Statistical Sampling in Audit Case

Kwon I-W G[*]

Saint Louis University, St. Louis, MO, USA

[*]**Corresponding author:** Kwon I-W G, Ph.D. Professor Emeritus Saint Louis University, St. Louis, MO 63108, USA, Fax 314-977-2067, Tel: 314-560-5487, E-mail: kwoni@slu.edu

Introduction

Sampling methods and procedures impact outcomes in many different areas whether they are in compensations in product injury cases, jury selection, population estimation in census, forecasting national gross products, etc. Economic penalty and emotional injustice from faulty sampling methods ranges from life or death to irreparable economic and emotional damages. Nevertheless, literature on sampling methods and procedures in court cases has been surprisingly silent. Statistical sampling method was not used until recently and sampling methods and procedure have been rarely discussed in academic circles as well as the judicial arena. Even in statistical sampling methods in court cases, several alternate sampling methods such as stratified sampling method and systematic random sampling methods have not been fully exposed. In addition, literature is also silent on determining optimum sample size that would produce credible results in extrapolating outcomes from sample into population.

Keywords: Auditing; Statistical Testing; Simple Random Sample; Stratified Random Sample

Absence of such debates is odd considering a series of decisions rendered by the United States Supreme Court on using statistical sampling methods. For example, the Supreme Court allowed statistical inference in a series of decisions beginning with a jury discrimination case, Castaneda vs. Parida, decided in March 1977 (Meier, 1986) [1]. The U.S. Supreme Court repeated the same decision in 1977 in favor of statistical sampling evidence to establish liability and damages in a "donning and doffing" overtime class action under the Fair Labor Standards Act (FLSA) and state wage law. *Tyson Foods, Inc. v. Bouaphakeo*, No. 14-1146, 2016 WL 1092414 (U.S. March. 22, 2016). The 8-0 decision by the Supreme Court on Want vs. Fort Bank and Trust in 1988 reaffirmed the legality and legitimacy of statistical sampling methods used in court cases.

It appears that the Supreme Court decisions, nevertheless, opened a door to statistical sampling in various cases. Illinois Physicians Union et al. vs. Jeffrey Miller et. al in1982 on reasons for sampling, Ratanasen vs. State of California Department of Health Services in 1993 on sampling methodology and sample size, Chaves County Home Health Services, Inc. et al vs. Louis W. Sullivan in 1991 on representative sample size, United States ex rel. Martin vs. Life Care Centers of America, Inc in 2014 extrapolation of statistical evidence to population, United States ex el. Wall vs. Vista Hospice Care Inc. et al in 2016 on case of sampling selections, just to name a few (Fienberg and Straf, 1991) [2].

The purpose of this paper is to review sampling methods and procedures used primarily in court cases. Literature review was a major research method to synthesize historical development in sampling techniques in court cases. Stratified random sampling method will be presented in the areas where population includes heterogenous element (Nathan, 2011) [3]. In addition, this paper discusses how to determine optimum sample size that yields sample values to be extrapolated into population. This paper concludes with a case study on sampling method.

This paper consists of; Introduction, methodology/literature review, proper sampling method and procedure, determining optimum sample size, and conclusion with case study.

Methodology and Literature Survey

This study uses a literature survey to review the extent of studies and research on sampling methods in auditing in general and in court cases in particular. Literature survey has been extensively used in the academic area to investigate the landscape of particular issues in a historical prospective. Not only we learn what has been done in an area of interest but more importantly we learn from it a trend that a discipline has progressed up to a certain time period. A new methodology could be discovered by synthesizing the existing theories.

