International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-4, Issue-9, September 2018
ISSN: 2395-3470
www.ijseas.com

Sampling Methods and Sample Size

Dr.S.Narayanan[1], Dr.C.Sabarigirinathan[2], Dr.K.Vinayagavel[3], Dr.D.Deepiha[4] , Dr.S.BalaSiddharth[5] , Dr.M.Saravanapriya[6]

1(PhD student of Pacific University of Higher Education and Research(PAHER), Udaipur, Rajasthan)

2(Professor and HOD, Dept of Prosthodontics, Tamilnadu Govt Dental College & Hospital, Chennai, India)

3(Professor, Dept of Prosthodontics, Tamilnadu Govt Dental College & Hospital, Chennai, India)

4&6(Post Graduate Student, Dept of Prosthodontics, Tamilnadu Govt Dental College & Hospital, Chennai, India)

5(MBBS – Private, Chennai, India)

Abstract

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed but may include simple random sampling or systematic sampling. This article details about the Sampling methods and Sample size.

Introduction

Most research studies involve the observation of a sample from some predefined population of interest. In epidemiological studies, for example, a sample of people is observed for exposure to various risk factors, health outcomes and other related variables. The conclusions drawn from the study are often based on generalizing the results observed in the sample to the entire population from which the sample was drawn. Therefore, the accuracy of the conclusions will depend on how well the samples have been collected, and especially on how representative the sample is of the population. In this chapter, we will discuss the major issues that a researcher has to face in selecting an appropriate sample.

Why sampling?

Sampling is a process of choosing a section of the population for observation and study. There are several reasons why samples are chosen for study, rather than the entire population. First and foremost, a researcher wants to minimize the costs (financial and otherwise) of collecting the information, processing this information and reporting on the results. If a reasonable picture of a population can be obtained by observing only a section of it, the researcher economizes by choosing such a section of the population. Obviously, when a sample is observed, the total information will be less than if one were to observe the entire population.

However, in some cases, the process of observing the entire population would take such a large amount of time and resources that (a) the results would not be timely, and (b) the observations might be less reliable. Consider the common approach to observing the entire population, the census. Most countries collect information on their population periodically (every five years, every ten years, etc.) through census. This involves enumerating every individual in the population and collecting a predetermined set of information. Even in a relatively small country such as Canada (population, 29 million), the process takes a substantial part of a year, and the tabulated observations are not available for several years after the census. When the population size is large, for example in India or China, the data analysis and reporting may be delayed even further. In addition, the census is never able to collect information on all the population: the homeless and nomadic sections of the population are often missed.

A major advantage of sampling over complete enumeration is the fact that the available resources can be better spent in refining the measuring instruments and methods so that the information collected is accurate (valid and reliable). Some information, such as monitoring of the body burden of toxic metals in the population, which may require specialized equipment and staff, cannot be collected from the entire population. A sample in such cases would provide a reasonable picture of the population status.

Process of sampling

What determines a proper sample? The primary concern in selecting an appropriate sample is that the sample should be representative of the population. Every variable of interest should have the same distribution in the sample as in the population from which the sample is chosen. This requires knowledge of the variables and their distribution in the population, which of course is why we are doing the study in the first place! Therefore, it is not often possible to ensure the representativeness of the population. However, statisticians have come up with ways in which we can give a reasonable guarantee of representativeness. We will discuss some of these methods briefly in later sections.

Before a sample is drawn, the population has to be clearly defined. In a population survey, this requires having a list (sampling frame) of all the individuals in the population. Probabilistic methods can then be developed to draw a sample in such a way that we can assure the representativeness of the various characteristics in which we are interested. In experiments (such as clinical trials) this list may not be explicit, and may evolve as the sampling progresses. For example, a list of inclusion and exclusion criteria would be specified at the beginning of the trial, defining the general framework for the population. Then, as patients are identified, they will be selected for study, and allocated to various experimental groups using probabilistic methods.

