Journal of INDEPENDENT MEDICINE

The Myth of the 'Correct' Sample Size: Why Formula-Based Justification Misleads Researchers and Reviewers

(Errors and simulations that benefit no one and harm many, depriving them of time and energy)

L. Prieto-Valiente[1], C. Carazo Diaz[2], C. Marín-Barcena[3]

Abstract

This text does not report the results of a research study nor is it a bibliographic review. It does not conform to the formal structure of Original Research Articles or Review Articles. It is a "teaching collaboration" that does not appeal to the authority criterion of any author, but rather to rigorous scientific reasoning, and aims to show that:

a) In most cases it is arbitrary and wrong to require "scientific" proof that the "N" used in a study is appropriate.
b) It is even more erroneous to accept as correct alleged demonstrations that demonstrate nothing.
c) The formulas for estimating the appropriate sample size for each study have little usefulness in most cases, because in reality there are many sizes that are valid and there is not one in particular that is "the appropriate one."
d) Contrary to common belief, the mathematical formulas used in this context do NOT say what N should be used, and therefore it makes no sense for reviewers to demand from authors a "scientific" justification of the size used.
e. Putting an end once and for all to this chain of misunderstandings and errors would save time, effort, and discomfort for many researchers. To do so they do not need to learn more mathematics or implement more sophisticated procedures. Quite the opposite, the correct approach is simpler and more direct than the incorrect one.
f. What is correct is for the researcher to use the number of individuals that their resources allow. If the study is well conducted, that sample will provide useful information. In light of the results provided by it, it will be decided whether it is appropriate to subsequently expand the study by incorporating more individuals.

Keywords: Sample size, interval estimation, common sense

[1] Biostatistics and Big Data. Catholic University of Murcia, UCAM. Scientific Society of Biomedical Research, SCIB, Palma de Mallorca, Spain
[2] Director of the Department of Biostatistics and Big Data. Catholic University of Murcia, UCAM. Scientific Society of Biomedical Research, SCIB, Palma de Mallorca, Spain
[3] Student of the Scientific Society of Biomedical Research (SCIB), Palma de Mallorca, Spain

Corresponding author:
Luis Prieto-Valiente
Rio Negro 2, Los Alcazares , Murcia 30710, Spain
Email: lprieto@ucam.edu

Introduction

When planning a medical research study, one of the first issues to decide is the number of individuals, "N", that we will use. Statistics books contain sections entitled "The minimum sample size required to…", with formulas of the type de $\mathbf{N = Z^2 \cdot \sigma^2 / d^2}$, and many physicians believe that one of these formulas will give them the "appropriate" or "minimum necessary" size for their research. But when attempting to apply it they encounter serious difficulties because they lack certain basic knowledge at that point. Therefore, in many cases they end up taking the number of individuals that their resources allow, although they believe - wrongly - that this is incorrect from the scientific point of view, and they believe - wrongly - that they should justify the chosen size with a mathematical formula. Unfortunately, some textbooks and biostatistics teachers do not sufficiently clarify the issue and help to spread that error. (3)

The same belief is held by reviewers of scientific articles for journals and evaluators of research projects, who often require authors to "rigorously" justify the sample size used. If one of these formulas or some phrase referring to them appears in the text they are judging, they usually assume that this rigorously endorses the size of the sample used. And if they do not see such formulas or phrases in the text, they understand that this endorsement is missing and ask the author to justify the choice of the size used. Generally the author corrects their text at that point, adding one or two sentences that the reviewer usually accepts.

In some cases it is a logical and necessary process, in which the evaluator detects a real deficiency and the author corrects it properly. But in not a few cases the reviewer's request is inappropriate and the sentences the author sends in response to justify the N used justify nothing and on many occasions are manifestly unintelligible. Both parties involved in this dialogue are unfamiliar with the topic and exchange sentences that neither the writer nor the judge understands.

We are facing a collective misunderstanding that affects a large part of the scientific establishment in all developed countries. Putting an end once and for all to this chain of misunderstandings and replacing it with correct methods would save time, effort, and discomfort for many researchers. To do so they do not need to learn more mathematics or implement more sophisticated procedures. Quite the opposite, the correct approach is simpler and more direct than the incorrect one. Any researcher can access it if it is explained clearly.