Literature in auditing in general and in court cases in particular reveals that non-statistical sampling methods (also called judgmental sampling) has been a major sampling tool. Although many factors could have contributed to the lagging understanding and use of statistical sampling techniques, Rogers and Elliott (1972) [4] list two main reasons; statistical techniques have generally been known as cumbersome to apply manually, and statistical techniques have not been adequately related to common audit objectives. Simply put it, auditors are not familiar with rigor of statistical frameworks for selecting an unbiased sample using a statistical framework. Messier, Jr. et al (2001) [5] suggest better understanding of nonstatistical sampling methods, which have gained increasingly an importance in the contemporary audit environment. It is no surprise, therefore, to see that statistical sampling procedures have nearly disappeared from practice (Gilbertson and Herron, 2003) [6]. They further argued that no studies then have directly tested the potential implications of sampling methods on jurors' verdicts, damage awards, or sample size expectations. According to their study, in a case where auditors were alleged to have used an insufficient sample size, the sampling methods (judgmental sampling vs. statistical sampling) did not affect the likelihood of a "guilty of negligence" verdict. However, damage awards were significantly higher when nonstatistical sampling was used compared with statistical sampling. Part of the reasons may be that nonstatistical sampling methods often fails to provide sampling errors, and are, therefore, prone to decision biases (Elder, et al, 2013) [7].

Ponomon and Wendell (1995) [8] published one of the first auditing studies comparing the performance of judgmental and random sampling methods, finding that statistical methods of determining bounds were superior to auditors' judgments of bounds. They go on to state that "firm management needs to be cautious when employing judgmental procedures for selecting and making subjective inferences about the underlying population under examination" (p. 33). While their study and several others highlight the potential danger in using judgmental sampling in practice, the superiority of statistical audit sampling has yet to be tested in courts

Nathan (2011) [3] suggests that sample size depends on characteristics of the population. According to his argument, in a population with homogeneous elements, the sample selected should be lower, in contrast for a population of heterogeneous elements, the sample selected should be higher, since for these latter elements there is a higher risk to be distorted as variance of population characteristics are too large and the amount of uncertainty and erroneous conclusion are higher than alternative. Accordingly, if an appropriate sample is not correctly selected, errors identified are wrongly extrapolated to the entire population. Extrapolation to population based on sampling results regardless the types of sampling methods carries injustice to the parties if the sampling process is flawed. Phillips et al (2017) [9] warned in a Medicare fraud case that by understanding the critical elements of appropriate statistical methods and how they should be tailored to the specifics of the case at hand, defendants can better respond to, and defend against, the use, and perhaps misuse, of sampling and extrapolation.

Sampling Methods: Simple Random Sample (SRS-1) vs. Stratified Random Sample (SRS-2)

When selecting a statistically acceptable sample size (n) from a pool of a large population (N where N > n with a great margin), most experts recommended a simple random sampling method (SRS-1). SRS-1 has been used extensively primarily because it is convenient, economical and simple to select and extrapolate. However, it should be noted that SRS-1 can be used only when

elements (also called units of analysis) in a population are homogenous (or same) across the entire population spectrum (Nathan, 2011) [3]. When and if the units of analysis differ by any known groups (also called stratum), e.g., age, race, gender, place of living etc., SRS-1 does not assure the samples so selected to represent the population. For instance, if a hospital is to be audited on Medicaid services by a state agency, it is assumed that the recipients of services may differ by types of service they receive (inpatients, outpatients, ER visits just to name a few), personal characteristics (gender, age, family history) and assortments of treatments they received measured by diagnosis related group (DRG). The amount charged to Medicaid recipients, accordingly, differs by the type of patients (inpatients costs more than outpatient visits), personal characteristics (older patients may use more resources than the younger patients) as well as types of DRG's.

If the population (patients) has different characteristics, SRS-1 does not assure a representative sample. Any statistical inferences based on such a non-representative sample may lead to biased interpretation of population value. Whether SRS-1 selects sample proportionally (representatively) to population can be tested through accepted statistical tests such as chi-square test and simple t-test.

Another sampling method that is often used includes a systematic sampling method. This method again has been practiced by some due in large part to a simple and convenient way of selecting samples. Suppose auditors need to take 100 samples or files(n) from a total 1,000 files (N, population) for auditing purposes, the auditor would take every 10th file (N/n = 1,000/100 = 10th). Although this sampling method seems simple enough for convincing many practitioners, it also has potential sources of bias as it does not assure selected samples to be representative of the population. It is especially true if and when the entire files (population) are arranged in a special order, e.g., patient records are arranged based on the total amount billed to Medicaid insurance.