The sampling frame consists of a list of elements (units) of the population. In population surveys, this is a list of people. In clinical trials for a disease, it is a list of patients with that disease. In a casecontrol study, it is a list of people with the disease and a list of people without the disease. The completeness and accuracy of this list is essential for the study to be successful. One of the major flaws in many research projects is a biased selection of the sampling frame. For example, if a telephone survey is conducted in India before a general election to predict which party will win, the results will most likely be wrong, since the sampling frame consists of only affluent people (who can afford a telephone), and their opinions are not likely to be representative of the entire population.

Once a sampling frame has been identified, one needs to have methods of selecting individuals from this frame to be included in the study. Two issues are important: how large a sample should be selected, and how the individual units should be selected. These issues are discussed in the following sections.

How large a sample?

One of the most difficult decisions facing the researcher is how large his sample should be. Two common approaches are employed in research studies: the empirical and the analytical. The empirical approach involves using sample sizes that have been used in similar studies. This has no scientific basis, and will only be satisfactory if the previous studies had acceptable limits on the errors of generalization, and the current study is very similar in its scope (objectives, design, study population, etc.). This method is not recommended and will not be discussed further.

The analytical (scientific) approach to determining the appropriate size of the sample to be included in the study depends on the assessment of errors of inference, and a desire to minimize 'sampling error'. Sampling error measures the amount of variability between sample results (as a proxy for closeness to the real situation in the population, and as reproduced in the sample results); the less variable the sample results are, the closer the sample results are to the population results.

The main determinant of the sample size is, therefore, how accurate the results need to be. This depends on the purpose of the study (descriptive study to determine a summary measure of a characteristic, or an analytical study where specific sets of hypotheses are being tested).

Sample sizes for descriptive studies

In the case of descriptive studies, often the object is to obtain an estimate of a population parameter. For example, in opinion polls, the market researcher may be interested in finding out what proportion of people prefer a particular brand. A nutritionist may be interested in the average daily caloric intake of the population. A health researcher may be interested in the proportion of people who smoke, or the median survival after coronary bypass surgery. The determination of the size of sample required to answer these questions depends on several factors:

i. What is the measure of interest? This would have been determined by the study objectives. The identification of the characteristic of primary importance determines the next steps in the process of defining the sample size. For example, if a prevalence rate in the population is to be estimated by observing a sample from the population, the measure is the proportion of people in the sample with the disease.

ii. What is the underlying probability distribution of the characteristic of interest? Most research questions fall into one of two possible scenarios: the binomial distribution (when one wants to estimate the proportion of a certain event), and the normal distribution (when one wants to estimate an average value). The market researcher above, for example, has the preference of a brand as the characteristic, with two possible outcomes. If one assumes that there is possibly a fixed proportion (π) of people with preference for the brand, then the number of people expressing this preference in any fixed set of people will follow a binomial distribution, with the proportion (p) of the people showing the preference as a good estimate of the population proportion. For the nutritionist, the daily caloric intake of individuals follows a normal distribution with some average (μ), and the average of the daily caloric intake of the sample of people (x) observed would be a good estimate of this population value.

iii. What is the sampling distribution of the measure? Drawing inferences from the sample to the population involves inherent errors, which are measured by the sampling distribution. If we observed several samples, under the same method of selecting the samples, the measures from each of these samples would vary, resulting in a 'probability distribution' for the sample measure. This distribution is called the sampling distribution, and it depends on the type of study design and on how the samples were obtained. In calculating sample sizes, it is often assumed that the sampling involves simple random sampling (discussed later in this chapter). Sometimes the sampling design is much more complicated (e.g. multistage cluster sampling techniques) and more complicated formulae will have to be used to calculate sample sizes appropriately.

iv. How accurate do you want the results to be? Basically, one is interested in obtaining an estimate as close to the population value as possible. Therefore, some measure of the difference between the estimate and the population value has to be considered. In most cases, a mean-squared error (average of the squared deviation of the sample value from the population value) is used. A concise way of expressing this error is to use the 'standard error of the estimate'. The standard error comes from the sampling distribution of the estimate. If the sampling is done properly (with appropriate probabilistic methods), one can predict what this distribution should be, and based on this, one can estimate how close to the population value the sample estimate will be:

For example, in the case of estimating the population proportion, the sampling distribution of the sample proportion, p is approximately normal, with mean π and variance π(1-π) /n, where n is the sample size. This gives the (1-α) confidence interval for π to be

$$p \pm z1-\alpha\sqrt{p(1-p)/n}$$

where $z_{1-\alpha}$ is the appropriate cut-off point on the standard normal distribution. (For example, for 95% confidence, $z_{1-\alpha} = 1.96$.) The accuracy of the estimate therefore depends on two quantities: how narrow this interval is (width of the interval) and how confident we are (e.g. 95%).