To achieve this goal the first condition is that we all recognize the problem. (1, 2, 17) The second is that researchers know that replacing incorrect concepts with correct ones does not require learning mathematics and that these are simpler than those, so that to master them one only needs to apply basic logic (4,5). In what follows we see the very limited usefulness of the formulas that give the sample size. (NOTE: Numerical probabilities are expressed in this text interchangeably as a proportion or as a percentage. If in a group of N = 40 there are 10 obese individuals, we will say that the proportion is 0.25 or 25%).

An anecdote to draw attention to an error

Dr. Julia Feynman is beginning her doctoral thesis work and decides to conduct a study to estimate the proportion of obese individuals in the population of Florida.

To determine the number of people she must survey, she visits a biostatistician who, after discussing the case with her, suggests interviewing 27250 people. But on the way back she meets another biostatistician on the train who, after discussing the case with her, suggests interviewing 50 people. Suspecting that she had made some mistake when giving the data requested by the experts or that one of them made an error in their calculations, she speaks again with each of them, but, to her surprise, both reaffirmed their number.

To resolve the conflict she convenes a committee of the best experts in the State. After a brief deliberation those convened unanimously agree that both biostatisticians had acted correctly, and they add that 374 would be an appropriate size.

Julia cannot get over her astonishment, perplexity, and anger, and decides to convene a committee of the best experts worldwide. After listening to her, the new committee members agree that the three sizes previously proposed were correct and add that 1040 would be an appropriate size. Finally, Julia gives up on the "scientific" calculation of the "appropriate size" of the sample and decides to act according to the human and economic resources available to her. She surveys a random sample of 1000 people and finds that 200 of them are obese, p = 0.20). She tries to publish an article saying that in the sample studied 20% were obese and that the Confidence Interval for the population percentage, , was: $IC_{95\%}(\Pi) = \mathbf{p} \pm 1.96 \cdot [\ p\ (1\text{-}p) / N\ ]^{1/2} = \underline{17.5\%}$ y $\underline{22.5\%}$

She sends the manuscript to a scientific journal and they tell her that the topic is interesting and the study is well designed, but they cannot publish it because the sample size is not "scientifically" justified. She then sends it to another journal that publishes it without raising objections.

Is the objection of the first journal justified? Does the second journal act with less "scientific rigor"? We will soon see that it does not. (8)

The formula for calculating the "necessary" sample size to estimate a population proportion

The only way to know the proportion of obese individuals in a population is to assess that value in all the individuals of that population. But in practice we take a sample and calculate the sample proportion, which in general does not have exactly the same value as the population proportion. We expect it to have a "similar" value, but how good will the approximation of the sample proportion to the population proportion be? The answer is:

In general, the larger the sample, the better that approximation will be. (4, 5)

Julia tries to take a sample large enough for the sample value to be very close to the population value, and at the same time as small as possible, to save time and money. It is a general belief that there is an ideal size that is the minimum, and therefore the cheapest, among those large enough to give a good approximation. But the reality is that in most situations there is no such ideal size, and the decision about how many individuals to include in our research does not depend on a statistical calculation.

When it is explained to Julia that, contrary to her belief, there is no size that is "appropriate" for the research she is planning, she responds that in statistics books there are formulas to calculate those sizes and she believes that by applying one of them the ideal size for her project will appear. It is true that those formulas exist, but we will see that each formula can give many different answers, all of them valid, for the sample size sought. Therefore, in the final decision of the sample size to use in a research study, these formulas carry very little weight.

From the sample proportion, p, the researcher calculates a Confidence Interval, CI, within which we believe the population proportion is found. In general, the larger the sample, the more precise the estimation will be; that is, the CI for the population value, CI(π), will be narrower.

This is the formula found in all books and on the Internet. (1,2)

$$N = \frac{p'(1-p')Z^2}{d^2}$$

We see that N depends on three quantities, p', Z and d, whose meaning is:

p´ Indicates the order of magnitude of the population proportion that we are trying to know with the study we are planning. That value is not known; that is why a study is going to be done to find out what it is. The quantity p' to be inserted in the formula must be one that we suppose will be close to the proportion we are trying to determine. If there is no information about it, we set p´ = 0.5.

