Education in Medicine Journal ISSN 2180-1932

EDUCATIONAL RESOURCE
Volume 5 Issue 2 2013
DOI: 10.5959/eimj.v5i2.130
www.eduimed.com

Introduction to sample size calculation

Wan Nor Arifin

Unit of Biostatistics and Research Methodology, School of Medical Sciences, Universiti Sains Malaysia.

ARTICLE INFO

Received	: 06/11/2012
Accepted	: 26/12/2012
Published	: 01/06/2012

KEYWORD

Sample size
Steps
Single mean
Single proportion

ABSTRACT

One of the most common reasons why researchers seek help from statistician is sample size calculation. However despite the common believe that it only involves formula and calculation, researchers often ignore other aspects of research design that leads to proper sample size calculation. In this article, the author outlines basic steps toward sample size calculation. The author also introduces the logic behind sample size calculation for single mean and single proportion in simplified and less intimidating forms to those not statistically inclined.

CORRESPONDING AUTHOR: Dr. Wan Nor Arifin, Unit of Biostatistics and Research Methodology, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia.
E-mail: wnarifin@kk.usm.my

Introduction

In my experience of statistical consultation, one of the most common reasons why researchers seek help from statistician is sample size calculation. However, one of the least common reasons is to discuss on the general planning and design of a research, including on what statistical analysis to use. Of note, by having discussion with statistician at initial stage of study would clarify many issues with regard to the general conduct of a research, more so the issues related to sample size calculation.

Sample size is very much related to other parts of a research (1) and it is not a stand-alone entity. As such, to handle the problem of sample size calculation, the other parts of a study should be taken into account. Despite misconception that the process only involves formula and calculation, researchers often ignore other aspects of a research that lead to proper sample size calculation. Most often, researchers come for consultation sessions with standard deviations or percentages from related journal articles and expect to calculate sample sizes for their studies with only that information. It is important to note that sample size calculation requires a number of preliminary steps with are related to general aspects of a research planning rather than plain formula and calculation.

In this article, I would suggest basic steps to obtain sample size. I do not include study design among them as the steps outlined are meant to be applied in general sense. In this introductory article also, I would go through sample size calculations for single mean and single proportion. I would show that sample size is not unreasonable and meant to complicate planning of a research and writing of a research proposal, but rather a logical and important part in quantitative research.

Basic steps toward sample size calculation

Step one: Objective

A clear (2) and achievable objective must be specified; most importantly it has to be quantitatively achievable. In quantitative research context, the outcome (dependent variable) and predictor (independent variable) that are stated in an objective can be measured or counted, not in form of abstract or subjective concepts. For example, the objective “To determine the mean systolic blood pressure among staffs in XYZ University” is clear and the outcome, “systolic blood pressure” is measurable (i.e. measured using sphygmomanometer in mmHg unit). Likewise, the objective “To determine the prevalence of HIV positive among drug addicts in XYZ district” is clear and the outcome is countable (i.e. the number of drug addicts and the number of HIV positive among them that would constitute the numerator and denominator of a prevalence are countable).

On the other hand, the objective “To look into the perception of medical personnel on the importance of sample size calculation”, although looks appealing, it is not clear as to how “perception”, a subjective concept, is measured. Restating the objective to “To determine the mean score of perception of medical personnel on the importance of sample size calculation using XYZ inventory” it is clear and quantifiable, as “perception” is measured by “XYZ inventory”.

As for the objective “To determine associated factors of smoking among school teenagers”, it is a well stated objective as the outcome is countable as we categorize the school teenagers as smokers or non-smokers (outcome) and count the number, given that the factors (predictors) are also quantifiable.

When an objective is stated in general form, for example “To determine the association between systolic blood pressure and demographic factors”, split the general objective into smaller specific objectives, such as “To determine the association between systolic blood pressure and age”, “To determine the association between systolic blood pressure and house income” and so on. Likewise, for objective “To determine the associated factors of smoking among school teenagers”, it has to be restated in form of specific objectives, such as “To determine the association between socioeconomic status and smoking status of teenagers”, “To determine the association between gender and smoking status of teenagers”, and so on. This process of restating the general objective into smaller specific objectives makes the objectives clear so as to facilitate sample size calculation (but it is not necessarily so in proposal).

