MINI REVIEW

SAMPLE SIZE CALCULATION FOR EPIDEMIOLOGICAL STUDIES

Maria Rahim[1*], Mir Ghulam Hayder Talpur[2]

ABSTRACT

Any research project's design needs to incorporate a definition of the sample size required to conduct the investigation. The number of patients needed to achieve the desired study goal is estimated by this sample size. For this purpose, we do sampling to achieve the desire result of research in term of time and money. This article explains how to determine the sample size in the types of studies that are most frequently observed in health research and how to estimate it using the many statistical programs that are included in the commonly utilized sample size calculation software. To ensure reliable and valid result determining an appropriate sample size is crucial. The formula for calculating sample size varies depending on whether you are estimating a mean or a proportion (percentage). It also depends on whether the population is finite or infinite. Similarly, commonly used statistical programs such as G*Power, R software, PASS, and OpenEpi are frequently mentioned.

Keywords: Sample Size, Epidemiological, Proportion Studies, Diagnostics Studies, Mean Difference, G-Power, R Software, Open-epi

Cite this article as: Rahim M, Talpur MGH. Sample size calculation for epidemiological studies. Baqai J Health Sci. 2025;26(1): 45 - 49

Date of Submission: Feb 20, 2025
Date of Acceptance: June 03, 2025
Date of Online Publication: June 30, 2025
DOI: https://doi.org/10.63735/baqa.v26i1.45-49

[1,2]Institute of Business Management (IoBM), Korangi 3, Karachi

INTRODUCTION

To conduct epidemiological study, determining the appropriate sample size is crucial step. To ensure that study is adequately powered to detect meaningful result and association between variables [1-3]. Along with an appropriate and comprehensive experimental design, the hypothesis of the study must be validated through a consistent relationship between the number of observations made, their potential repetitions, their representative, and the overall quality of the evidence. This quantity of observations or samples is known as the sample size (SS) and likewise denoted with the letter *n* [4]. A set of mathematical formulas that have been created to ensure accuracy in predicting population parameters or to produce meaningful results in studies that compare multiple treatment plans or groups are applied in the SS calculation.

Establishing the SS prior to the study's implementation is crucial since it ensures that a sufficient number of patients will be recruited. If this is not done, there is a risk of collecting insufficient data, leading to imprecise results and a failure to detect significant differences that may actually exist. On the other hand, collecting excessive data can result in unnecessary tests, wasting time and resources. Accurate calculation of sample size requires relevant background information, which can be obtained from previous research articles, books, or pilot studies. Pilot studies mean small-scale investigation conducted in the same settings as the global or larger study, but with a restricted 10% of actual sample size. Pilot studies help identify

***Corresponding Address**: PhD Scholar (Statistics and Scientific Computing), Institute of Business Management (IoBM), Korangi 3, Karachi. Email: rhm.maria@gmail.com

potential implementation issues during the study. They also provide preliminary data that can be used to determine the final sample size. Certain studies have trouble finding the required number of participants; these studies typically involve rare diseases, such idiopathic solar urticaria, for which there aren't many cases. Even in these situations, though, it is still wise to ascertain the SS, making an effort to conduct the study across different situation, with each situation contributing a specific number of instances. In cases when such research is not feasible, the number of subjects who can be recruited for the research study is taken into account when making decisions on SS; nevertheless, this entails a significant inconvenience in terms of decreased precision [5-7].

It is not uncommon for studies to define two main aims and several supplementary objectives. Theoretically, each primary objective should be associated with a distinct sample size (SS) calculation. Selecting a smaller sample size reduces statistical power. Ideally, the primary objectives requiring the largest sample sizes should be prioritized. Otherwise, those main objectives with insufficient sample sizes may ultimately be downgraded to secondary or exploratory aims [8-10].

The size of the origin population affects the SS to some extent. To determine the required sample size for the research study, we typically begin by assuming that subject participated in study involve populations of unknown or infinite size. We may need to sample populations of limited size, or N, in some investigations. This is especially important in descriptive surveys, where the population size must be considered when calculating the sample size. In practice, sample size formulas that include the population size (N) tend to converge with those that do not when N is large. Most researchers consider a population to be finite if it consists of fewer than 100,000 individuals.

Therefore, it is typical for the investigator to determine the number of observations based on the time available for conducting the study, the available financial and human resources, and other factors.

