Simple sample size calculations for cross-sectional studies

James Beard

Independent Consultant

Correspondence:
james@beard.net

Submitted: October 2024
Accepted: October 2024
Published: November 2024

Introduction

In any research study, it is necessary to decide how much data to collect: too little and the study may not produce a useful result; too much and resources will be wasted. In either case, data collection may be considered unethical. This applies both to the amount of information collected about each participant and the number of participants in the study. The current article is about the latter.

Studies with different outcome measures require different approaches to calculating sample size. Below, we look at some simple scenarios, mainly relevant to cross-sectional studies. Formulae are given, and calculations can often be done by hand (or calculator). However, online calculators exist, and some statistical analysis programs include the ability to calculate sample sizes for different problems. Do not be surprised if sample sizes given by the formulae below differ slightly from those given by online calculators or by analysis software. Such calculators may use more complicated formulae that take into account other factors, such as the size of the population being studied, or may be based on different methods of analysis.

Before looking at specific formulae, we need to consider the so-called normal distribution or bell curve (red line in Figure 1). Many statistical tests and formulae

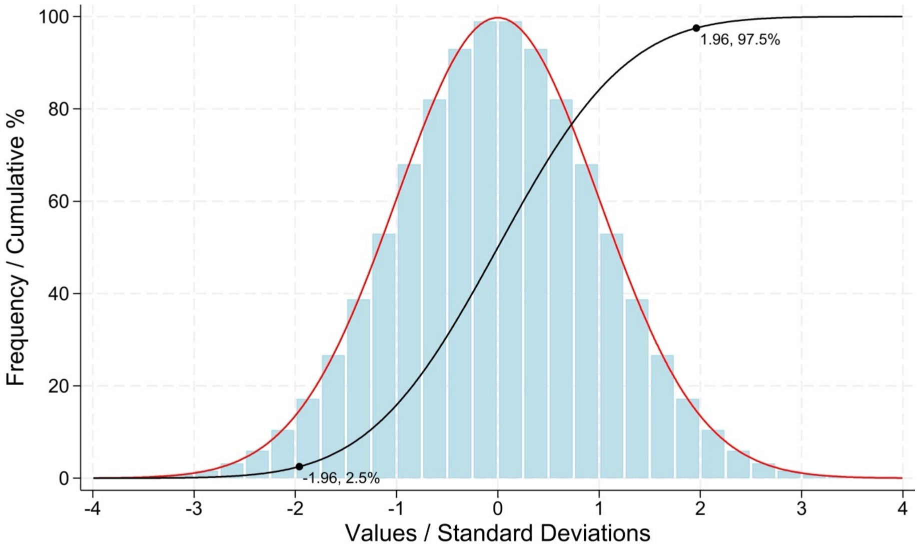

Figure 1. Red line: normal distribution based on 400 observations of a hypothetical variable with mean 0 and standard deviation 1. Black line: cumulative percentage of observations. Blue bars: numbers of observations in 0.25 ranges

Citation: Beard, Simple sample size calculations for cross-sectional studies, South Sudan Medical Journal, 2024;17(4):213-216 © 2024 The Author(s) **License:** This is an open access article under CC BY-NC DOI: https://dx.doi.org/10.4314/ssmj.v17i4.12

are based on the assumption that the observed values of a variable of interest follow a normal distribution, at least approximately. In Figure 1, the standard deviation (a measure of dispersion of the data) is one, so the X-axis shows both the data values and the number of standard deviations from zero, the mean value. The black line in Figure 1 shows the percentage of observations with a value less than or equal to the value on the X-axis. Note the two highlighted points on that line. These tell us that only 5% (2.5 + (100-97.5)) of observations are more than 1.96 standard deviations below or above the mean value. So, we can have 95% confidence that the true value of our variable lies somewhere in the range mean ±1.96 standard deviations.

