The THAI Journal of SURGERY

Official Publication of the Royal College of Surgeons of Thailand

Vol. 33 April - June 2012 No. 2

Review Article

Principles of Statistics for Surgeons III: Factors Influencing Sample Size Estimation

Panuwat Lertsithichai, MD., MSc. (Medical Statistics)

Department of Surgery, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

Abstract **The present article, the third in a series, describes some important ideas underlying the practice of sample size estimation. These include types I and II errors, the effect size, and data variability. Each of these ideas is described and illustrated in some detail without too much technical development, and should be relatively easy to understand for most surgeons. In particular, important concepts in elementary statistics are highlighted and should be useful for self-study. The necessity of some prior context-specific knowledge, when estimating the sample size, is emphasized. Simple numerical examples are presented at the end of the article to strengthen the emphasis. We hope the reader will be sufficiently informed in preparation for consultations with statisticians when such needs arise.**

Key words: **sample size; estimation; factors**

Introduction

In the present article, some theoretical factors that can influence sample size estimation will be discussed. At the end, we will provide some simple numerical examples.

What is the justification for sample size estimation? There are four important points to consider. If the sample size is too small, the study will lack power to detect important differences or other associations, and the study is unscientific. If, however, the study is too large, then we are placing more research volunteers at risk whether from withholding treatment or from harm due to new interventions, and the study is unethical. If the study is too large it will also be a waste

Correspondence address: *Panuwat Lertsithichai, MD, MSc. (Medical Statistics), Department of Surgery, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; Telephone:+66 2201 1315; Fax:+66 2201 1316; E-mail: raplt@mahidol.ac.th*

of resources and hence uneconomical. Finally, from a practical point of view, a study cannot be too large because of time limitations and constraints on available resources. This latter is especially pertinent for the student.

A sample size estimation procedure must take into account all these considerations. The resulting sample size estimate is therefore a compromise of conflicting requirements. It is fair to say that sample size estimation begins with scientific idealism and ends with everyday pragmatism.

A recent article says much about the current skepticism in the practice of sample size estimation.[1] For example, "Sample size calculations are frequently based on inaccurate assumptions for the control group, calculations are often erroneous, and the hypothesized treatment effect is often fixed a posteriori." In addition, "the current calculation of sample size is actually mainly driven by feasibility." Recent emphasis on systematic reviews are also making sample size estimates for single trials more or less irrelevant, and some Bayesian approaches to data analysis as well as some adaptive trial designs do not require a priori fixed sample size.[2]

So, why do we bother with sample size calculations? Perhaps mainly for educational value and to provide insight into some important (frequentist) ideas in statistics. However, in clinical trials where ethics and economics are of real concern, sample size estimation is still commonly done, or is still considered essential. For simplicity, let us limit the present discussion to the case of two-arm comparative trials. Also, further assume that we are interested in differences in outcomes between the two arms. For example, the difference might be in terms of some laboratory values, such as changes in bone mineral density before and after some treatment, or in terms of pain score. The two arms might correspond to two treatments whose effects include bone density alteration or changes in pain level. Each treatment arm will have the same number of research volunteers, there will be no interim analyses, and all analyses will be firmly within the "frequentist" framework.

The sample size question can be stated essentially as, "What is the smallest sample size such that, if a true difference exists, the study can detect or demonstrate such a difference?" Of course, the question as stated might not have a definitive answer, but it captures the essential idea. The idea is similar to that of the sensitivity

"2×2 Table for Research Studies"

	Conclude: Different	Conclude: Not different
No True difference	Type I error (α)	"Specificity"
True difference	Power = 1 – β "Sensitivity"	Type II error (β)

Figure 1 The definitions of types I and II errors.

of a diagnostic test used to detect a disease given that it exists.

There are three factors influencing sample size estimation. The first is the false negative and false positive detection rates that the researcher is willing to accept. The second factor is the expected size of the treatment difference, generally called the "effect size". The third factor is the variability in the data, due to both systematic and random sources.[3]

1. Types I and II errors

The false negative and false positive detection rates can be shown in a familiar 2×2 table (Figure 1). The columns denote the conclusions of the study, similar to the results of a diagnostic test. The rows denote the existence of a true difference between the treatments, similar to the disease status in diagnostic testing. A false positive detection is the conclusion that there is a difference between treatments even though there is none, similar to the conclusion that a patient has a disease even if he or she does not. In research terminology, this is called the "type I (alpha) error", as shown in the cell so marked in the figure. Conversely, a false negative detection is the conclusion that there is no difference even though there is. This is called the "type II (beta) error". Ideally, the research study should have no chance of false negative and false positive conclusions. But, as in diagnostic testing and in the real world, this is impossible.

