Pakistan Journal of Applied Sciences 1 (4): 504-507, 2001

Sampling Devotion for Efficient and Effective Representative Sample Surveys

S.M. Seeletse
Department of Statistics, University of South Africa, P.O Box 392, Pretoria, 0003, South Africa

Abstract: Sample surveys are important in finding information about a population of interest. However, the objective of sample surveys might fail to produce the intended results if not properly administered. In this article important elements that could enhance efficiency and effectiveness in survey sampling are exposed. The intention is to discuss selection of a sample that utterly represents the targeted population. The discussion does not envelop methods to find the information from the sample; it goes only up to the selection of a sample that has a great probability of resembling the targeted population. The idea is to select a representative sample so that the information contained in it can be adequate to answer any question about the population.

Keywords: Representative Sample, Efficient and Effectiveness in Survey Sampling

Introduction

Researches in business and marketing, economic, health and social sciences (among others) depend on sample surveys and other techniques that can be performed well by statisticians. Cohen (1988) supports the above assertion by pointing that postgraduate research projects (qualitative and quantitative) use sample surveys frequently. Statistical consultants normally assist these projects. Modern developments show that a mix of qualitative and quantitative (enumerative) study has potential to add value to the research results. Hence, this discussion does not confine itself to a single type (qualitative or quantitative), any of the types or a combination of them is also covered.

Sarndal *et al.* (1992) describes a sample survey as a process that consists of planning (as starting point), data collection, analysis of survey results to find appropriate statistical procedures to analyse the collected data, deriving the findings and presenting results. The quality of results depends on the appropriateness and correctness of each of these steps. Further, any step that might be done erroneously have potential to reduce this quality.

Lindgren (1999) emphasizes that the quality of data is an important constituent in deriving worthy survey results. While statisticians can improve the procedures used in data collection, it is not possible to control the quality of the already collected data. Rousseeuw & Leroy (1987) enlighten that if the collected data shows to be "dirty" (i.e. contaminated by undesirable elements or lacking quality), the only appropriate data analysis approach to ensure reliable and valid research results is to use robust methods, which are methods that resist influence of undesirable elements. Robust methods and their use is a specialized skill, where not every statistician is equipped to perform. Therefore, the option to permit dirty elements in the data with the view to use robust methods is not recommended in general.

Experience shows that data collected without using a well-informed person such as a statistician is a worthless exercise. It can lead to use of inappropriate data analysis methods. Following a process that is confined to appropriate planning can enhance collection of good quality data. Appropriate planning leads to employing a beneficial sampling framework and consequent selection of a good sampling method.

Framework for sampling in sample surveys: A framework is a guidepost that includes both the ingredients and the methods or processes used in an application. In this case, a sampling framework is an inclusive set of the entities that starts from the population or universe targeted for selecting a sample, the study objectives and the sampling method. To ensure that a survey has all that it requires, there needs to be a clear and known sampling frame, which is defined by Patel & Lumetta (2001) as a list of all the units in the population that is under investigation.

There are cases where the definition of the target population is not clear, such as situations where the research objectives have not been clearly defined. However, according to Hajek (1981) the guide is that the beginning should be specification of what needs to be accomplished (objectives), how it should be done (method) and by whom (skills identification). This emphasises that objectives should be clearly defined; skills required should be identified and then the target population can be specified. Practical experience shows that data analysis becomes more charming when the objectives are known and the manner in which data was obtained is clear to the data analyst.

Data analysis can also be difficult or impossible. For example, when the researcher is not able to provide the necessary information, there is a possibility that the analysis could break down. Thompson (1992) states that a statistician can caution the researcher that statistical inference is not possible when information is inadequate regarding the objectives and the data collection. Even though some researchers could interpret such a response as sarcastic, Thompson believes that it is an appropriate reaction.

Greenfield (1998) points out that it is impossible to design a helpful questionnaire when the given information is insufficient for the task. Further, an advice by Pace (2000) is that the fundamental aspects for a proper statistical analysis start with clearly defined objectives and a clear definition of the target population. In support of this, Cochran (1977) puts forward the contention that if the target population is not clearly defined, it will also not be easy to ensure random sampling because each random sampling is a probability sampling method based on a specific format in the data. Richards *et al.* (2000) portray the most uncomplicated and effective sampling frame as the one consisting of all the entities starting from the target population, not including any replicates and is recent. It should also not contain any elements outside of the target population.