If the population (e.g., patient records) is a mixture of different backgrounds based on, for example, types of service rendered (e.g., inpatients, outpatients, ER visits, etc.) which in turn are greatly influenced by personal characteristics such as gender, age, race, etc., then a stratified random sampling method (SRS-2) is an appropriate sampling method. SRS-2 first classified (stratified) population (N) based on unique characteristics into several strata (groups). The population now is separated by several stratum where each stratum contains homogenous elements (e.g., male for one stratum and female patients for the other stratum). Once the population is stratified into different strata, it is assumed that the population in each stratum is homogenous in major attributes, e.g., gender, DRG, age groups so on. Once each stratum houses a homogenous population, then SRS-1 can be applied. The sum of samples from each stratum should be equal to the total sample for the study. For example, if the total population of medical records shows 70:30 on gender in favor of female patients and 100 samples are planned, then 70 samples should be collected from female strata through the SRS-1 method and 30 from male strata. SRS-2 assures 100 samples were selected in proportion to population through unbiased SRS-1 method.

Once sampling for the audit has been properly carried out, then statistical analysis on major attributes, e.g., the amount of over/underpaid on Medicaid services, can be estimated and extrapolated into the population. Extrapolation could be a point estimate or confidence estimate usually using a 95% confidence level.

Determining Optimum Sample Size

Another area that has not been explicitly explored and discussed in sampling methods in court cases include the optimum sample size that yields a reasonably credible estimate of the population value. Although there are many theories how to determine optimum sample size, ranging from few samples (based on judgmental assessment primarily used in life science research) to as many as possible (census for every 10 years) and between. If uncertain, yet aiming at credible and statistically defensive sample size, one could use a statistically valid sample size method called optimum sample size. All one needs is several pieces of relevant information; confidence level measured by Z-score, level of uncertainty measured by p (if unsure, then use 0.5 to maximize possible sample size) and margin of error (how much one can tolerate error) measured by B. then, the statistical formula for optimum sample size is given by:

$$n = [(Z^2 * p(1-p)]/B^2$$

A simple illustration may help understanding this process. Suppose you would like to determine an optimum sample size statistically valid and a universally accepted manner. You would like to have the result within a 95% confidence (Z) with margin of error (B) at 0.01 or 1%. Since you are not sure the population distribution in a stratum (p), you decided to use the maximum uncertainty of 0.5 or 50%, meaning statistically speaking "I do not know". With the above background information, the optimum sample size that would withstand any challenge in a court case would be:

$$n = [1.96^2 * 0.5 (1-0.5)]/0.01^2 = 122.5 \text{ or } 123 \text{ sample sizes}$$

Mukherjee (2019) [10] illustrates in details the process to select an optimum sample size.

Conclusion with a Case Study

This paper reviews the current status of statistical sampling methods in court cases. It appears that experts seem to agree that statistical sampling methods produce better results compared to nonstatistical sampling methods. Yet, literature has been not convincing. Most cases use simple random sampling methods even in cases where the population consists of heterogenous elements. Stratified simple random sampling method which produces a superior outcome was rarely mentioned in the literature. The following case study illustrates the points.

Case Study

Up to 10% of the United States health care spending is estimated to be lost from overpayments in the form of fraud, waste and abuse (CMS (Center for Medicare and Medicaid Services), 2015) [11]. These health care overpayments range from upcoding to sophisticated kickback networks (Ekin and Musal, 2021) [12]. Comprehensive auditing is almost impossible and uneconomical considering the size and complexity of the healthcare reimbursement systems. Therefore, statistical sampling methods have become an integral part of health care audits [13]. The following court case study addresses sampling methods and the subsequent decision rendered by the Illinois 5th Appellate Court in 1998 on Protestant Memorial Medical Center vs. the Department of Public Aid of the State of Illinois*.