Note that if p´ = 0.10 then p´ (1-p´) = 0.1 x 0.9 = 0.09

Note that if p´ = 0.30 then p´ (1-p´) = 0.3 x 0.70 = 0.21

Note that if p´ = 0.50 then p´ (1-p´) = 0.5 x 0.50 = 0.25

That is, N will be larger when the population proportion takes a value around 50% than when it is very small or very large.

d The formula allows us to calculate the probability that p differs from by less than any quantity, d, that we choose. That is, that the proportion that appears in the sample does not differ from the population proportion by more than "d": $|\pi - p| < d$. This quantity must be decided by Julia and there are no "mathematical" or "scientific" criteria for doing so. Depending on the value chosen for "d", the formula yields an N.

Z But there is no N that totally guarantees that it will be $|\pi - p| < d$. Julia's confidence that this will occur is greater the larger N is. The Z value in the formula depends on how large we want that confidence to be. For example, for 90% confidence, Z = 1.65. For 95% confidence, Z = 1.96; for 99% confidence, Z = 2.58, etc. (Readers familiar with the NORMAL DISTRIBUTION will recognize that Z is the value that leaves between it and its negative that % of the values of the Normal Distribution. But that knowledge is not necessary to follow the logical reasoning we are dealing with). (13)

Let us apply the formula to have 90% confidence (Z = 1.65) that the p that appears in the sample does not differ by more than 7 percentage points (d = 0.07) from the true population proportion, π, assuming that the population proportion is on the order of 10% (p´= 0.10).

$N = (0.1 \times 0.9)\ 1.65^2 / 0.07^2 = 50$

The chosen values of p', d and confidence are reasonable, as are many other possible ones, and there is no objective reason that makes those quantities more valid than other similar ones. But with other values of p', d and Z the result can be very different. Other possible choices are shown in **Table 1**.

What does each of these sizes guarantee us? If we take a sample of N = 50 we have probability 0.90 that the proportion found in the sample will not differ from the population proportion, π, by more than 0.07, if π is on the order of 0.10. What does that probability 0.9 mean?? That in 90 out of every 100 times that we repeat taking a sample of that size, the sample p differs from the population p by less than

0.07. (16, 14)

And if we take a sample of $N = 374$ individuals we have a 99% probability that the proportion we will obtain in that sample will differ by less than 0.04 from the population proportion we want to estimate, if it is on the order of 0.10. But if the population proportion is on the order of 50%, to have a 99% probability that the sample proportion will differ by less than 0.04 from the population proportion, we need $N = 1040$.

We see, then, that the formula is not to determine the "appropriate size", but rather to calculate the "performance" that each size provides. And in order to use it we have to:

a) Estimate the population proportion, p´
b) Choose the maximum difference that we would not want to exceed between the sample proportion we obtain in the sample we are going to study, p, and the population proportion, π.
c) Choose the confidence we want to have that the above occurs. (10)

Note the difference between "estimating" and "choosing". If someone asks me for the number of inhabitants of my city, I will say that I do not know it exactly, but I ***estimate*** that it is around 500 000. And if they ask me what number of inhabitants I consider most appropriate for a city to be comfortable and have all services, I will say that I ***choose*** around 70 000.

What is the most reasonable and correct attitude?

When Julia asks statisticians what the appropriate sample size is for the research she plans to conduct, they must answer that there is no such "appropriate" size. In the vast majority of cases, the size proposed by the researcher based on common sense and on the availability of time and material resources is perfectly valid. Although, of course, the larger N is, the more information it provides. In many cases it may help the researcher to have a table showing the sizes corresponding to various values of Z, d, and p'. For example, the following table gives N for $d = 0.06$ or 0.02, and $p' = 0.10$ or 0.50, and for confidence levels of 90% or 99%.

In this way, Julia sees that any size between the smallest and the largest in this table, and also others that do not differ much from those extremes, are acceptable. If she decides to study 68 people and the population proportion is on the order of 10%, she will have a 90% probability that π differs from p by less than 6 percentage points, above or below. And that is quite useful information. But if she has a great deal of time and money available, she may study, for example, 4160 people and then she will have a 99% probability that π differs from p by less than 2 percentage points, if π is on the order of 50%. This information will be much more precise, but this does not imply that the sample of $N = 68$ is useless. (11)

In summary: researchers and reviewers must know that there is no appropriate sample size for the research they are planning. And very rarely can it be said that the size is insufficient. In the next article we will see that a study with $N = 5$ can be completely valid and very useful. (9)

In other cases the researcher's objective is to estimate the population mean of a numerical variable, such as blood pressure or blood glucose. And in other cases the objective is to perform a statistical test comparing two means or two proportions, etc. In each of these situations the corresponding formula relates N to other parameters that the researcher must agree upon, and the N obtained varies markedly depending on the values given to those parameters. The evaluators of the second journal acted correctly and those of the first made a mistake that, unfortunately, is very common in our time.

