Sample Size Calculation and Sampling Techniques

Syed Asad Ali
Department of Paediatrics and Child Health, Aga Khan University, Karachi.
Email: asad.ali@aku.edu

Sample Size Calculation and Sampling Techniques:

Calculating and justifying sample size for a study can be an intimidating task for new researchers. Nonetheless, it is one of the most important aspects of any study. In this article, we aim to provide basic and fundamental information about calculating sample size which should be helpful for most research studies.

In research studies, a sample has to be obtained if it is not possible to include all the subjects in the study population. For example, if you want to study the prevalence of smoking in the first year medical students of your medical college, you could just ask all of the students in that class to fill the questionnaire, without doing any sampling. However, if you want to know the prevalence of smoking in all the medical students of Pakistan, then it will be very difficult for you to have each and every medical student to fill your questionnaire. In that case you will need to do sampling.

For sampling to work correctly, it is critical that the sample is free from bias. This means that everybody in the study population should have a more or less equal chance of being included in the sample. If sample is collected in such a way that some members of the study population have less chance of being included than others, then the resulting sample is a biased sample, and the results obtained from such a sample will not be valid or generalizable to the entire study population.

While calculating sample size, the first thing to note is that methods to calculate sample size vary depending on the study design. For example, calculating sample size for a survey has a different methodology than to calculate sample size for a case control study or a clinical trial. However, the fundamental principles remain the same.

There are many statistical softwares and online websites which can help you in calculating sample size for your studies (Figure). However, you need to have following four parameters in hand before you can use those resources:

(1) The effect size

(2) The population standard deviation (for continuous data);

(3) The desired power of the experiment to detect the

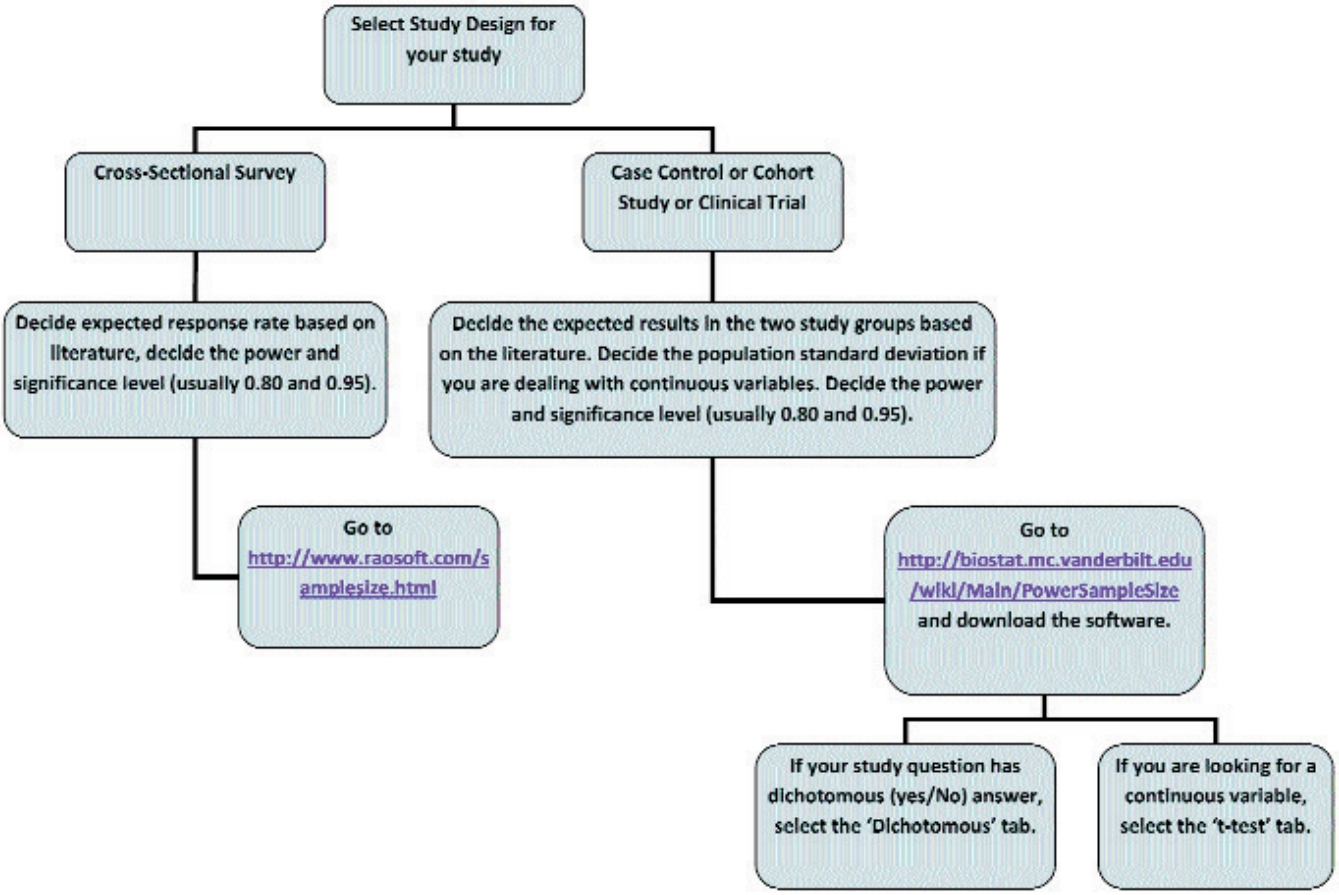

Figure: Approach and online resources for calculating sample size for different study designs.

postulated effect;

(4) The significance level.

