SBV AHEAD

REVIEW

HOW BIG A SAMPLE DO I REQUIRE?

Dr.G.Ezhumalai, *Sr. Statistician & Research Consultant*

Sri Balaji Vidyapeeth - Mahatma Gandhi Medical College and Research Institute Campus

Pillaiyarkuppam, Puducherry - 607403, India.

ABSTRACT ▶ The objective of this article is to create awareness on the importance of sample size which decides the validity and quality of the research outcome. It also provides ideas on what information is needed when consulting a statistician for sample size determination. Sample size calculation is scientific, should be reported with relevant formula and justification in every proposal/report of a research project.

INTRODUCTION

When planning a research, the investigator is worried about *"how many samples should I include in my study"* and the statistician is also very particular about it. One should never compromise in fixing the sample size for any research activity. Why does it is important? Usually, sample data will be collected from a population to answer a question or to test a hypothesis. Hence, by doing sampling, one wants to be able to generalize the findings to the whole population, if possible. The sample estimates will be generalisable to the population values only when the sample represents the population. The question is to get a reliable estimate of the population, how many sample does the researcher need to have?

THEORY

Sample size depends upon the aims, nature and scope of the research. All these need to be carefully addressed[1]. The quote "you need 30 samples for statistical significance" may hit millions of hits in whatsApp or similar applications. First, the 30-sample rule-of-thumb was originated by William Gosset, a statistician after being bamboozled by a correlation coefficient experiment for 750 times. Student t distribution was his own concept and it follows normal distribution as the sample size nears thirty[2]. The second point is "law of large number". It says that when we take more samples, the more likely the estimates will be reliable to the population value. In view of these two rules, are we now comfortable with a sample size of thirty? By law of large number, we may think that 30 samples will give a good result; but 40 samples will give better estimate than 30 samples: 50 samples would yield results better than 40 ….. What is the limit, then?

WHY IT SHOULD BE SCIENTIFIC?

Estimation of sample size is one of the important aspects in conducting a research study. An excessive sample size may result in waste of materials, time and money because equally accurate estimates can be obtained from a smaller sample size. On the other hand, a lower sample size is also wasteful, since an insufficient sample size has a low probability of detecting a statistically significant difference even when a difference is really present[3]. Further, when an investigator does not find a significant result, he can't say whether the intervention doesn't work or the insufficient sample size has not able to find the difference.

Finally, population parameters are represented by a confidence interval rather than point estimate from a sample. The confidence interval has a component called "standard error" which determines the narrowness of the interval. This error tends to a minimum with higher sample size and we get a reliable, estimate of the population. Hence, estimation of sample size requires a scientific approach to get better estimates of the population parameter.

KEY COMPONENTS REQUIRED TO ESTIAMATE SAMPLE SIZE?

It is neither feasible nor practical to study the entire population for any research problem. The aim of calculating an adequate sample size is to estimate the population values with a good precision. It can be estimated by using a simple formula with relevant inputs. Thus, the sample size obtained by calculation will adequately represent the population, provided, an appropriate sampling technique is also used. The results obtained from this sampling procedure will signify the true value of the population and the inferences are quite likely to be realistic[4].

The parameter required for calculating sample size are,

1. Variable of interest
2. Type I error (α)
3. Type II error (β)
4. Whether one tailed or two tailed
5. Study design
6. Effect size
7. Precision
8. Prevalence

1. Variable of interest

The scale of measurement of variables is mainly grouped into three as categorical, ordinal and continuous. Description of these variables are distinct (percentage and mean) and hence the formulae for calculation of sample size also differ. Generally, categorical outcome variables require a higher sample size compared to continuous variable for the same precision[5].

2. Type I error (α)

Type I error known as α-error occurs when we "reject the null hypothesis when it is actually true". When we fix significance level of α as 0.05, and obtain a probability value of 0.04 in two tailed, there are two ways to explain. First, really there exists a difference between the two groups and the second due to chance alone; but it is only 4%. When the p value is close to zero, the difference found in the study will be very low. Level of significance is set at 0.05 in many of the studies by convention. Lower the set alpha level, larger the sample size[5].

3. Type II error (β)

Type II error is defined as "we do not reject the null hypothesis when it is false". Type II error is related to power of the study. In many occasions, the power (1-β) of the study is set at 80% and increasing the power will give a higher sample size[6].

4. One tailed or two tailed

The direction of effect between two groups is important since the corresponding standard normal Z values are used in sample size calculation formula. Generaly, two tailed test of hypothesis is used in inferential analysis, unless the direction of effect is known. For example, when the claim is a superiority trial, single tail can be used and when the direction of change is not known two tailed is the choice. The sample size calculated using two tailed is always higher than that of one tailed. One tailed tests are more powerful, but two tailed have a stronger justification[7].

5. Study design

Study design controls the power of a study. Based on whether the study uses one group or many, whether it is observational or experimental, the sample size formula will vary. Descriptive studies need a larger sample to give acceptable confidence intervals.

6. Effect size

Effect size is defined as the ratio of difference between means to the standard deviation. Higher effect size yields a higher power of the study.

7. Precision

How precisely, the sample statistic is estimated is called precision. Standard error is a measure of precision and when it is small, estimates from the sample will be nearing the population values.

While estimating the sample size, the researcher is free to allow some percent of error. Based on the situation, it can be fixed to 5% or 10% with justification. For example, suppose an anticipated population proportion is 20% and a 5% precision level is fixed, the expected population proportion will lie between 15-25%.

8. Prevalence

The proportion (prevalence) of outcome variable related to the objective may be collected from review of literature. If one has several proportions from literatures, the value with similar study design, study population and the most recent values will be used.

In case, it is not available, an assumed prevalence of fifty percent (p=0.5) can be used in the formula or estimated by doing a pilot study.

There is not a single formula to calculate sample size but several according to the situation many. The investigator can choose some of the above requirements, the rest by doing review of literature or by doing a pilot study. A number of online/downloadable softwares are available to estimate sample size. After collecting the required inputs, and substituting the values in the formula, the minimum sample size can be calculated. Sample size will also be determined by negotiation based on the availability, cost involved, duration of research, risk involved and ethical issues in the study. Attrition, withdrawal, drop-out or death of animals is also kept in mind while finalizing the sample size.

CONCLUSION

The sample size must be adequate to answer a research question; but too large is waste of resources while too small inadequately represent the population parameters. Sample size calculation with relevant input and utilization of correct formula is more scientific and reasonable. It has its own merits in deciding the accuracy and precision of the estimates. The relevant precision must be decided by the researcher with up-to-date knowledge by doing a review related to their objectives. The researcher should always mention the process of sample size calculation in her/his proposal/report with justification.

REFERENCES

1. S.K.Lwanga and S.Lemeshow. Sample size determination in health studies – A practical manual. World Health Organization, 1991
2. Student. Probable error of a correlation coefficient., Biometrika, 6(2-3),302-310, 1908
3. Jayakaran charan, N.D.Kantharia. How to Calculate Sample Size in animal studirs? Journal of pharmacology and pharmacotherapeutics, I4(4), 303-306, 2013
4. Baoliang Zhong, MD. How to Calculate Sample Size in Randomized Controlled Trial? Journal of thoracic disease 1:51-51 2009
5. KP Suresh, S Chandrashekara. Top of Form Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci. 2012 Jan-Apr; 5(1): 7-13.
6. Prashanth Kadam, Supriya Bhalerao, Sample size calculation, International journal of Ayurvedha research vol 1(1), p55-57, 2010
7. Stata power and sample size reference manual release 13, A stata press publicaiotn, Stata corporation Ltd., College station, Texas, 2013