ESSENTIAL CONSIDERATION FOR CALCULATING OF SAMPLE SIZE

When estimating sample size, researchers must first consider several key factors, as these determine which formula is appropriate. The main components include the following; [10,11]

1. The Study type involved: observational, descriptive, experimental, animal and in-vitro. In descriptive studies with finite populations, researchers need to know the population size, N.
2. We are prepared to accept errors of type I (á) and type II (â). With the following exclusions, we use á= 0.05 and á = 0.10 or 0.20 as standard values in case of uncertainty, For descriptive research, the precision (width or amplitude of the confidence interval) and the á error, or estimation confidence interval, are all that are needed. On the other hand, in experimental or observational studies, both á and â errors are required to calculate sample size.
3. The response indicates the variables that need to be monitored along with their degree of measurement (that is, whether they are means or proportions (%) or quantitative or qualitative).
4. The smallest variation to be detected between the treatment groups or between the null hypothesis and the alternative hypothesis. Depending on the study at hand, this will vary. The more participants we need to include in the study, the smaller the difference we hope to find. This difference ought to be both reasonable and clinically relevant. The amplitude of the confidence interval determined during the estimate process in descriptive research indicates the difference.
5. When analyzing quantitative variables, one must take into account their variability, which is expressed in terms of variance or standard deviation (SD). The number of participants needed is significantly less when there is less variability than when the variability related to the trait under analysis is high. Pilot studies or

sources in the literature can provide the variability.

6. The hypothesis test's skewness, or laterality, determines whether it has two or one tails. One-tailed tests frequently call for a smaller sample size than two-tailed tests, however the first kind should only be taken into consideration if the test's direction is clear
7. Losses reported for follow-up or patient localization. The sample calculation should account for these losses.
8. The variety of groups that require to be contrasted and the contrasts that should be drawn between them when comparing multiple groups, the formulas used to calculate SS need to specify how many groups were taken into account during the investigation. If this isn't done, there's a chance that ? mistake will spread and beyond the 5% original threshold [11,12]

Sample Size Calculation

For calculating sample size, different formula required according to objective to research, we can distinguish them accordingly as descriptive studies, experimental studies, in-vitro, animal studies.

Descriptive Studies

Studies that aim to determine demographic parameters typically mean and proportions or percentages are referred to as descriptive studies. There is a discernible difference between the studies that have finite populations and those that have infinite populations [12].

Finite Population:

Estimation of a proportion is, percentage or prevalence should calculate through the following formula.

$$n = \frac{t_\alpha^2 * p * q * N}{(N-1) * e^2 + t_\alpha^2 * p * q}$$

Where, n = sample size to be calculated; N = the population's size that the sample is taken from; p = estimated proportion of the population with the characteristic of interest q = 1 - p (this represent the proportion without the characteristics); e = accepted margin of error (usually between 5 and 10%). (This error is one-half of the width of the confidence interval calculated for the parameter, or equivalently 2*e is the width or amplitude of the interval. It is expressed as a percentage); t^2 = value of the normal curve associated to the confidence level. For a confidence of 90%, the value is 1.64 ,95%, this value is 1.96; and for a confidence of 99% it is found to be 2.57 (for two-tailed testing).

Mean Estimation:

The formula for calculating mean estimation is given below, in situation we find out the different in mean formula for calculating sample size (n) when sampling from a finite population. This specific formula is often referred to as Yamane's Formula for Sample Size for Finite Populations where s population variance.

$$n = \frac{t_\alpha^2 * s^2 * N}{(N-1) * e^2 + t_\alpha^2 * s^2}$$

The above mentioned situation can be calculated in all available software of sample size calculation, G-Power, open epi, PASS and Rpackages by inserting the result from previous studies.

Infinite Population

The size of the population has no bearing in the case of infinite populations, and the sample size equations can be reduced to provide the following expressions: Calculation for a proportion:

$$n = \frac{t_\alpha^2 * p * q}{e^2}$$

Calculation for a mean variable

$$n = \frac{t_\alpha^2 * s^2}{e^2}$$

Observational and Experimental Studies:

Sample sizes are calculated when comparing two proportions and two means in experimental or observational research. These computations are not affected by the population size N.

In-case of two proportions:

In case of comparison of two proportions, for example the percentage of improvement after the vaccine of two different clusters of patients, the formula to estimate the sample size would be

$$n = \frac{\left(t_{\alpha} * \sqrt{2 * p * q} + t_{\beta} * \sqrt{p_1 * q_1 + p_2 + q_2}\right)^2}{(p_1 - p_2)}$$

Where, n would be number of patient required per group,

p_1 = proportion in the first cluster group. This often comes from prior research, pilot studies, or clinical assumptions.

p_2 = proportion in second cluster group, â should be between 10 and 20%.

In – case of comparison of two means

In case – of comparison of two mean, the SS formula would be,

$$n = \frac{2 * s^2 * (t_{\alpha} + t_{\beta})^2}{(\overline{x_1} - \overline{x_2})^2}$$

In essence, this formula helps you determine the minimum number of participants needed in each of two independent groups to detect a specified difference between their means, with a given level of confidence (á) and a desired probability of detecting that difference (power, 1?â), assuming a common variance (s2) across the groups.