The reason why types I and II errors should affect the sample size is as follows. Consider a clinical trial comparing 2 operations in terms of post operative pain, measured on a visual analog scale (VAS) from 0 to 10. We wish to test the hypothesis that the 2 treatments are different, that is, one treatment is more painful than the other on average. Let us call the hypothesis

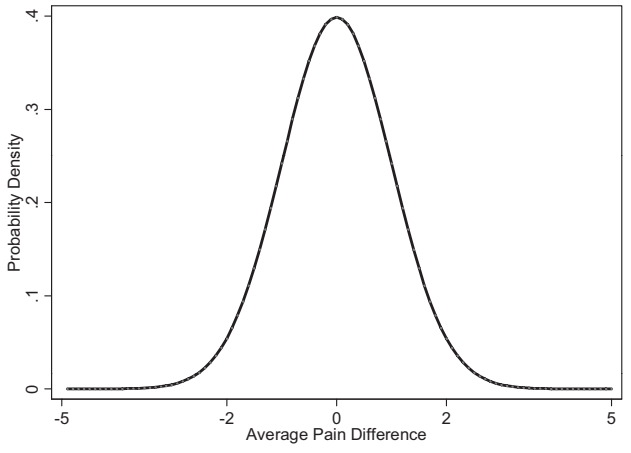

Figure 2 The sampling distribution of the average pain difference under the Null Hypothesis

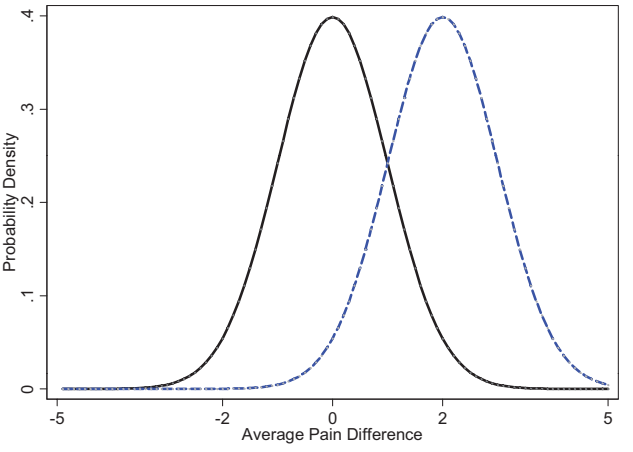

Figure 3 Sampling distributions of the Null (solid line) and Alternative (dashed line) Hypotheses.

that there is an average difference the "alternative hypothesis", and the converse hypothesis that there is no average difference the "null hypothesis". Then a type I error occurs when we conclude that there is a difference even though the null hypothesis is true. Conversely, a type II error occurs when we conclude that there is no difference, even though the alternative hypothesis is true.

Suppose we wish to perform a study with 50 patients per arm, or 100 patients overall. After completion of the study, we calculate the outcome: the average or the mean pain difference. Suppose also that a clinically meaningful pain difference is 2. If we perform such studies - with 50 patients per arm drawn randomly from a defined population - many, many times, ideally an infinite number of times, we can plot a histogram, or a distribution curve, of the average pain difference. This distribution curve is called the "distribution curve of the sample mean". We can also call it the "sampling distribution of the mean", or "sampling distribution" for short. It is crucial to understand that this distribution curve is not the distribution curve of the data. The sampling distribution is, in a sense, a theoretical construct.

