COMMENTARY

Sample size calculations in health research: Contemporary issues and practices

Amiel Nazer C. Bermudez*[1,2,3], Kim L. Cochon[4]

*Corresponding author's email address: acbermudez@up.edu.ph

[1]*Department of Epidemiology and Biostatistics, College of Public Health, University of the Philippines Manila, Manila, Philippines*
[2]*Department of Epidemiology, School of Public Health, Brown University, Providence, Rhode Island, United States of America*
[3]*Philippine Initiative for Research, Service, and Training, Brown University, Providence, Rhode Island, United States of America*
[4]*JC School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR*

ABSTRACT

Sample size computations, which should be done at the planning stage of the study, are necessary for research to estimate a population parameter or test a hypothesis. For causal analysis of observational databases, sample size computations are generally not needed. Post-hoc power analyses, which are typically done with non-significant findings, should not be performed since reporting post-hoc power is nothing more than reporting *p* values differently. While sample size calculations are typically based on the tradition of significance testing, sample size calculations based on precision are feasible – if not preferred – alternatives. Sample size calculations depend on several factors such as the study objective, scale of measurement of the outcome variable, study design, and sampling design. Computing for sample size is not as straightforward as presented in textbooks but specific strategies may be resorted to in the face of challenges and constraints.

Keywords: *sample size, power, precision*

Introduction

An essential aspect of any quantitative health or health-related study is determining the appropriate sample size needed to answer research questions. However, in most settings, sample size calculation is complex since the investigator has to consider other factors aside from the technical aspects of the study. In addition, the appropriateness of some current practices in sample size calculation has been increasingly scrutinized, with mounting literature supporting the need to abandon these practices. In this commentary, we review basic principles in sample size calculation and introduce some contemporary perspectives or approaches which can serve either as supplements or alternatives to current practices. We also provide a list of tools that health researchers might find useful.

Why and when should we compute for sample size?

Sample size calculation is usually needed when planning a study that employs quantitative methods. A sufficient sample size and an appropriate sampling design improve the study's external validity and can mitigate the distortion caused by outliers and influential observations. For studies that test hypotheses, a sufficient sample size also ensures that statistical tests have sufficient power to correctly reject the null hypothesis. In addition, underpowered studies generally yield imprecise estimates that are only weakly informative.

However, not all studies that employ quantitative methods require sample size calculations. Descriptive studies (*e.g.*, case series) generally do not require a minimum sample size, as do studies that merely describe the distribution of some phenomenon (*e.g.*, distribution of a prognostic factor such as age in the entire trial sample). In addition, for causal analyses involving pre-existing observational databases (*e.g.*, electronic health records), power and sample size calculations are not necessary. Since the goal of any causal analysis is to quantify a causal effect as unbiasedly and as precisely as possible, observational analyses that yield imprecise estimates - from a sparse sample, for example - should not be avoided [1]. In such settings, it is preferable to have multiple studies that yield imprecise estimates – that can be combined in a meta-analysis later on to yield more precise estimates – rather than having no study at all, and this is particularly important for

observational analyses of rare outcomes [1]. Also, exact pre-analysis calculations of power and sample size are often tedious and impractical in settings where different biases have to be simultaneously adjusted for, and/or when study variables are time-varying [1].

Sample size estimation must be done *a priori* or during the planning stage of a study that employs primary data collection. Determining sample size *a priori* ensures better planning of the implementation and financial aspects of the study. From an ethical perspective, including in the study, too many or too few participants than necessary can be viewed as unethical as participants may be exposed to risks in a study that is potentially non-informative. It is a good idea to involve a biostatistician or epidemiologist in planning the study to ensure that sample size calculations are done correctly. While some basic sample size calculations (*e.g.*, estimation of a proportion) can be done quite easily using web-based calculators and with minimal supervision, in the majority of instances, sample size calculation is not too straightforward, and may require, for example, balancing the need for precise estimates with resources available for the study.

Commonly, when study findings are not significant, some researchers perform post-hoc power analyses to distinguish between true negatives (*e.g.*, a conclusion of no effect when there is actually no effect) and false negatives (e.*g.*, a conclusion of no effect when there is actually an effect). When post-hoc (or observed) power is low, researchers usually interpret this as evidence that an effect actually exists, only that the sample size was too small to detect it. However, reporting post-hoc power is nothing more than reporting *p* values in a different way - that is, low post-hoc power is to be expected when the results are non-significant, and vice versa. Thus, contrary to the purported intent, post-hoc power analyses do not actually differentiate between true negative and false negative results and should therefore not be performed when faced with non-significant findings [2-4].