Appellate Court of Illinois, Fifth District.
PROTESTANT MEMORIAL MEDICAL CENTER, INC., d/b/a Memorial Hospital, Petitioner-Appellee, v. The DEPARTMENT OF PUBLIC AID and Robert W. Wright, Director of Public Aid, Respondents-Appellants.
No. 5-96-0611.
Decided: March 25, 1998

The Memorial Hospital (Hospital) is a participating service provider for the Medicaid program, a medical assistance program administered and monitored by the Department of Public Aid (Department). The Department's rules and regulations require it to audit service providers and recover excess payments made under the program. 89 Ill. Adm. Code § 140.30 (Supp.1987). Pursuant to its rules, the Department conducted an audit of all outpatient services rendered to Medicaid recipients by the Hospital during the period of September 1, 1986, to February 29, 1988. Following two reaudits, the Department determined that the Hospital received $110,477.82 in overpayments, and the Department notified the Hospital that the Department planned to recoup this amount. To determine the amount of overpayment, the Department used a sample and extrapolation method, which is allowed under the Department's rules and regulations. 89 Ill. Adm. Code § 140.30(b) (Supp.1987). The Hospital disagreed with the Department's findings, and a hearing was held before an Administrative Law Judge (ALJ) on 46 separate days over a three-year period. At that hearing, the Hospital disputed the Department's determination of the amount of overpayment for specific emergency services

rendered and also challenged the validity of the Department's sample used to extrapolate the amount of overpayments. Evidence regarding the sampling and extrapolation method used by the Department in its audit was presented through the testimony of three experts.

Expert Testimony

Dr. Nosari on behalf of Department, defined the Department's "universe" as the number of recipients who received outpatient treatment by the Hospital during the audit period, i.e., 6,134 recipients. Nosari agreed that it is important to relate the universe selected to the purpose of the audit and that different aspects of a universe have different characteristics.

Nosari did not believe there is any method for testing the randomness of a sample, and he was unaware of the Wald–Wolfowitz Runs Test (WW test). Nosari explained that the variation of a universe's characteristics determines the spread of the sample from the mean. Nosari agreed that if a sample is not representative of a universe, that sample is not reliable and the results of the sample cannot be projected to the entire universe. Nosari admitted that services could be used as the unit for the universe and that using a stratified sample would extrapolate to the universe with more accuracy;

Dr. Ik–Whan Kwon, chairman of the Management and Decision Services (a division of the School of Business) at Saint Louis University, testified for the Hospital as an expert in statistics. Kwon reviewed the Department's audit data and determined that the Department should have used stratified services, i.e., grouping services into the three categories of institutional, general medical, and pharmacy, as the unit of analysis for the universe and the sample, rather than recipient. Kwon based his opinion on the fact that the audit evaluated the measurement of the payment based upon the services received and not upon the recipient. Kwon explained that the unit of analysis is important because any statistical information based on the unit of analysis will be applied to the entire universe. After the unit of analysis is selected, then the size of the sample, the sampling methodologies, the statistical tools to be employed, and the statistical techniques to be used are selected.

Kwon also explained that simple random sampling, which was used here, is an acceptable statistical method, but only when the unit of analysis is homogeneous, i.e., when each recipient received a similar amount of either services or payments. If the unit of analysis is not homogeneous, then a simple random sampling of the recipient's results in grossly exaggerated variations and the randomness creates a biased result. Additionally, Kwon stated that a computerized random sample does not guarantee randomness, as randomness is determined by whether a sample is selected in the fashion that the population is distributed.

Kwon found that the proportions of the three categories of services in the Department's sample were not proportionate to those of the universe. Kwon performed a computerized WW test on the Department's sample and population. Kwon stated that the WW test can be used to determine if a sample is randomly selected from a known population. Kwon testified that the results of the WW test revealed that the Department's sample was not truly random and therefore was not representative of the universe. Similarly, the Chi Square test Kwon performed on the Department's data revealed that the sample did not reflect the universe. Kwon concluded that if a sample is not random, efficient, and representative of the universe, or if any of these three criteria are violated, then a sample is invalid and extrapolation is useless.