An impression made by Dutka & Frankel (1990) is that a perfect or ideal population usually does not exist in practice, but only in theory. These authors point out that if good sampling frames can be found, they are very expensive for most researchers, especially the student researchers. Regardless of these obstacles, there should always be a sampling framework to enhance collection of good quality data. The fact that good sampling frames

are difficult or sometimes impossible to find should not be used as a reason for not following the appropriate guidelines. A sampling frame should include a representative section of the target population. If a sampling frame misses a fundamental aspect, it could misrepresent the reality of the entire population. The next example shows a distortion by sampling frames.

Example of distortion in sampling frames: A market researcher wanted to investigate the market potential for a new fruit juice among households in the Gauteng Province of South Africa. Due to time and budget shortages he decided to limit the extent of the survey. His surveys were confined to Soweto township. A sample was drawn from Soweto, and the analysis was done. It revealed an overwhelming demand. In one summer month in 1997, a hefty scale preparation was made for the fruit juices, and huge sales were expected from seven townships that are in the radius of 25 km from the borders of Soweto. To the surprise of management, only two townships (including Soweto) had vast sales. In one township the sales were just below half of the stock for that area. In four of the townships the sales were far below 25% of the stock for these areas. On aggregate, the sales made for the entire effort could not reach 30%, it was about 26.76%. The result was that over 70% of the stock was to be repacked and sold at lower quality and less price. It was also scary because the repacked stock had a shorter duration than the original ones. The entire setback was attributed to the use of an inadequate sampling frame. It was found out that the sampling frame was simply not representative of the target market.

Magnussen (2000) cautions that a sampling frame that does not cover all the features of the target population leads to results that cannot be generalized to the entire target population. The sampling frame so used is called a partial frame, and the resulting problem is referred to as "non-coverage", or "non-representativity".

Another serious problem could occur due to a complete absence of a sample frame. This problem forces the researcher to use a sampling method with a planning that is not based on any target population. This implies that planning does not depend on any reliable information, except for guesswork. It normally compels the researcher to use a non-probability sampling method simply for convenience of obtaining data, and resulting in paying the price of non-generalizability.

Researchers are advised to use inventiveness, plus any available resource to unearth an optimal sampling frame for the survey that is being planned. Prominent researchers emphasise that the choice of a sampling frame depends significantly on the originality of the researcher. This quality can enable the researcher to realise that some random sampling procedures do not require an entire inventory of the target population. The statement of objectives can assist in deciding the extent of population requirements in sampling.

It is recommended that researchers should take care in defining the research objectives, and in the specification of the population. S/he should be careful that if a partial frame is used, the consequence could be dangerous due to lack of representativity and omission of valuable information. Sample results can only be generalized to the units in the frame from which the sample is selected. There is no validity in research results based on partial sampling frame and then supplemented by additional information to legitimize it. In general, a possible price that can be paid by limiting the extent of a survey could include lack of reliability and of validity.

Probability vs non-probability samples: According to Patel & Lumetta (2001), collection of quality data depends typically on an optimal sampling frame. Further, skillful use of a sampling frame in selecting the sample is equally vital, and is the next important step after identifying the frame. In the selection process, human bias should not be allowed to play a role. Use of random (probability) methods in sampling addresses the problem of bias to a considerable extent. Probability samples require that every unit in the sampling frame should have a positive chance of being selected/included in the sample. There is also no general requirement of equal probability for being selected. The basic requirement for probability sampling is a positive chance/probability for all population units. Sample surveys are based on these types of sampling.

In fact, for fairness in data analysis and the deriving of results, a desirable feature in probability samples is that sampling methods be unbiased, i.e. should not favour certain units at the expense of others. The problem is that in general, the nature of sampling frames and population cannot guarantee total fairness. Nonetheless, there are numerous approaches to ensuring restricted (and sometimes no) biases in the selection process. Two common, accessible approaches are the use of a table of random numbers and a computer algorithm that generates random samples from the sample frame.