A sample size estimate does not arise from thin air

Commonly, sample size calculations are based on statistical power, which is heavily influenced by the tradition of statistical significance testing. With some statisticians and epidemiologists veering away from statistical significance testing in favor of confidence interval estimation [5-7], an alternative worth pursuing is to estimate sample size based on precision, for example, by specifying the desired width of the confidence interval for the expected population parameter [8]. For additional information, we refer the reader to Rothman and Greenland (2018); in addition, the {precisely} package in R can be used to perform sample size estimation based on precision [9].

Regardless of the approach used, generally, sample size estimates depend on several factors such as the objectives of the study, scale of measurement of the outcome variable, study design, and sampling design.

- **Objectives of the study.** Existing sample size formulae can be classified according to purpose: (1) estimation of a population parameter and (2) hypothesis testing. Sample size calculations for estimating a population parameter require specification of the confidence level desired, tolerable margin of error, and expected magnitude of the parameter to be estimated. On the other hand, sample size calculations for hypothesis testing require specification of the type of testing to be performed (*i.e.*, one-tailed or two-tailed), level of significance, desired power of the test, and effect size to be detected. Note that the desired power of the test is required in sample size calculations for hypothesis testing but not for the estimation of a population parameter.
- **Scale of measurement of the outcome variable.** Quantitative outcome variables (*i.e.*, ratio and interval scales) are summarized differently from qualitative outcome variables (*i.e.*, nominal and ordinal scales). For example, mortality (a qualitative variable) is summarized using frequencies and proportions, while fasting blood glucose (a quantitative variable) is usually summarized using the mean and standard deviation. The scale of measurement of the outcome variable dictates the summary measures that can be used to summarize the variable, which, on the other hand, determines the sample size formulae to use. Sample size formulae for estimating or testing a population mean are different from those used for estimating or testing a population proportion.
- **Study design.** Study design is an important consideration in sample size calculations because it determines the number of groups compared, the number of times the outcome variable is measured, as well as the measure of disease frequency and measure of causal effect that is appropriate for the study (*i.e.*, risk/rate ratios can be directly estimated from cohort studies and randomized trials only). Consequently, sample size formulae for estimating or testing the risk/rate ratios are different from those used for estimating or testing the odds ratio.

- **Sampling design.** The choice of sampling design can be affected by how large the required sample size is. In the same manner, determining sample size requirements should also consider the efficiency of the chosen sampling design. For example, cluster sampling is less efficient than simple random sampling, and would therefore require larger sample sizes.

Some common issues in sample size calculations, and how to address them

Presented in Table 1 are some issues or concerns regarding sample size calculations commonly encountered in practice and suggestions on how to address them.

Table 1. *Common Issues in Sample Size Calculations*

Issue / Concern	Suggestions on how to address them
I do not have information on some of the parameters needed to compute for sample size.	Although it is ideal to use information from your target population to compute for sample size, pre-existing data are usually absent. In these instances, your assumptions for sample size calculations can be obtained from: 1. Similar studies conducted in other populations 2. Pilot studies 3. Expert opinion consensus 4. Educated guess (ideally a range)
My research project has several specific aims. *I am planning to conduct a cohort study and I have several outcomes for which I want to obtain precise estimates of relative risks.* *I am planning to conduct a case-control study and I have several exposures for which I want to obtain precise estimates of odds ratios.* *I am planning to conduct a cross-sectional study and I have several exposure-outcome combinations for which I want to obtain precise estimates of prevalence ratios.*	These scenarios are typical in health and health-related research projects. Even for studies in which there is only one exposure variable and one outcome variable of interest, you may wish to also obtain precise estimates for related specific aims (e.g., estimate incidence proportions). In these instances, it is advisable to perform sample size calculations separately for each specific aim (or each exposure-outcome combination of interest), then select the largest sample size. This strategy will ensure that your study will be sufficiently powered to address all your aims.
I wish to estimate causal effects in subgroups of the population defined by some variable of interest.	Often, especially for analytic studies (*e.g.*, randomized experiments, cohort studies, case-control studies), you might be interested in determining whether estimated causal effects differ in subgroups of your study population. For example, in a randomized controlled trial of a vaccine, you might be interested in estimating vaccine efficacy (VE) separately for patients with comorbidities and those without. If VE estimates substantially differ between the subgroups, the presence of comorbidities is an effect measure modifier (or moderator) of the causal effect of the vaccine on some outcome of interest – or vaccine effect is heterogeneous. Testing for effect measure modification should be planned beforehand; thus, there should be a separate sample size estimation for this objective since the sample size needed to estimate moderated effects (i.e., effect in the presence of effect measure modification) is usually substantially larger than the sample size needed to estimate the main effect (*i.e.*, effect in the absence of effect measure modification).
I am interested in estimating a population prevalence but my computed sample size requirement is substantially larger than my target population.	When you are interested in estimating a population mean or a population proportion, and if your initially computed sample size is larger than the size of the population from which you will derive your sample, you can apply finite population correction to adjust your final sample size requirement [10]: $$n = \sqrt{\frac{n_0 N}{n_0 + (N-1)}}$$ – n is the adjusted sample size – n_0 is the initially computed sample size – N is the size of the population from which you will derive your sample