Inadequate sample sizes

Although there is no sample size that is "the appropriate one", since very different sizes are valid, certain sizes are clearly "inadequate", because they are excessively large or excessively small. For example, if in the case at hand Julia had used a sample of $N = 50{,}000$ and found $p = 0.30$ in the sample, the confidence interval for the population proportion would be $IC95\%\ (\pi) = 0.3 \pm 1.96\ [0.3 \times 0.7 / 50{,}000]^{1/2}$ = 29.6% and 30.4%, which is unnecessarily narrow. Probably an interval two or three times wider would be sufficiently informative. And if she had found that same sample proportion in a sample of $N = 20$, it would be $IC95\%\ (\pi) = 0.3 \pm 1.96 \cdot [0.3 \times 0.7 / 20]^{1/2}$ = 10% and 50%, which is clearly imprecise.

Inadequate sample sizes

Although there is no sample size that is "the appropriate one", since very different sizes are valid, certain sizes are clearly "inadequate", because they are excessively large or excessively small. For example, if in the case at hand Julia had used a sample of $N = 50{,}000$ and found $p = 0.30$ in the sample, the confidence interval for the population proportion

would be IC95% (π) = 0.3 ± 1.96 [0.3 × 0.7 / 50,000]$^{1/2}$ = 29.6% and 30.4%, which is unnecessarily narrow. Probably an interval two or three times wider would be sufficiently informative. And if she had found that same sample proportion in a sample of N = 20, it would be IC95% (π) = 0.3 ± 1.96 · [0.3 × 0.7 / 20]$^{1/2}$ = 10% and 50%, which is clearly imprecise.

Here, as in many areas of science and life in general, it happens that there are extreme values that are clearly inadequate, but this does not imply that there is a value that is "the correct one". If we talk about the number of dollars needed to buy a new car, it is obvious that $1,000 is insufficient and $500,000 is manifestly excessive, but this does not mean that there is a "correct" amount. For example, $10,000 and $20,000 and $60,000 and many other amounts are valid. (10)

Certainly, there are contexts, for example randomized trials with planned hypothesis testing, where sample size planning may help avoid underpowered or overpowered studies. But even in those cases, formulas depend on unobservable assumptions and still do not produce a single "correct" sample size. The only reasonable option is to manage a board with different options.

Furthermore, most studies involve interest in more than one variable and the formula will give different results depending on the variable to which it refers. It must also be taken into account that many comparisons of interest will affect subgroups of the sample, whose sizes cannot be predicted in advance.

A succession of lies and simulations that harm everyone and benefit no one

As we have said, many members of the committees responsible for approving and funding projects require that sample sizes be justified with mathematical formulas. Most researchers do not know how to use those formulas nor how to interpret their result. When they are forced to handle them in research projects they include sentences they do not understand, copied from other projects, whose authors also did not understand them and had, in turn, copied them from others. This gives rise to a chain of simulations that leads to an absurd situation. The steps of that sequence are:

1. In the process of "blind" copying, typographical errors accumulate that the copyists do not detect because they do not understand what they are copying.
2. Most of the evaluators mentioned above do not understand those sentences and accept them as valid, including the typographical errors.
3. The approved projects fall into the hands of other authors of new projects, who repeat the copying process adding new typographical errors.
4. The repeated practice of this ritual has led to the fact that in many projects one reads totally absurd sentences, but approved by evaluators.
5. On some occasions this simulation is also carried out orally, at scientific conferences and in the defense of doctoral theses. With regard to sample size, the researcher confidently states an unintelligible sentence and the examining committee solemnly nods. One could speak of a "double-blind" agreement, where neither party understands anything and both simulate understanding. (11)