Step two: Hypothesis testing or estimation

After clarifying the objective, be clear about whether the objective requires use of statistical test or just in form of descriptive statistics. In other words, either we are testing out hypothesis (using statistical test) or estimating (using confidence interval), which are two approaches of inferential statistics (3).

This dichotomy is reflected in the objective. For estimation, the objective would be stated in form of determination or measuring outcome of interest. As an example, for the objective “To determine the mean systolic blood pressure among staffs in XYZ university”, the outcome of interest is “systolic blood pressure”, and it aims to estimate the mean systolic blood pressure in the population. Notice the absence of predictor in the objective. It is also helpful to preview the way the result would be presented. For example, if a researcher wishes to know the prevalence of a particular disease in a population, it would be presented in form of percentage followed by respective confidence interval. Thus, it falls under the category of estimation.

For hypothesis testing, the objective would typically consist of outcome and predictor. The decision on suitable statistical test to test the hypothesis depends on the relationship of the outcome to its predictor, so it should be decided accordingly. For example, the objective is “To compare the means of systolic blood pressure between staffs in University A and University B. The outcome is “systolic blood pressure” and the independent variable is staffs' category (University A or B). Comparison between the staffs of the universities is done by comparing the means of systolic blood pressure. It is hypothesized that there are no difference between the populations and to test this hypothesis, it requires a suitable statistical test, which is independent t-test. Another indicator that the objective falls into the category of hypothesis testing is that in the presentation of the result, it would include p-value of respective statistical analysis. It should be stressed that for objective involving hypothesis testing, it is important to decide on statistical test to use as sample size calculation depends on the test used.

Step three: Sample size formula

After deciding whether the objective falls into estimation category or hypothesis testing category, we can decide on appropriate formula to use to calculate sample size.

For example, continuing from the previous example, a researcher wishes to know the prevalence of a particular disease in a population. He visualizes his result in form of ##.#% (95% confidence interval: ##.##%, ##.##%). In other word he want to estimate the prevalence of disease of interest in a population based on the sample he collected. There are a number of information that we can extract from the preceding sentences: 1. Prevalence is essentially proportion. 2. Estimate with 95% confidence. 3. Single sample proportion is involved. Even if you are not familiar with sample size calculation, by going through commonly used sample size formulas you would be able to guess that to calculate sample size for this objective, the most appropriate formula is single proportion formula. Following preceding two steps, it is easy to decide on which sample size formula is appropriate for the objective.

Next, a researcher wishes to compare the means of systolic blood pressure between population A and population B. He decided to use independent t-test to compare samples from these two populations. Essentially, he wishes to test his hypothesis that the populations are different (or not different) in term of means of systolic blood pressure. From the sentences, we can extract: 1. Two means are to be compared. 2. Two sample means are involved. 3. Hypothesis testing is involved. Again, it is clear that two means formula is appropriate to calculate the sample size for the objective.

After deciding on appropriate sample size formula to use, in both cases it is thus only a matter of deciding and finding applicable values to put in the formula.

In the next part, I would introduce reader to the basis of sample size formula for single mean and single proportion. It would be a good introduction toward understanding of how sample size is obtained and why it is important part of planning of a research.

Sample size calculation for estimation

Single mean

For single mean, the objective of a study is to estimate the mean of an outcome of interest in a population from data obtained from sample, of which the outcome is measured on numerical continuous scale.

For example, a researcher is interested to know the mean weight of young children aged 10 to 12 years old in Malaysia. He wishes to estimate the mean weight of the population by taking a sample representative of the population with 95% confidence. He previews that the result would be presented in form of xx.x (95% confidence interval: xx.xx, xx.xx). He wants to know the sample size that he should take to achieve his study objective.