In order to generalise from the sample to the population, random sampling is necessary. In actual fact, it is of crucial importance. Benefits of random sampling include the fact that estimation of population characteristics can be done, and the relevant confidence intervals evaluated. On the other hand, non-probability sampling restricts the scope of statistical analysis because there cannot be any justifiable generalizations, unless there is a further investigation, or additional information is provided. The use of non-probability sampling is in fact unfortunate if there is a need to generalise. This is because in general, non-probability methods place severe and often unnecessary limitations on the potential impacts of a survey.

The researcher is advised to take extra care of the selection process to ensure that the sampling frame has been used as intended. There are cases where a convenience sample is selected by the person tasked to select a sample, and giving the researcher the impression that a probability sample has been selected. Results that could be made without the researcher knowing that it was not the probability sampling used, may mislead users later to the extent that an impression is made that "it is easy to lie with statistics". Inference is not appropriate when non-probability methods were not used in sampling.

A probability sample can only be derived from a probability sampling method, and the researcher needs to be certain that when it is concluded that a probability sample was used in the selection, that is indeed the case. Even if a slight deviation was made in the selection, such deviation should be reported to the data analyst.

Probability samples and representativity: Probability sampling methods do not necessarily guarantee that all the selected samples using the methods are wholly representative of the entire population. The patterns in the data enable appropriate choices of random sampling methods to enhance representativity. Common random sampling methods are simple random sampling, cluster sampling, stratified sampling and systematic sampling, each one with its own conditions and guidelines. They are discussed in the next paragraphs.

Simple random sampling: Bless & Kathuria (1993) explicate that simple random sampling is based on that all the units in the population have an equal chance of being included in the sample. However, even though simple random sampling is popular, it is not the only

probability sampling method used. Further, where data elements are not homogeneous with respect to the characteristic being investigated, simple random sampling is simply not an appropriate method to use in sampling for representativity.

Therefore, there could be certain compositions in some target populations that will forbid use of simple random sampling. More probability sampling methods are given in the next discussions.

Stratified random sampling: Populations could consist of various subgroups such as nationality, age, gender and other characteristics that are being studied. That is, the population may contain subgroups that differ in terms of the subject being studied, and each of these subgroups will need to be represented in the sample. It is in fact of crucial importance that each of the subgroups is represented in the sample. This can be accomplished by dividing the population into these subgroups before conducting the survey. Henry (1990) notifies that these subgroups are called strata (stratum for singular) and the sample design is referred to as stratified random sampling. It should be noted that as with other probability sampling methods, stratified sampling constrains the sample to be representative of the various strata of the population. Construction of strata becomes possible or straightforward when there is additional information to make the researcher and the data handler aware. For example, a sampling frame that provides information about gender, race, nationality, educational qualification, geographic location, job category, or even age, has information that justify stratified sampling.

Systematic random sampling: Suppose there are N elements in the population of interest, and a sample of size n is required. To collect a systematic random sample, Stoker (1988) advises that the sample requires that a unique number be assigned to each of the N sample units, (say 1, 2, ..., N) in a 1-in-k systematic sample (where k = N/n). Here, either k is a natural number or select the closest rounded natural number K, and pretend that the numbers are in a circle. Select using a probability sample a number between 1 and K, say the i^{th}. This means that the i^{th} unit has been chosen. To select the complete sample, starting from the i^{th} unit, select every K^{th} unit until n elements have been chosen. The sample is i^{th}, $i + K^{th}$, $i + 2K^{th}$, . . . , $i + (n - 1)K^{th}$.

Cluster sampling: According to Hague & Harries (1993), cluster sampling is useful in drawing a representative sample from a population where there is no specific knowledge about a population, or a sampling frame does not exist and there are shortages of time and cost. Cluster sampling procedure requires dividing the population into natural subgroups called clusters, where the elements in each cluster are geographically close to one another. Few clusters are then chosen using a probability sampling method, and the elements from the chosen clusters form the sample called cluster sample.

The cluster sample is representative if the internal composition of each cluster is diverse with respect to the population characteristic of interest. The variance within a cluster should therefore be large but the variance between clusters should be small. Each cluster should thus represent the population.