Conclusions

1. In every study it is necessary to decide how many individuals to include, and many researchers believe — erroneously — that this decision is made by applying a mathematical formula.
2. The formulas related to this topic do not say what size the sample of a research study should have; they only help calculate the performance that the sample will have depending on its size.
3. Some researchers do not know how to use these formulas nor how to interpret their result, but when they are forced to handle them in research projects, they include sentences they do not understand, copied from other projects whose authors also did not understand them and had, in turn, copied them from others.
4. The reality is that there is no "more appropriate" size that optimizes the cost/benefit ratio. Many sample sizes, excluding the extremely small or extremely large ones, are valid.
5. The larger the sample, the more information it gives us about the population. This increase in information as the sample size increases is gradual and there is no cutoff point that separates sufficient sizes from insufficient sizes.
6. The formulas do not tell us which size should be used, but rather how the confidence that the estimation error does not exceed a certain amount depends on the size. This confidence and the magnitude of that maximum error must be chosen by the researcher according to their good judgment. There are no mathematical or biological arguments to help them in that choice.
7. The size yielded by the formula is also influenced by the population proportion of dichoto-

mous variables. Since the researcher does not know that value, they must assume it, which introduces another factor of imprecision.

8. Although it is not appropriate to speak of adequate sizes, it is appropriate to speak of inadequate sizes, either because they are extremely small and give rise to estimation errors with very wide margins, or because they are extremely large and give rise to unnecessarily small estimation errors.
9. In the next article we will see that a study with, for example, N = 5 can be completely valid and very useful. (15)

Table 1. Sample Size Options for Different Precision Targets and Confidence Levels (p' = 0.10 to 0.50)

Confidence	d	p´		N
90%	.07	.10	1.65 2 (0.1 x 0.9) / 0.07 2	50
99%	.04	.10	2.58 2 (0.1 x 0.9) / 0.04 2	374
99%	.04	.50	2.58 2 (0.5 x 0.5) / 0.04 2	1040
99.9%	.01	.50	3.3 2 (0.5 x 0.5) / 0.01 2	27,225

Table 2. Confidence Interval Width as a Function of Sample Size: Showing Both Excessive Precision and Inadequate Precision

	p´= 0.10		p´= 0.50	
	Conf.= 90%	Conf.= 99%	Conf.= 90%	Conf.= 99%
d = 0.06	68	166	267	462
d = 0.02	613	1498	2401	4160

References

1. Abellán-Huerta J, Prieto-Valiente L. The myth of sample size. Rev Esp Cardiol. 2020;73(9):735–738. doi:10.1016/j.recesp.2020.04.017
2. Altman DG. Statistics and ethics in medical research: III How large a sample? Br Med J. 1980;281(6251):1336–1338. doi:10.1136/bmj.281.6251.1336
3. Altman DG, Moher D, Schulz KF. Peer review of statistics in medical research. Reporting power calculations is important. BMJ. 2002;325(7362):491–492.
4. Armitage P, Berry G. Statistical methods for medical research. 4th ed. Oxford: Blackwell Science; 2005.
5. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ. 1995;311(7013):1145–1148.
6. Carazo-Díaz C, Prieto-Valiente L. El valor p de un test no es un índice matemático, es simplemente una frecuencia relativa. Rev Neurol. 2024;78(7):209–211. Spanish.
7. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988.
8. Fisher RA. The design of experiments. New York: Hafner Press; 1935. p. 1–2.
9. Herranz I, Martínez-Sellés M, Alonso R, Prieto-Valiente L. Tests of significance vs tests of hypothesis. Far East J Theor Stat. 2007;21(1):97–111.
10. Guyatt GH, Mills EJ, Elbourne D. In the era of systematic reviews, does the size of an individual trial still matter? PLoS Med. 2008;5(1):e4. doi:10.1371/journal.pmed.0050004
11. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci. 2014;9(6):641–651. doi:10.1177/1745691614551642
12. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing clinical research. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2013.
13. Kelly PJ, et al. How many patients do we need for a clinical trial? Demystifying sample size calculations. Nephrology. 2010;15(8):725–731.
14. Lenth RV. Some practical guidelines for effective sample size determination. Am Stat. 2001;55(3):187–193.
15. Martínez-Sellés M, Prieto-Valiente L, Herranz I. Frequent mistakes in statistical inference of biomedical data. Ital Heart J. 2005;6(2):90–95.
16. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharm Stat. 2007;6(3):161–170. doi:10.1002/pst.301
17. Senn S. Sample size considerations for n-of-1 trials. Stat Methods Med Res. 2019;28(2):372–383. doi:10.1177/0962280217726801

This page is intentionally left blank

 https://doi.org/10.71189/JIM/2026/V02N01A06