). As cited by Diamond, *Id*. at p. 244.

13 A survey of members of the Council of American Survey Research Organizations revealed that 95% of in-person marketing research studies conducted during September 1985 were conducted in shopping malls using non-probability sampling methods. Jacoby, Jacob and Amy H. Handlin. 1991, "Non-Probability Sampling Design for Litigation Surveys," *Trademark Reporter*, 81: 169-173, pp. 172-173.

14 As stated by Shari Diamond, "The key limitation [of web surveys] is that the respondents accessible over the Internet must fairly represent the relevant population whose responses the survey was designed to measure." See Shari Diamond, 2000. "Reference Guide on Survey Research," in *Reference Manual on Scientific Evidence*, 2nd ed. Washington D.C: Federal Judicial Center, p. 264.

15 Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches," *Public Opinion Quarterly*, 64:464-494.

16 See "Table 2B. Presence of a Computer and the Internet at Home for People 18 Years and Over, by Selected Characteristics: October 2003" from the US Census Bureau, Current Population Survey, October 2003.

17 Demand effects occur when the survey procedure provides cues about the intent of the study and respondents react to them by guessing what they believe to be the "correct" answer. See: Orne, Martin T. 1962, "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications," *American Psychologist*, 17(11): 776-783.

18 Miller, Jeff. 2006. "Online Marketing Research," *Handbook of Marketing Research: Uses, Misuses, and Future Advance's*. Eds. Rajiv Grover and Marco Vriens. Sage Publications, pp. 117-118.

19 See, for example: http://www.realclearpolitics.com/polls/, accessed 30 December 2008. See also http://www.washingtonpost.com/wp-srv/politics/polls/poll_response_rate.html.

20 A study by the Pew Research Center found that 88% of respondents to the survey have an answering machine, caller ID, or call blocking technology; 47% had at least two of these technologies. http://people-press.org/report/?pageid=813, accessed 11 December 2008.

21 Diamond, Shari Seidman. 2000,"Reference Guide on Survey Research" *Reference Manual on Scientific Evidence*, 2nd ed. Washington, DC: Federal Judicial Center, pp. 245-246.

22 *Id*. at p. 245.

23 *Id*.

24 Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004, *Survey Methodology*, New Jersey: Wiley-Interscience, p. 178.

25 *Id*., at p. 384 (emphasis original).

26 *Id*., at p. 178.

27 Horvitz, D.G, M.F. Weeks, W. Visscher, R.E. Folsom, P.L Hurley, R.A. Wright, J.T. Massey, T.M. Ezzati. 1990, "A Report of the Findings of the National Household Seroprevalence Survey Feasibility Study," Proceedings of the American Statistical Associations, Survey Research Methods Section, pp. 150-159.

[28] Beaulieu, Martin. 2006, "A Study of Nonrespondents in the Canadian Vehicle Survey," Proceedings of the American Statistical Associations, Survey Research Methods Section, pp. 2734-2740.

[29] Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000,"The Effects of Response Rate Changes on the Index of Consumer Sentiment," *Public Opinion Quarterly* 64:413-428.

[30] Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000,"Consequences of Reducing Nonresponse in a National Telephone Survey," *Public Opinion Quarterly* 64:125-148.

[31] A study conducted in 2003 by the Pew Research Center had quite similar results. There was little difference between a standard survey that obtained lower response rates and a survey using rigorous techniques to obtain a high response rate. http://people-press.org/report/211/, accessed 11 December 2008.

[32] Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill, 2006, "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," *Public Opinion Quarterly*, 70:759-779.

[33] http://www.prweekus.com/Polls-drive-news-narratives-to-the-election/article/120080/, accessed 11 December 2008.

[34] AAPOR is the leading association of public opinion and survey research professionals. http://www.aapor.org/responseratesanoverview, accessed 11 December 2008.

[35] For example, see: Levy, Paul S. and Stanley Lemeshow. 2008. *Sampling of Populations: Methods and Applications*, 4th ed. Hoboken, NJ: Wiley Interscience, pp. 373-381. See also: Lohr, Sharon L. 1999. *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press, pp. 298-303.

[36] Levy and Lemeshow 2008, Id. "Chapter 7: Ratio Estimation," pp. 189-222. Lohr *Id*. at pp. 60-71.

[37] For a basic reference on the bootstrap, see: Efron, Bradley and Robert J. Tibshirani, 1993. *An Introduction to the Bootstrap*, New York: Chapman & Hall.

[38] *ConMed Corporation v. Johnson & Johnson, Inc., Ethicon, Inc., Ethicon Endo-Surgery, Inc., Johnson & Johnson Healthcare Systems, Inc*. No. 03 Civ. 8800 (JES) (DF) (S.D.N.Y. 2007).

[39] For this example, the design effect equals 1.11 so the loss of precision is about 10%.

[40] Federal Judicial Center, 1995. *Manual for Complex Litigation*, 3rd ed., p. 102.

[41] The 95% confidence interval stretched from $26 million to $42 million.

[42] *Tiffany (NJ) Inc. and Tiffany and Company v. eBay Inc*., 576 F. Supp. 2d 463 (S.D.N.Y 2008).

[43] A systematic random sample is different than a simple random sample. In a systematic random sample, all items being considered for inclusion are first numbered sequentially. Selections are made by calculating the sampling interval, "k," equal to (number of desired items in the sample) / (total number of items).A random number between 1 and k is then selected; the numbered item that corresponds with this random number is the first selection. Subsequent selections are made by selecting each kth item as specified by the sampling interval.

[44] A description of an appropriate sampling method for this expert's situation is given by Eugene P. Ericksen and Joseph B. Kadane. 1986, "Using Administrative Lists to Estimate Census Omissions," *Journal of Official Statistics*, 2:397-414.

NERA
Economic Consulting

About NERA

NERA Economic Consulting (**www.nera.com**) is a global firm of experts dedicated to applying economic, finance, and quantitative principles to complex business and legal challenges. For nearly half a century, NERA's economists have been creating strategies, studies, reports, expert testimony, and policy recommendations for government authorities and the world's leading law firms and corporations. We bring academic rigor, objectivity, and real world industry experience to bear on issues arising from competition, regulation, public policy, strategy, finance, and litigation.

NERA's clients value our ability to apply and communicate state-of-the-art approaches clearly and convincingly, our commitment to deliver unbiased findings, and our reputation for quality and independence. Our clients rely on the integrity and skills of our unparalleled team of economists and other experts backed by the resources and reliability of one of the world's largest economic consultancies. With its main office in New York City, NERA serves clients from over 20 offices across North America, Europe, and Asia Pacific.

Contact

For further information and questions, please contact:

Dr. Eugene P. Ericksen
Special Consultant
NERA Economic Consulting
+1 215 864 3878
eugene.ericksen@nera.com