Let say based on a hypothetical literature search, in one Asian country, it was estimated that the mean weight among children aged 10 to 12 years old was 20.0 kg (95% CI: 19.75, 20.25), based on a sample of 250 children. The standard deviation of the weight was 2.00 kg. Looking at the result also it was precise to plus and minus 0.25 kg.

Let us trace back the result to its basic formula used to obtain the 95% confidence interval and also the precision. Basically, a confidence interval consists of lower confidence limit and upper confidence limit, in our example the limits were 19.75 and 20.25 respectively. The lower confidence limit was obtained by subtracting the precision from the mean, in which 20.0 kg minus 0.25 kg equals 19.75 kg. On the other hand, the upper confidence limit was obtained by adding the precision to the mean, in which 20.0 kg plus 0.25 kg equals 20.25 kg. As such, confidence interval formula for mean is given by,

$$\text{mean} \pm \text{precision}$$

as applied to our example,

$$20.0 \pm 0.25$$

$$19.75, 20.25$$

So, where are the component of 95% confidence and the standard deviation? Precision can be further deconstructed into (3),

$$\text{precision} = \text{reliability coefficient} \times \text{standard error}$$

Reliability coefficient value depends on our preset level of confidence, for example the reliability coefficient for 95% confidence is 1.96. The reliability coefficients for other confidence level are 1.645 for 90% confidence and 2.58 for 99% confidence [A].

As for the standard error, you can view it as adjustment for standard deviation when we are dealing with sample instead of population. To be exact, it is the standard deviation of sampling distribution. Interested reader can read further under sampling distribution topic in statistics textbooks, for example (3). Standard error consists of,

$$\frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

with the sample standard deviation and its size used in the formula.

So, from our hypothetical literature search, the mean weight was 20.0 kg, the standard deviation was 2.00 kg and the sample size was 250. So, putting everything together with 95% confidence,

$$\text{mean} \pm \text{reliability coefficient} \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

$$20.0 \pm 1.96 \times \frac{2.00}{\sqrt{250}}$$

$$20.0 \pm 0.25$$

$$19.75, 20.25$$

or presented in form of 20.0 kg (95% CI: 19.75, 20.25), which we already encountered in our hypothetical literature search before.

[A] The reliability coefficient is the corresponding *z*-value of standard normal distribution at a particular probability value. For example, given 95% confidence level, it corresponds to covering 95% area of cumulative probability distribution of standard normal distribution, leaving only 5% area for our lack of confidence (usually denoted as α, typically known as type I error or significance level). As we want to divide our uncertainty into two parts (lower and upper limits), as such we allocate 2.5% of the area at the lower region of the distribution (left most) and also 2.5% of the area at the upper region of the distribution (right most) so that our confidence area lies in the middle. So, please look up in any standard normal distribution table, you will find *z*-value of –1.96 corresponds to 0.025 cumulative probability (area), and *z*-value of 1.96 corresponds to 0.975 cumulative probability. As the z-values are only different in direction, it is easier for us to just find the z-value for the upper limit, usually written as $z_{(1-\alpha/2)}$, and also because our confidence interval formula already accommodate that.

After all, what does this calculation of confidence interval mean to us? In the calculation, we did not pre-specify the precision (0.25), but instead we just put in the value of the respective components of precision obtained from literature. So, what if we pre-specify the precision that we deem acceptable and we would like to know the sample size required to achieve that level of precision? By algebraic manipulation we obtain,

$$\text{precision} = \text{reliability coefficient} \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

$$\sqrt{\text{sample size}} = \frac{\text{reliability coefficient} \times \text{standard deviation}}{\text{precision}}$$

$$\text{sample size} = \frac{\text{reliability coefficient}^2 \times \text{standard deviation}^2}{\text{precision}^2}$$

To recapitulate objective put forward by a researcher in our example, he wishes to estimate mean weight of young children aged 10 to 12 years old from a representative sample with 95% confidence level. Additionally he wishes to estimate with precision of 0.5 kg, which he deems acceptable. From literature, it was found that the standard deviation of weight in that age group was 2.00 kg. The sample size to achieve his objective is,

$$\text{sample size} = \frac{1.96^2 \times 2.00^2}{0.5^2} = 61.47 \approx 62 \text{ children}$$