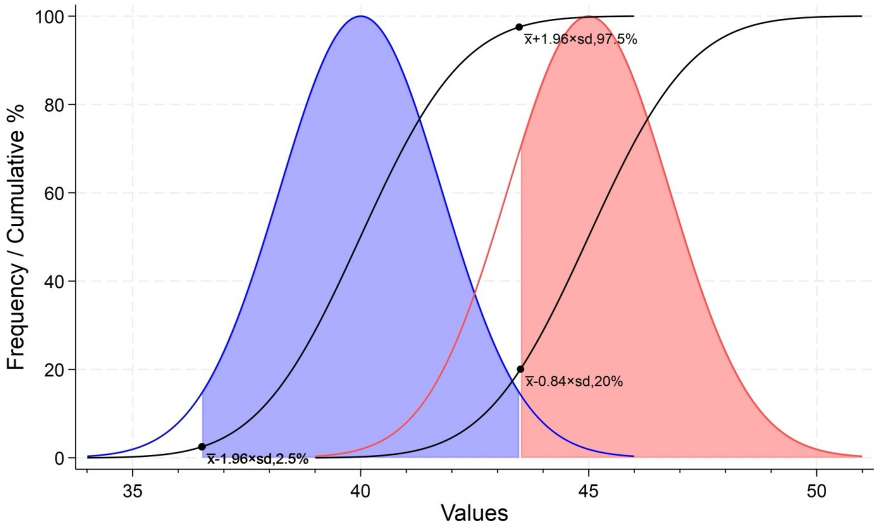

Figure 2. Blue and red lines: normal distributions centred around 40 and 45 respectively. Black line: cumulative percentages of observations for the two distributions. Blue shaded area: 95% of area under blue line. Red shaded area: 80% of area under red line.
$\bar{x}$ = mean, sd = standard deviation

The number of standard deviations from the mean is often referred to as *Z*, which is the meaning of *Z* in the formulae below. One type of *Z* comes from the required confidence interval in our sample size calculations. We have seen above that mean ±1.96 standard deviations gives us 95% confidence, so *Z* is 1.96 for 95% confidence.

Another *Z* is related to the *power* of a study to detect a difference between two groups, which is often set to 80%. For example, we might conduct a study in which we think two groups may have mean values of 40 and 45 for a particular variable. We must ensure that we collect enough data to be able to claim that a detected difference is likely to be real. Figure 2 shows distributions centred around 40 and 45, our supposed mean values for the two groups. In Figure 2, if the red- and blue-shaded areas do not overlap, we get 95% confidence and 80% power to show a difference (the larger the sample, the narrower the distributions will be).

Some useful values of *Z* are shown in Table 1.

Table 1. Values of *Z* for some power and confidence percentages. Commonly used values in bold

Power %	Conf. %	*Z*
80.0	60.0	0.84
90.0	80.0	1.28
95.0	**90.0**	1.64
97.5	**95.0**	1.96
99.0	98.0	2.33
99.5	**99.0**	2.58

Note that the formulae assume that the sample to be taken will be representative of the population studied. There may be practical reasons why this is difficult to achieve; for example, it may be easier to recruit women than men. This has to be allowed for in the study design and, hence, sample size calculations.

The formulae require an estimate of the quantity the study is attempting to measure (a value or a difference). This can often be found in the literature, but it may be necessary to carry out a small pilot study to get an appropriate value. In simple prevalence studies, it is safe to assume 50% prevalence, as this will result in the largest sample size.

Some Formulae

1. The prevalence of an attribute in a population

For example, the prevalence of type 2 diabetes in 18- to 59-year-olds in urban Juba. Cochran's formula is almost always used for such studies:

$$\frac{Z^2 * p * (1-p)}{e^2}$$

The formula's inputs are Z (the first Z mentioned above) for the desired confidence level, p an estimate of the outcome prevalence, and e, the desired absolute precision of the result. Common values used in this calculation are 1.96 (for 95%), 0.5 (50%), and 0.05 (±5%), respectively, which gives a sample size of 385. If ±3% precision was wanted instead, the required sample size would be much bigger, 1068.