The calculation of the size of the sample for a descriptive study therefore depends on the two parameters – the width of the confidence interval and the confidence coefficient. Computer programs are readily available (e.g. EPIINFO has a module that allows for the computation of sample sizes). The two common scenarios, estimating a population proportion and estimating a population mean, are illustrated below:

a. Estimating a population proportion (p). Suppose we want to conduct a survey to determine the prevalence (π) of a relatively common disease in a community. We want to determine how many people should be observed to obtain a reasonably accurate picture of the prevalence. The following steps are necessary: · Specify the parameters of error:

Confidence coefficient (1-α)	95%
Width of the interval (δ)	10%
· Make a guess as to the value of π	30%

The problem is to calculate the sample size required for estimating the prevalence of the disease within ± 5% of the true value, with 95% confidence. Since the confidence interval actually depends on the true value, p, we have to make a guess as to what this value might be. This is done based on prior experience; if no guess is available, use the value 50%, which will give the largest sample size. Using the fact that the sample proportion (p) has the confidence interval given above, the sample size (n) can be calculated using the formula:

$$n = (z_{1-a}/d)^2 p(1-p)$$

In the above example, therefore, n = $(1.96/5)^2(30*70)$ = 323; we need a minimum of 323 subjects observed to assure that the 95% confidence interval for the estimated proportion will be within 5% of the true prevalence. If the true prevalence is less than 30%, the confidence interval will be narrower. The maximum sample size required will occur when the true prevalence is 50%, in which case, n = 385.

The above calculation assumes a simple random sample from a relatively large population. In practice, the population from which the samples are drawn may be fixed and small, in which case corrections to the above formulae are required. (See EPIINFO program for variations of this formula, and use under different sampling designs.)

b. Estimating a population average (μ). Suppose we want to estimate the average daily caloric intake of people in a community. The daily caloric intake is assumed to have a normal distribution around μ, with a standard deviation (σ). The sample measure used to estimate μ is the sample mean. The sampling distribution of the sample mean is also normal, with the same mean, μ and standard deviation, σ/√n (the standard error of the mean). Notice that we need to know the value of σ to proceed further. It is either obtained from other similar studies, or by actually obtaining a small number of observations at random in a test study. If neither of these is possible, one may make a reasonable guess by taking the maximum range (maximum value possible – minimum value possible) and dividing this range by 4. (Using the supposition that for normal distribution, 95% of values will be within ± 2 standard deviation from the mean, and the mean will be the central value.) Then the following steps will help calculate the sample size:

· Specify error parameters:

Confidence coefficient (1-α):	95%
Width of the interval (δ):	50 cal.

· Obtain the standard deviation (σ): 150 cal.

· The 95% confidence interval for the sample mean is:

$$\bar{x} \pm z(1-a)s/\sqrt{n}$$

· Therefore the required sample size in the example is:

n = (1.96*150/50)^2 = 35.

c. Estimating relative risks or odds ratios. The formulae for calculating sample sizes in these situations are much more complicated, since the sampling distribution of the estimates of relative risks and odds ratios are not simple. Various computer programs are available to calculate the appropriate sample sizes.

The principles are essentially the same: determine the formula for confidence interval, and by specifying the two parameters, calculate the sample size from this formula.

Sample sizes for analytical studies

Since the primary purpose of an analytical study is to test (one or more) null hypotheses, the determination of the sample sizes requires the specification of the limits of errors one is willing to accept in accepting or rejecting the null hypothesis (type I and type II errors). As in the case of descriptive studies, one has to determine the sample measures used (a proportion, a sample mean, an estimate of RR or OR, etc.) and their sampling distribution (on the basis of which, a decision to accept or reject null hypothesis is taken). By equating the two types of errors based on the sampling distribution to the pre-set limits on these errors, we can work out the sample size.