In case of more than two groups, proportion Chi-square formula would be applied , furthermore for more than two mean ANOVA formula would be applied for comparing mean of more than two groups, both calculation only available in G-Power software, other software is unable to provide the per group sample size to require the study efficiency.

Sample size for animal studies:

For animal studies resource equation method would be applied based on crude method of law of diminishing return. It is employed in situations where it is impossible to estimate the effect size, when many endpoints are measured, when sophisticated statistical procedures are used for analysis, or when it is impossible to estimate the standard deviation in the absence of prior findings. This approach can also be applied in certain exploratory investigations when the researcher's main goal is to identify any degree of variation across groups rather than to verify a hypothesis [13-15].

According to the above mention law a value “E” should be calculated for the sample size of animal studies, which is the degree of freedom of analysis of variance (ANOVA). The value of E must fall within the range of 10 and 20. If E is less than 10, then having more animals will increase the likelihood of getting more meaningful results; but, if E is more than 20, then it won't increase the likelihood. This method can be applied to any animal experiment, regardless of whether it is ANOVA-based. Any sample size that keeps E between 10 and 20 should be considered adequate. The following formula can be used to compute E:

E= Total number of animals – Total number of groups.

This is an easy method but cannot be compared with the power analysis method. Therefore, it is recommended to apply proper formula.

We would like to recommend that researchers include a statement in their publication that explains how they calculated and justified the sample size.

CONCLUSION

Calculation of appropriate sample size and power analysis have always a major issue in epidemiological studies and analysis. This article provides the proper guideline for the calculation of sample size during planning and designing the study. Other, more difficult to use software programs14, like MBESS (for social and behavioral sciences) or pwr (calculation of power), can also be used to determine sample size. In any event, we can always compute any sample size in research by either specifying a special function, as we have seen in some circumstances, or by just applying the formula straight from the instructions line.

Conflict of interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgment: None
Funding: This research did not receive any funding.

Author's Contributions:
MR: Concept, design, writing and responsible for integrity of research.
MGHT: Final approval of manuscript, editing of manuscript, and responsible for integrity of research.

REFERENCES:

1. Del Águila MR, González-Ramírez AR. Sample size calculation. Allergologia et immune pathologia. 2014;42(5):485-92. DOI: https://doi.org/10.1016/j.aller.2013.03.008.
2. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. John Wiley & Sons. 1991. Weblink: https://books.google.com.pk/books?hl=en&lr=&id.
3. Anderson-Cook CM. Experimental and quasi-experimental designs for generalized causal inference. Weblink: https://www.tandfonline.com/doi/pdf/10.1198/jasa.2005.s22.
4. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiol. 2003;229(1):3-8. DOI: https://pubs.rsna.org/doi/full/10.1148/radiol.2291010898.
5. De Maiois FG, Llovet I, Dinardi G. Latin American perspectives on the sociology of health and illness. 2020. Weblink: https://www.taylorfrancis.com › chapters › edit › introdu.
6. Cohen J. Statistical power analysis for the behavioral sciences. Academic press; 2013. Weblink: https://www.utstat.toronto.edu › CohenPower.
7. Zimmermann FJ. Estadística para Investigadores. Bogotá, D.C., Colombia: Editorial Escuela Colombiana de Ingeniería; 2004. Weblink: https://biblioteca.ucatolica.edu.co/bib/22964.
8. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J biomedic informatic. 2014;48(1):193-204. DOI: https://doi.org/10.1016/j.jbi.2014.02.013.
9. Kang H. Sample size determination and power analysis using the G* Power software. J education evaluat healt professions. 2021;18(1): 1-5. DOI: https://doi.org/10.3352/jeehp.2021.18.17.
10. Ko MJ, Lim CY. General considerations for sample size estimation in animal study. Korean J Anesthesiol. 2021;74(1):23-9. DOI: https://doi.org/10.4097/kja.20662
11. Lakens D. Sample size justification. Collabra: Psychol. 2022;8(1):33267. DOI: https://doi.org/10.1525/collabra.33267.
12. Zhang X, Hartmann P. How to calculate sample size in animal and human studies. Front Med. 2023;10(1): 1-5. DOI: https://doi.org/10.3389/fmed.2023.1215927.
13. Charan J, Kantharia N. How to calculate sample size in animal studies?. J Pharmacol Pharmacotherapeut. 2013;4(4):303-306. DOI: https://doi.org/ 10.4103/0976-500X.1197.
14. Eglen SJ. A quick guide to teaching R programming to computational biology students. PLoS computat biol. 2009;5(8):e1000482. DOI: https://doi.org/ 10.1371/journal.pcbi.1000482.
15. Kang H. Sample size determination and power analysis using the G* Power software. J educat evaluat healt prof. 2021;18(1): 17-23. DOI: https://doi.org/ 10.3352/jeehp.2021.18.17.