Multi-stage random samples: Lehtonen & Pahkinen (1994) present multi-stage sampling as the combining of sampling methods one after the other in selecting a sample. Such a combination may be taken if its use suits the requirements of a specific application best. A three-stage random sample example is as follows: Divide the target population into clusters, select a random sample of clusters (called primary sampling units, which is the first stage). For the second stage, use stratification to divide the selected clusters into strata, and select a random sample (called secondary sampling units) from each stratum. Proceed to the final stage by using simple random sampling on the strata to select the final or ultimate sampling units.

There are possibilities that the methods presented could lack applicability due to realities of the situation under which sampling is to occur. This calls for a skilled and experienced investigator to understand possible modifications to the approach without distorting the meaning and intentions of the research. The next discussion presents a case that shows how a skillful practitioner can do to find a representative sample. A discussion follows the example to point at the imperfections that could have occurred if an incorrect sampling procedure had been used.

Example: a restructuring survey: Case study:

This case study explains the set of circumstances that took place at a business company in South Africa. The company details, and details of the endeavour, including the choice of a sample, are given in the following commentary. A discussion is given at the end of the case study to present a critique.

Background: This discussion reports the multipurpose terminal (MPT) of Valqual, a company that transports cargo and people, which also provides services in environmental issues and dealing in scrap exchange & other services. Valqual is situated in the Brits area. Brits is a small town in the North-West Province of South Africa. The sampling objective was to undertake a survey, and select a sample that would assist in finding information about the perceived changes to address affirmative action and other new legal requirements.

In January 1999 Valqual was been investigated for a number of compliances with new legislation, including equity in the senior ranks of employment. The objective was to evaluate changes that took place since 1996 when the directive for restructuring was received by the firm. In addition to the objective of finding perceptions about the extent of restructuring for Valqual, the surveys used were expected to investigate developments (or changes) in employee training approach as well.

It was decided that the sampling frame be composed of MPT employees who were already in the employ of MPT since 1994. At the beginning the MPT manager proposed that simple random sampling be used, and the sampling frame as the entire MPT workforce at the time. He was, however, advised that the workforce is not homogeneous, and that not all the employees had the required information. New employees lacked information, and were more likely to lie. Further, the different departments differed in degrees of change.

During planning it was noted that some trade unions in the firm did not serve their constituencies as loyal as expected, and this led to division in thoughts between union executives and the constituencies. Two groups emerged from this. Other strata that were formed came from the three levels of management; junior, middle and senior (3 strata), and the 6^{th} stratum was made up of the general (so-called unskilled) workers. The casual, part-time and temporary staffs (CPT) were put in the 7^{th} stratum in case they had different impressions from the groups with which they identified or worked during deployment.

Therefore, the sampling frame consisted of 7 strata. Stratum 1 consisted of trade union office bearers with 13 members. Stratum 2 of the dissatisfied trade union members had 103, and strata 3 to 5 of junior, middle and senior management with 37, 29 and 14 members. The 6^{th} stratum consisted of CPT with 127 members. Finally, the 7^{th} stratum of general workers had 1024 members. This made the 1347 units in the sampling

frame.
The next paragraph explains how the whole undertaking was done. It demonstrates that skill and knowledge, coupled with experience and insight, are often vital in undertaking a job.
Method: It was decided that the sample size be 10% of the sampling frame. Thus, about 135 respondents were required. For each stratum, 10% of units were required, which meant 1, 12, 4, 3, 1, 13 and 102 respondents from the respective stratum. This implied that 136 units were required, which was considered fine.
A computer was used to split strata, and names in strata 1 to 5 and 7 were ordered alphabetically. For strata 1 to 5, simple random sampling was used with random numbers used to target or identify respondents.
In stratum 6, the CPT consisted of diverse and relatively unknown personnel. Thus, cluster sampling was used to select the 13 respondents. The stratum was subdivided into residents of surrounding residential areas, and eight such geographical areas were identified. Eight clusters were formed. The 127 CPT members were placed according to these eight clusters, and four clusters were selected by simple random sampling. The clusters did not differ much in sizes. From the first three clusters, 3 units were selected using random numbers. In the last cluster 4 units were selected using random numbers. These contributed the required 13 respondents from CPT.
From the last stratum, the last 102 from the general workers were selected using systematic sampling. For this, $N_6 = 1024$, $n_6 = 102$, $k = N_6/n_6 = 10.04$ so that $K = 10$. A 1-in-10 systematic sample was selected after finding the 1st element between 1 and 102 from random numbers. Thus strata 1 to 5 used simple random sampling, stratum 6 used cluster sampling and stratum 7 used systematic sampling.
The sampling procedure used was a complex multistage probability sampling method.