For example, suppose we decide to accept a type I error, or α (probability of making a false conclusion that the two proportions are not equal in the population, when they are in fact equal). The calculation of a type II error, or β (probability of making a false decision that the two proportions are equal when they are not) depends on a precise definition of 'null hypothesis is not true'. The simplest way to do this is to define the smallest difference (δ) in the two proportions that we consider meaningful (clinically significant difference) and calculate β under this hypothesis. Clearly, if the difference is larger than δ, the probability of type II error will be less. Using this approach, formulae have been derived for calculating sample sizes for various types of statistical tests. [Note: In statistical tests, the discussion of type II errors may be worded in terms of 'statistical power', which is simply 1-β: i.e. having a 5% type II error is the same as the study having 95% 'power'.] The more common of these situations are summarized below. (As before, computer programs are readily available for most of these cases, and the computation here is presented solely for illustrative purposes.)

a. Testing equality of two proportions: $p_1 = p_2$.

The sample measures used are the sample proportions, and the sampling distribution used in testing this null hypothesis is either the standard normal distribution (z), or equivalently the chi-square (χ^2).

· Set type I error: α;

· Determine 'minimum clinically significant difference': δ;

· Make a guess as to the 'proportion' in one group (usually 'control'): π_1 ;

· Determine the power required to detect this difference: (1-β).

The sample size required is:

$$n=[\{z_{1-a}\sqrt{2\bar{p}(1-\bar{p})}-z_b\sqrt{p_1(1-p_1)+p_2(1-p_2)}\}/d]^2 \text{ where } \bar{p}=(p_1+p_2)/2$$

For example, suppose we are interested in determining the sample size required in a clinical trial of a new drug that is expected to improve survival. Suppose the traditional survival rate is 40%, i.e. $\pi_1 = 0.4$. We are interested in detecting whether the new drug improves survival by at least 10%, i.e. $\delta = 0.10$, therefore $\pi_2 = 0.50$. Suppose we want a type I error of 5%, i.e.$\alpha = 0.05$, therefore $z_{1-\alpha} = 1.96$; we also want the type II error (β) to be 5%, or we want to detect a difference of 10% or more with a probability of 95%: therefore $z_\beta = -1.645$.

Substituting these values in the above equation gives n = 640. Thus the study would require 640 subjects in each of the two groups to assure a probability of detecting an increase in the survival rate of 10% or more with 95% certainty, if the statistical test used 5% as the level of significance.

b. Sample size for a case-control study

Suppose that long-term use of oral contraceptives (OC) increased the risk for coronary heart disease (CHD) and that one wished to detect an increase in relative risk of at least 30% (equivalently, OR>1.3) by means of a case-control study. What would be the proper sample size?

The test of hypothesis in the study will be equivalent to testing if the proportion of women using OC is the same among those with CHD and those without CHD. We need to determine what proportion of women without CHD (controls) use OC; let us say 20%. Then we decide what will be the minimum difference that should be detected by the statistical test. Since we need to detect an OR>1.3, this translates to an increased use (24.5%) among the CHD patients, to give a difference of 4.5% to be detected. Choosing α and β to be 5% each, the sample size, using the above formula, would be 2220, i.e. we need to study 2220 cases and 2220 controls for the disease.

Sometimes the ratio of cases and controls may not be one-one, e.g. when the disease is rare, the number of cases available for study may be limited, and we may have to increase the number of controls (1-2, 1-3 etc.) to compensate. In such cases, the calculation of the sample size will incorporate these differences. Computer programs such as EPIINFO allow for these variations.

c. Comparison of two population means

When the study involves comparing the means of two samples, the sample measure that is used is the difference of the sample means. This has an approximately normal distribution. The standard error of difference depends on the standard deviations of the measurements in each of the population, and depending on whether these are the same or different, different formulae have to be used. In the simplest (and most commonly used) scenario, the two standard deviations are considered to be the same. We will illustrate the procedure.