We often round up the sample size when dealing with human being, as we cannot simply sample only part of a person just to be precise with our calculated sample size. We may also add additional subjects to the calculated sample size to accommodate for possible dropouts,

$$\frac{\text{sample size} + \text{dropouts}}{\text{calculated sample size}} = \frac{100\% \text{ subjects}}{\% \text{ subjects} - \text{dropouts}}$$

$$\frac{\text{sample size} + \text{dropouts}}{\text{calculated sample size}} = \frac{1}{(1 - \text{proportion of dropouts})}$$

$$\text{sample size} + \text{dropouts} = \frac{\text{calculated sample size}}{(1 - \text{proportion of dropouts})}$$

with, let say with 10% drop out rate,

$$\text{sample size} + \text{dropouts} = \frac{62}{(1 - 0.1)} \approx 69 \text{ children}$$

thus the required sample size for his study is 69 children after accommodating for 10% drop out rate.

Single proportion

For single proportion, the objective of a study is to estimate the proportion or percentage of an outcome of interest in a population from data obtained from sample, of which the outcome consists of two categories (dichotomous).

For example, a researcher is interested to know the percentage (or prevalence) of obesity among young children aged 10 to 12 years old in Malaysia. He wishes to estimate the percentage of obesity in the population by taking a sample representative of the population with 95% confidence. He previews that the result would be presented in form of percentage: xx.x% (95% confidence interval: xx.xx%, xx.xx%). He wants to know the sample size that he should take to achieve his study objective.

Again, let say based on a hypothetical literature search, in one nearby Asian country, it was found that percentage of obesity among children aged 10 to 12 years old was 30.0% (95% CI: 25.00%, 35.00%), based on a sample of 320 children. Note that the result was precise to 5% (or 0.05 in form of proportion).

Before going into the sample size calculation, we need to go through the detail of confidence interval calculation for the proportion given in by the literature. Confidence interval formula for proportion is given by,

$$\text{proportion of a factor} \pm \text{precision}$$

or simply,

$$\text{proportion} \pm \text{precision}$$

in our example,

$$0.3 \pm 0.05$$

$$0.25, 0.35 \quad \text{or} \quad 25.00\%, 35.00\%$$

As for the precision, it is given by,

$$\text{precision} = \text{reliability coefficient} \times \text{standard error}$$

which is similar to the precision for mean in term of basic formula. For the reliability coefficient, we still use the same *z*-value that corresponds to our confidence level. However, the standard error part needs minor changes to the formula. In general, standard error is given by,

$$\frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

which looks similar to that of single mean. But notice that for proportion, we are not presented with standard deviation in literature. It does not mean that there is no standard deviation for proportion, but because it is not commonly presented in article. Standard deviation of proportion can be easily obtained by,

$$\sqrt{\text{proportion with outcome} \times \text{proportion without outcome}}$$

in other words,

$$\sqrt{\text{proportion with outcome} \times (1 - \text{proportion with outcome})}$$

By putting our standard deviation for proportion into our standard error formula, it becomes,

$$\frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

$$= \sqrt{\frac{\text{proportion with outcome} \times (\text{proportion without outcome})}{\text{sample size}}}$$

$$= \frac{\sqrt{\text{proportion with outcome} \times (1 - \text{proportion with outcome})}}{\sqrt{\text{sample size}}}$$

or simply,

$$= \sqrt{\frac{\text{proportion} \times (1 - \text{proportion})}{\text{sample size}}}$$