The sampling distribution of the mean is approximately Normal. This is true even if the distribution of the data, i.e., of the pain score, is not Normal, given that the number of patients in each study is large enough[4]. In our example, 50 patients per arm are sufficient. If the null hypothesis is true, then the sampling distribution curve will have a mean of zero, and the standard deviation will have a value depending on the sample size and the variability of the data. The standard deviation of the distribution of the sample mean will decrease as the sample size increases. This is in contrast to the standard deviation of the data, which does not decrease with increasing sample size, but remains essentially constant. Thus, suppose the sampling distribution for the null hypothesis is as presented in Figure 2. The curve is symmetric and bell-shaped, with an average value of zero and a standard deviation of 1 (the exact value of the standard deviation is irrelevant to the argument). We have located the clinically meaningful pain difference on either side of the curve with two ticks on the horizontal (or "x") axis, at minus 2 and plus 2.

The sampling distribution for one alternative hypothesis can be represented by a curve centering on the pain difference of 2, because this is the smallest clinically important difference. The sampling distribution for this alternative hypothesis will have an average of 2 and a standard deviation identical to that of the curve for the null hypothesis. The standard deviation of the sampling distribution is usually called the standard error, to distinguish it from the standard deviation of the data.

An aside: a simple intuitive explanation for the decreasing standard error with increasing sample size is as follows. Imagine a population from which we randomly choose a sample. Suppose we wish to estimate the average height of the population using the sample mean. Obviously, the sample mean is unlikely to be the

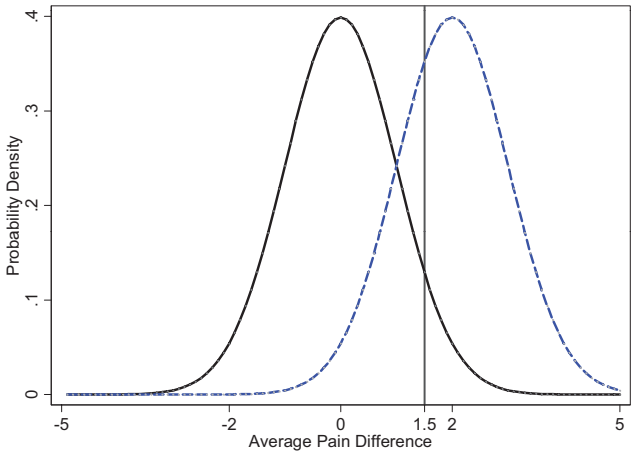

Figure 4 Showing that a particular study with an average pain difference of 1.5 units is compatible with both Hypotheses.

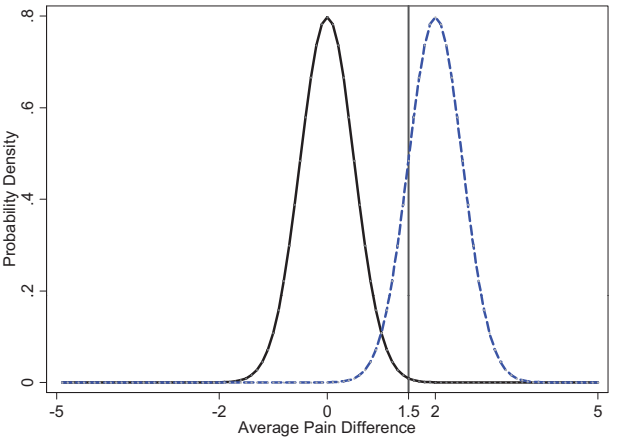

Figure 5 Showing that an average pain difference of 1.5 units is more compatible with the Alternative hypothesis (dashed line) in this figure.

same as that of the population, but it could be close. The sample mean can vary according to sampling variation (i.e., by chance). But with increasing sample size, the likelihood that the sample mean will approach the true population mean will also increase. The limit is reached when the sample is the same as the population: the sample mean is then identical to the population mean. At the limit, the standard error of the mean is zero. Therefore, the standard error of the mean must decrease as the sample size increases.

We can present the sampling distributions for both the null and alternative hypotheses in one graph as in Figure 3. With a sample size of 50 per arm, there is a large "overlap" (in the values of the average pain difference) between the two sampling distributions. This overlap implies that if the null hypothesis were true, there is a considerable chance that a study with 50 patients per arm might produce an average pain difference falling within the range of values attributable to the alternative hypothesis. Hence, the researcher might conclude that there is a real difference, even though there is none, which is a type I error.

Conversely, with such a large overlap, if the alternative hypothesis were true a study with 50 patients per arm might produce an average pain difference falling within the region attributable to the null hypothesis. The researcher might then falsely conclude that there is no real difference, even though there is. This is a type II error.