2. The difference in prevalence between two groups

If the researchers were not only interested in the overall prevalence discussed in the previous section but also in differences between (say) men and women, a different calculation is needed:

$$\frac{(Z_a + Z_b)^2 * (p_1 * (1 - p_1) + p_2 * (1 - p_2))}{(p_1 - p_2)^2}$$

Here, we have both Zs mentioned above. Z_a is the Z related to the desired confidence level, while Z_b is related to the required power of the study to detect a difference between the two groups. p_1 and p_2 are the estimated proportions for each group. As before, the confidence level is often 95%, so Z_a is 1.96, and the power is commonly 80%, so Z_b is 0.84. Although again, the nearer the estimated proportions are to 50%, the larger the required sample size, the hypothesised difference has more effect on the required sample size. A safe option is to choose two values, around 50%, that are different by the expected difference between the groups. Say the hypothesised difference was 10%, then 0.45 and 0.55 would be appropriate values for p_1 and p_2. These inputs give a sample size of 389 *per group*. Reducing the difference to 5% would give a much larger sample size, 1565 per group.

3. The mean of an attribute in a population

For example, the systolic blood pressure of men aged 60 to 69 in rural South Sudan. The required sample size can be calculated using the formula

$$\frac{Z^2 * \sigma^2}{e^2}$$

where Z is related to the required confidence interval, σ is the estimated standard deviation of the attribute in the population, and e is the desired absolute precision of the result. It does not matter what units are used for σ and e, but the same units must be used for both. The estimate for σ would typically come from other studies conducted in similar populations. If we again want 95% confidence in our result, Z would be 1.96. If we say σ is 20mmHg and we want a result accurate to ±2mmHg, then 385 would be the required sample size.

4. The difference between two means

For example, the difference between mean systolic blood pressure for men and women in the same population. In this case, the formula is

$$\frac{(Z_a + Z_b)^2 * 2 * \sigma^2}{d^2}$$

where Z_a and Z_b are related to the confidence interval and power required respectively, σ is the estimated standard deviation as in Formula 3 above, and d is the hypothesised absolute difference that the researchers wish to detect. Again, σ and d must be measured in the same units. For 95% confidence, Z_a will be 1.96; for 80% power, Z_b will be 0.84. Assuming that the standard deviation σ is again 20mmHg and we wish to detect a difference between the groups of 10mmHg, we would need a sample of 63 *per group*.

Summary

The scenarios above are just a few of the many that could be relevant to a real research study, and they are among the simplest. More complicated scenarios include:

- Comparing more than two groups
- Multiple outcome measures
- Different sample sizes in different study groups
- Different standard deviations in different study groups
- Clustered data (by village or hospital, for example)
- Paired data

In such cases, a competent statistician is needed to advise on the analysis to be done and to calculate an appropriate sample size (this should all be done before any data are collected).

Having calculated the statistically necessary sample size, researchers must consider whether this sample would be enough in practice. For example, the assumptions on which the sample size calculation was based may have been optimistic. Some potential participants may refuse consent, and others may be lost to follow-up (depending on the design of the study). Thus, the target sample size may need to be greater than the calculated minimum.

Smith, Morrow, and Ross provide a more detailed discussion of this topic.[1]

Internet Resources

There are many websites with a selection of easy-to-use sample size calculators, including:

- https://select-statistics.co.uk/calculators
- https://epitools.ausvet.com.au/samplesize
- http://www.openepi.com/Menu/OE_Menu.htm (Sample Size in the left-hand pane)

Note that some calculators ask for the estimated population variance of the outcome variable rather than the standard deviation - this is just the standard deviation squared, σ^2.

Acknowledgement: The author would like to thank Dr Indranil Saha of the Indian Council of Medical Research for his helpful comments on a draft of this article.

Reference

1. Chapter 5: Trial Size in Smith PG, Morrow RH, Ross DA (Ed.). Field Trials of Health Interventions: A Toolbox: 3rd ed. Oxford University Press, Oxford; 2015 https://doi.org/10.1093/med/9780198732860.001.0001