Thus, reconstructing the confidence interval given the literature,

$$\text{proportion} \pm \text{reliability coefficient} \times \sqrt{\frac{\text{proportion} \times (1 - \text{proportion})}{\text{sample size}}}$$

$$0.3 \pm 1.96 \times \sqrt{\frac{0.3 \times 0.7}{320}}$$

$$0.3 \pm 0.05$$

$$0.25, 0.35$$

Having understood the process of calculating confidence interval, similar to what we did for precision formula of single mean, by algebraic manipulation we derive the sample size for single proportion,

$$\text{precision} = \text{reliability coefficient} \times \sqrt{\frac{\text{proportion} \times (1 - \text{proportion})}{\text{sample size}}}$$

$$\text{precision} = \frac{\text{reliability coefficient} \times \sqrt{\text{proportion} \times (1 - \text{proportion})}}{\sqrt{\text{sample size}}}$$

$$\sqrt{\text{sample size}} = \frac{\text{reliability coefficient} \times \sqrt{\text{proportion} \times (1 - \text{proportion})}}{\text{precision}}$$

$$\text{sample size} = \frac{\text{reliability coefficient}^2 \times \text{proportion} \times (1 - \text{proportion})}{\text{precision}^2}$$

Recall the objective of our researcher, in which he wishes to estimate percentage of obesity among young children aged 10 to 12 years old in Malaysia from a representative sample with 95% confidence level and precision of 1%. From literature, it was found that the prevalence was 30.0%. The sample size to achieve his objective is,

$$\text{sample size} = \frac{1.96^2 \times 0.3 \times 0.7}{0.01^2} = 8067.36 \approx 8068 \text{ children}$$

As you can see, with very small precision, the sample size is inflated to 8078 children as compared to the study from literature with sample size of only 320 children. The researcher may need to reduce the precision to, for example 2% or 3% if he feels that it is impossible for him to collect a sample that large, or possibly due to budget constrain or other considerations pertaining to conduct of research. After deciding with an optimal sample size, he can inflate the sample size further to adjust for expected drop out rate.

Conclusion

In this short article, we have gone through the basic steps of sample size calculation. We also have gone through the basis of sample size formula to estimate true values (parameters) of a population, specifically single mean and single proportion formulas. I intentionally show the derivation of sample size formulas so that you can appreciate the reason why sample size is so important in planning of a research and it is not meant to complicate the process of conducting a research. I did not cover sample size calculation for two means and two proportions in this article as it is more suitable to be discussed in another article which would follow this introductory note on sample size. Throughout this article, formulas are written in full sentences or words instead of using Greek's alphabet or symbols or letters to foster the understanding of the formulas and to avoid unnecessary fear of statistical notations commonly encountered while reading statistics textbooks. The formulas in their commonly used forms are included in Appendix for those who are statistically inclined.

Conflict of interest

None to be declared.

Reference

1. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials. 1981;2(2): 93-113.
2. Lwanga S, Lemeshow S. Sample size determination in health studies: a practical manual. England: World Health Organization; 1991.
3. Daniel WW. Biostatistics: A foundation for analysis in the health sciences. 6th ed. USA: John Wiley & Sons. Inc; 1995.

Further reading

1. Machin D, Campbell MJ, Beng TS, Tan SH. Sample size tables for clinical studies. Singapore: Wiley-Blackwell; 2009.
2. Naing N. A practical guide on the determination of sample size in health sciences research. Kota Bharu, Malaysia: Pustaka Aman Press; 2010.

Appendix

Confidence interval for single mean:

$$\bar{x} \pm d$$
$$\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$$
$$\bar{x} \pm z_{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

Precision for single mean:

$$d = z_{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

Single mean formula:

$$n = \frac{z^2_{(1-\alpha/2)}\sigma^2}{d^2}$$

Confidence interval for single proportion:

$$\bar{x} \pm d$$
$$\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\hat{p}}$$
$$\bar{x} \pm z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}}$$

Precision for single proportion:

$$d = z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}}$$

Single proportion formula:

$$n = \frac{z^2_{(1-\alpha/2)}p(1-p)}{d^2}$$

Symbols:

$\bar{x}$ – mean
d – precision
$z_{(1-\alpha/2)}$ – reliability coefficient
σ – standard deviation
p – proportion
$\sigma_{\bar{x}}$ – standard error (single mean)
$\sigma_{\hat{p}}$ – standard error (single proportion)
n – sample size