Table 2. *Tools and Resources for Sample Size Estimation*

Tools	Comments
OpenEpi [11]	One of the most accessible web-based calculators for sample size. Latest version of the calculator can be accessed at https://www.openepi.com/Menu/OE_Menu.htm.
G*Power [12]	Has better functionality than OpenEpi but requires a steeper learning curve. Latest version of software installers can be accessed at https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower
{pwr} package in R [13]	Can be used to estimate sample size based on desired power. Can be installed from within the R/R Studio environment using the command install.packages("pwr", dependencies = TRUE)
{precisely} package in R [9]	Can be used to estimate sample size based on precision rather than power based on the work by Rothman and Greenland. Can be installed from within the R/R Studio environment using the command install.packages("precisely", dependencies = TRUE)

Tools and resources

In recent years, there has been an influx of tools and resources for sample size estimation including web-based calculators, software, and statistical packages. Regardless of the tool used, researchers should refer to the accompanying documentation of these tools not only to be familiar with how to use them but to ensure that they are used correctly. Listed in Table 2 are some of the tools which the authors have found useful in practice. In addition, researchers who do not have continuous or stable internet access may also refer to the manual on sample size estimation for health studies published for free by the World Health Organization [14].

References

1. Hernán MA. (2022) Causal analyses of existing databases: No power calculations required. J Clin Epidemiol. 144:203–5.
2. Althouse AD. (2021) Post-hoc power: Not empowering, just misleading. J. Surg. Res. United States; p. A3–6.
3. Goodman S, Berlin J. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. United States; 121:200–6.
4. Hoenig JM, Heisey DM. (2001) The abuse of power. Am Stat. Taylor & Francis; 55:19–24. Available from: https://doi.org/10.1198/000313001300339897
5. Wasserstein RL, Schirm AL, Lazar NA. (2019) Moving to a world beyond "p < 0.05." Am Stat. Taylor & Francis; 73:1–19. Available from: https://doi.org/10.1080/00031305.2019.1583913
6. Wasserstein RL, Lazar NA. (2016) The ASA statement on p-values: Context, process, and purpose. Am Stat [Internet]. Taylor & Francis; 70:129–33. Available from: https://doi.org/10.1080/00031305.2016.1154108
7. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, *et al.* (2016) Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol. 31:337–50.
8. Rothman KJ, Greenland S. (2018) Planning study size based on precision rather than power. Epidemiology. United States; 29:599–603.
9. Barrett M. (2021) precisely: Estimate sample size based on precision rather than power. Available from: https://cran.r-project.org/package=precisely
10. Ramachandran KM, Tsokos CP. (2021) Mathematical Statistics with Applications in R (3rd ed.). United States: Academic Press. Available from: https://doi.org/https://doi.org/10.1016/C2018-0-02285-9
11. Dean A, Sullivan K, Soe M. (2013) OpenEpi: Open source epidemiologic statistics for public health. Atlanta: Emory University; Available from: www.OpenEpi.com
12. Faul F, Erdfelder E, Lang A-G, Buchner A. (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods. Kiel, Germany: Psychonomic Society; p. 175–91.
13. Champely S. (2020) pwr: Basic functions for power analysis. Available from: CRAN.R-project.org/package=pwr
14. Lwanga SK, Lemeshow S. (1991) World Health Organization. Sample size determination in health studies: A practical manual. Geneva: World Health Organization. Available from: https://apps.who.int/iris/handle/10665/40062