The effect size reflects the expected difference in the two study groups of your study. For example, if you know that the current cure rate for disease X with drug A is 20%, and you think that the new drug B will have cure rate of 40%, then your expected effect size for drug B is double that of drug A. The larger your expected effect size, the smaller will be your required sample size. However, if you overestimate your effect size, you will be at risk of having false negative results. While you cannot know the effect size for sure before doing the study, you must have some rationale for choosing a particular estimated effect size. This is done based on previously published studies, or your own preliminary data.

If you are doing a survey, then the effect size means the estimated response rate to your study question in the population. For example, if you are surveying the <15 years old population of Karachi for measles antibodies, you will need to know the total < 15 years old population of Karachi and the estimated prevalence of measles based on your literature review. This information will need to be entered in the sample size calculation formulas available at the listed websites or softwares. If you have no idea about the expected response rate for your study question (measles sero-prevalence in the above example), then a response rate (measles seropositivity) of 50% is selected as a convention, as this response rate gives you the maximum sample size.

Population Standard Deviation is only needed if you are measuring a continuous variable. It is a measure of how spread out your data is in the study population. This is also obtained from previously published studies or your own preliminary data. For a yes/no (dichotomous) question, it is not required to find the standard deviation.

The power of the study means the ability to find a difference between the two groups being studied, if there is actually a difference. In the above example of drug A vs. drug B for the Disease X, even if drug B was actually better than drug A, it is still possible that by chance, they may appear to have same efficacy in your study which is based on a sample. This possibility is called beta error. If we want to be close to 100% sure of our finding (i.e. no beta error), then we will need a very large sample size. As a convention, scientists are

Table: Overview of Sampling Techniques.

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. *Examples of Probability Sampling* In a simple random sample, all members of the study population have an equal chance of being selected in the sample. Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element from then onwards. Stratified sampling is used when the study population has number of distinct categories or strata. For example, if we are drawing a sample of Pakistan, then each of its provinces could be considered as a stratum. In stratified sampling, each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.	Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/ 'undercovered'), or where the probability of selection can't be accurately determined. *Example of Non-probability Sampling* Convenience sampling involves the sample being drawn from that part of the study population which is close to hand. That is, a population is selected because it is readily available and convenient. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough.

happy to be 80% sure that if there is a true difference in the two groups, their study will find it. Hence conventionally, 80% power (or 90% in some cases) is taken as standard. 80% power corresponds to a beta error of 20%.

The final thing to decide on while calculating sample size is the significance level. Taking the same example of two drugs A and B for Disease X, it is possible that while in truth, there was no difference in the cure rates of the drugs A and B, just by chance in your study, drug B shows better cure rate than drug A. This is called the alpha error. If we have the alpha error of 0, then we will need a very large sample size. In general scientists accept the alpha error of 5% in their studies. Alpha error of 5% is also written as significance level of 95%.

In calculating sample size, the effect size and population standard deviations have to determined (and justified) by the scientists with the help of previous literature or their own preliminary data, while the power and significance level are taken as 80% and 95% respectively as a convention.

Once you have the information and references for your expected effect size and your population standard deviation, you should go to a relevant statistical software (e.g Epi Info, SPSS etc) or website as indicated in Figure 1. Once you feed in your numbers, the programme will give you the required sample size. The websites also often provide the text that you can use in your study proposals and papers to communicate the methodology of how you calculated your sample size. However, please note that you have to give the justification of how you calculated your effect size yourself, with the help of previous research studies in the field.

After you get the required sample size estimate from the above websites, the next step will be for you to select an appropriate sampling technique. Sampling frame is the source material from which a sample is drawn. There are different methods through which a sample can be selected from a sampling frame. These methods can be broadly classified as probability and non-probability sampling methods. A good example of probability sampling technique is simple random selection. For example in the above study, if you had a list of all the medical students of Pakistan, you could randomly select the students (total number based on your sample size calculation) and have them fill the study questionnaire. If this not possible to do because of operational challenges, then there are various other options available. These methods are defined in the table.

It is important that the correct sample size calculations be done before initiating the study. This is critical for answering the study questions correctly. Good journals do not accept studies for publication without proper sample size calculations. Sample size calculation is also a requirement for all study proposals. The little effort done in doing formal sample size calculation and defining proper sampling techniques for your study are one of the most important aspects of your study.