Discussion

At the beginning the MPT manager had proposed simple random sampling, and the sampling frame as the entire MPT workforce at the time. However, he was advised against it because the workforce is not homogeneous, and not all the employees have the required information. The new employees were probably lacking information, and instead of fact they were more likely to give perceptions. Further, the differences at different departments in terms of changes were not going to be captured enough. Simple random sampling was also more likely to over-represent some strata and under-represent others.
None of the sampling methods was appropriate on its own, and the complex approach followed ensured equitable representation (at least in theory) of all the strata.

Conclusion and Recommendations

Research is important for gathering information. However, gathering of inadequate information is a serious upset for pursuing any task. It costs effort, money, time, and probably distortion of truth, among others. This study discourages sampling that does not lead to accomplishment of the desired objective, and advocates following the procedures that add value to research efforts. Use of knowledgeable persons is advisable. It typically leads to more benefits than was originally anticipated by the researcher. On average, these benefits do not surface when non-experts are involved, instead there is potential to damage the entire research initiative. Sometimes use of non-expert gets to the extent of destroying the benefits that were originally anticipated by the researcher.
This discussion recommends use of sampling experts in sample surveys. Experienced experts would also advise the researcher on the definition of objectives and sampling frame, they know sampling (and are acquainted particularly with the distinction of probability and non-probability sampling). When probability sampling is required and the researcher is not a statistician, the discussion recommends use of a statistician. If s/he is a statistician, s/he should still consult using available networks to get her/his methods endorsed by a second expert. In the end, a statistician should explain the details of the survey sampling efforts to the researcher, and details about the choice of sampling methods. Expert conduct of the expert advisor can comfort the researcher to be assured that a representative sample has been selected. Efficiency and effectiveness depend on use of the sampling resources, including experts with appropriate knowledge.
The case study in the article demonstrated that other manipulations that could be necessary should be done. Without insight understanding the researcher might fail to realise this, or could follow an incorrect approach if advised but not assisted fully.

References

Bless, C. & R. Kathuria. 1993. Fundamentals of social statistics- an African perspective. Johannesburg: Juta & Co.

Cochran, W.G., 1977. Sampling techniques. New York: Wiley.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dutka, S. & L. Frankel. 1990. Misuses and abuses of statistical techniques in market research surveys. Chance, 3: 18-23.

Greenfield, T.K., 1998. Evaluating competing models of alcohol-related harm. Alcoholism: Clinical and Experimental Research, 22: 52-62.

Hague, P. & P. Harries. 1993. Sampling and statistics. London: Kogan Page.

Hajek, J., 1981. Sampling from a finite population. New York: Marcel Dekker.

Henry, T.G., 1990. Practical sampling. Sage: California.

Lehtonen, R. & E.J. Pahkinen. 1994. Practical methods for the design and analysis of complex surveys. London: Wiley.

Lindgren, O., 1999. Quality control of measurements made on fixed-area sample plots. (In: Integrated tools for natural resources inventories in the 21st century. Boise: Idaho USDA Forest Service.

Magnussen, S., 2000. Unequal probability sampling in fixed area plots of stem volume with and without prior inclusion probabilities. J. Applied Statistics, 27: 975-991.

Pace, E., 2000. Leslie Kish, 90; improved science of surveys. New York Times, 150(51541): A17.

Patel, S.J. & S.S. Lumetta. 2001. Replay: a hardware framework for dynamic optimization. IEEE Transactions on Computers, 50: 590-608.

Richards, T., J. Gallego & F. Achard, 2000. Sampling for forest cover change assessment at the pan-tropical scale. Int. J. Remote Sensing, 21: 1473-1491.

Rousseeuw, P.V. & A.M. Leroy. 1987. Robust regression and outlier detection. New York: Wiley.

Sarndal, C-E., B. Swensson & J. Wretman. 1992. Model assisted survey sampling. New York: Springer.

Stoker, D.J. 1988. The analysis of complex sample data. Pretoria: HSRC Publication, WS-40.

Thompson, S.K., 1992. Sampling. New York: Wiley.