Applet Court Decision

The only interpretation of what constitutes a statistically valid sample is provided by the expert testimony given in this case. Nosari determined that a simple random sample was sufficient. Won defined a valid sample as being composed of three criteria: randomness, efficiency, and representativeness. This court finds that Kwon's definition provides the most comprehensive, most just, and fairest legal interpretation of the phrase "statistically valid sample." If a sample is not representative, efficient, and random, then it would seem that any sample would meet the Department's rules, as long as the methods applied to that sample are statistically sound. It is a reasonable inference that if the basic underlying selection of a sample does not meet the three criteria set forth by

Kwon, then no matter how sound the statistical methods applied are, the result would be useless and invalid upon extrapolation. It appears that the trial court accepted Kwon's interpretation. In so doing, when the evidence at the administrative hearing is considered, the trial court's decision is not against the manifest weight of the evidence, and the Department's determination that its sample was statistically valid is against the manifest weight of the evidence.

CONCLUSION. For the foregoing reasons, the judgment of the circuit court of St. Clair County is affirmed. WELCH, P.J., and CHAPMAN, J., concur.

*Retrieved in part from https://caselaw.findlaw.com/il-court-of-appeals/1038024.html. Author of this article is a member of expert witnesses. This ruling was quoted extensively in Journal of Health Law, Spring 1999 Vol. 32, No. 2:348-349

References

1. Meier P (1986) Damned liars and expert witnesses. Journal of the American Statistical Association 81.

2. Fienberg S, Straf M (1991) Statistical evidence in the US courts: an appraisal. Journal of Royal Statistical society 154: 49-598.

3. Nathan DW (2011) Assessing the Validity of Statistical Samples in Medicare Audits. Journal of Health Care Compliance. Cited from Tatiana Danescu and Chris Anca-Oana in Procedia Economics and Finance 3: 1128-33

4. Elliott, Robert K, Rogers JR (1972) Relating statistical sampling to audit objectives. The Journal of Accountancy 46-55.

5. Messier WF Jr., Kachelmeier SJ, Jensen KL (2001) An Experimental Assessment of Recent Professional Developments in Nonstatistical Audit Sampling Guidance. AUDITING: A Journal of Practice & Theory 20: 81-96.

6. Gilbertson DL, Herron TL (2003) Audit Sampling Methods And Juror Negligence Awards: An Expectation Gap?. Journal of Applied Business Research 19: 10.19030/jabr.v19i1.2152.

7. Elder RJ, Akresh AD, Glover SM, Higgs JL, Liljegren J (2013) Audit Sampling Research: A Synthesis and Implications for Future Research. AUDITING: A Journal of Practice & Theory 32: 99-129.

8. Ponomon L, Wendell J (1995) Judgmental versus random sampling in auditing: An experimental investigation. Auditing: A Journal of Practice & Theory 14: 17-34.

9. Phillips J, Reese J, Younts J (2017) Using and Defending Against Statistical Sampling in False Claims Act Cases. American Health Association 21: 42-5.

10. Mukherjee R (2019) How to choose your sampling strategy to guarantee relevant results. The Akvo Foundation, Netherland.

11. Center for Medicare and Medicaid Services (2015). Medicare fee for service 2014 improper payments report, The Centers for Medicare & Medicaid Services, USA.

12. Ekin T, Musal RM (2021) Integrated statistical and decision models for multi-stage health care audit sampling. Journal of Applied Statistics 10.1080/02664763.2021.1900797.

13. Megerdichian A (2019) Economic analysis of optimum sample size in legal proceedings. International Journal of Statistics and Probability 8: 13-24.

Submit your next manuscript to Annex Publishers and benefit from:

> Easy online submission process
> Rapid peer review process
> Online article availability soon after acceptance for Publication
> Open access: articles available free online
> More accessibility of the articles to the readers/researchers within the field
> Better discount on subsequent article submission

Submit your manuscript at
http://www.annexpublishers.com/paper-submission.php