We need to determine, as in case a, the minimum difference (δ) in the means that we are interested in detecting by statistical test: the two types of statistical errors (α and β)

and the standard deviation (σ). Then the sample size required is calculated using the following formula:

$n = [(z_{1-a} - z_b)s/d]^2$

For example, suppose we want to test a drug that reduces blood pressure. We want to say the drug is effective if the reduction in blood pressure is 5 mm Hg or more, compared with the 'placebo'. Suppose we know that systolic blood pressure in a population is distributed normally, with a standard deviation of 8 mm Hg. If we choose $\alpha = 0.05$ and $\beta = 0.05$, the sample size required in this study will be: $n = [(1.96+1.645)8/3]^2 = 34$ subjects in each group.

If the design is such that the two groups are not independent (e.g. matched studies or paired experiments) or if the standard deviations are different for the two groups, the formulae should be adjusted accordingly.

d. Comparison of more than two groups and multivariate methods

When considering sample size calculations for studies involving comparison of more than two groups, either comparing proportions or means, several other issues (e.g. which comparison is more important than the others: whether errors of paired comparison, or for the study as a whole are more important, etc.) have to be taken into account. Accordingly, the formulae for each of these situations will be much more complicated.

In multivariate analyses, such as those using multiple linear regression, logistic regression, or comparison of survival curves, simple formulae for the calculation of sample sizes are not available. Some attempts at estimating sample sizes using nomograms, or by simulating experiments and calculating sample sizes based on these simulated experiments, have recently appeared in the statistical literature. We will not discuss these here. When planning experiments, one of the crucial steps is in deciding how large the study should be, and appropriate guidance should be sought from experts.

Sampling methods

Once the population has been identified and the size of the sample determined, we need to decide how we are going to choose the sample from the population. [The size of the sample will also depend on this choice and therefore, the issue of sample size may have to be revisited after the choice of the sampling method; most of the discussions in the earlier section on sample size assumed a simple random sample.]

a. Simple random sample

This is the most common and the simplest of the sampling methods. In this method, the subjects are chosen from the population with equal probability of selection. One may use a random number table, or use techniques such as putting the names of the people into a hat and selecting the appropriate number of names blindly. Recently, computer programs have been developed to draw simple random samples from a given population. The simple random sample has the advantages that it is easy to administer, is representative of the population in the long run, and the analysis of data using such a sampling scheme is straightforward. The disadvantage is that the selected sample may not be truly representative of the population, especially if the sample size is small.

b. Stratified sampling

When the size of the sample is small and we have some information about the distribution of a particular variable (e.g. gender: 50% male/50% female), it may be advantageous to select simple random samples from within each of the subgroups defined by that variable. By choosing half the sample from males and half from females, we assure that the sample is representative of the population with respect to gender. When confounding is an important issue (such as in case-control studies), stratified sampling will reduce potential confounding by selecting homogeneous subgroups.

c. Cluster sampling

In many administrative surveys, studies are done on large populations which may be geographically quite dispersed. To obtain the required number of subjects for the study by a simple random sample method will require large costs and will be inconvenient. In such cases, clusters may be identified (e.g. households) and random samples of clusters will be included in the study; then every member of the cluster will also be part of the study. This introduces two types of variations in the data – between clusters and within clusters – and this will have to be taken into account when analysing data.

d. Multi-stage sampling

Many studies, especially large nationwide surveys, will incorporate different sampling methods for different groups, and may be done in several stages. In experiments, or common epidemiological studies such as case-control or cohort studies, this is not a common practice. For details of these methods, see Levy and Lemeshow.

References and further reading

Cochran W.G. Sampling techniques. New York, John Wiley and Sons, 1977.

Fleiss H. Statistical methods for rates and proportions. New York, John Wiley and Sons, 1981.

Kish L. Survey sampling. New York, John Wiley and Sons, 1965.

Levy P.S., Lemeshow S. Sampling of populations: methods and applications. New York, John Wiley and Sons, 1991.

Sample size determination: a user's manual. Geneva, World Health Organization, 1986 (WHO/HST/ESM/86.1).

Schlesselman J.J. Case-control studies. Oxford, Oxford University Press, 1982.

Yates F. Sampling methods for censuses and surveys. London, Charles Griffin, 1981