As an example, suppose the alternative hypothesis is true, that is, there is a true average difference in the pain score of 2 units. Also, suppose that in one particular study, with 50 patients per arm, the average pain difference is 1.5. As can be seen from Figure 4, the average pain difference of 1.5 is compatible with both the null and alternative hypotheses. Now, if a statistical test based on the null hypothesis were to be performed, and setting the significance level at 5%, the difference of 1.5 is not statistically significant (to be significant the difference must be > 1.96). That is, we will not reject the null hypothesis ("no real difference"), and we will be committing a type II error.

To reduce the possibility of both types of error, we must reduce the overlapping of the sampling distributions. To do so, we must reduce the standard errors of the sampling distributions, to make the distributions more "slim" or more "tight". This is achieved by increasing the sample size. As shown in Figure 5, with a much larger sample size of 200 per arm or 400 overall, the two sampling distributions have much less overlap, and hence a smaller chance of types I & II errors. Referring to our previous example, a difference of 1.5 is now clearly not compatible with the null hypothesis at a significance level of 5%, and we will correctly conclude that the difference is significant and we should accept the alternative hypothesis. We will thus no longer commit a type II error.

Similarly, a study will require even larger sample sizes if one demands the types I and II errors to have lower values. Unfortunately, the increase in sample size is out of proportion to the reduction in types I & II errors (see the sample size formula, below). In

theory, one way to reduce types I & II errors to zero is to make the sampling distributions so slim that they become concentrated at only one point. This implies an impossibly large sample size, that is, a size of infinity. Since limiting either type I or type II error or both to zero is equivalent to setting the sample size to infinity, by convention the type I error (or significance level) is fixed at 5% (a false positive detection rate of 0.05), and the type II error is usually fixed at 10% or 20%. In this way, the calculated sample size will be finite. Therefore, how one sets types I & II errors will have a large impact on the sample size of the research study.

We need to define one more word: power. The "sensitivity" of the research study is (1 - type II error), as in the sensitivity of a diagnostic test. It is the probability of detecting a difference if there truly is a difference. In research terminology, this is also called the "power" of a research study. Thus, if the type II error is set at 20% or 0.2, then the power or sensitivity of the study is 80% or 0.8.

2. The effect size

Intuitively, it is clear that the larger the true difference between two treatments, the smaller is the sample size needed to detect it. Imagine if the difference between treatment outcomes is that between life and death. If a previous treatment has failed to treat a condition and all patients died, and if a new treatment cured all patients and all survived, then only a few patients are needed to convince anyone of the efficacy of the new treatment. On the other hand, if the old treatment is associated with, for example, 60% survival and the new with 70% survival, many more patients will be needed to show or convince others of a difference in treatment outcomes.

In a similar way, to detect a true average difference of 1 unit in pain score / bone density change between two treatments, a larger sample size is required compared with that needed to detect a true difference of, say, 2 units, all else being equal.

3. Data variability

The final factor influencing sample size estimation is data variability. In fact, it is difficult to separate this factor from the effect size or treatment difference. Some statisticians combine both factors into a measure known as the "standardized effect size". For educational (pedagogical) purposes, however, it is better to separate them.

Intuitively, it is clear that if there is little variability in the data, for example, if one treatment always cures patients and another always fails, then only a few patients are needed to detect this difference. Similarly, if a treatment always produces a pain level of 6 units and another always 4 units, then only a few patients are needed to establish that the two treatments are different in this respect. Even the magnitude of the difference, i.e. 2 units, can be reliably inferred from only a few patients.

However, with large variability in the data, the differences can be overwhelmed. That is, "the noisiness will drown the signal". With a large variability, even if the true average pain score is 6 units for one treatment and 4 units for another, the overlap in the range of pain values for the two treatments will be so great that it might be difficult to discern any differences unless the sample size is large enough. (Note that the issue of data variability is not directly related to the issue of types I & II errors, even if there are some similarities in the reasoning. One difference is the use of sampling distributions for latter, and the use of data and population distributions for the former.)

To illustrate graphically, consider Figure 6. The outcomes of two treatments, A and B, are pain scores, assumed to have Normal distributions with identically large standard deviations but different means, i.e., 4 and 6 units, respectively. There is a true average treatment difference of 2 units, and the two distributions

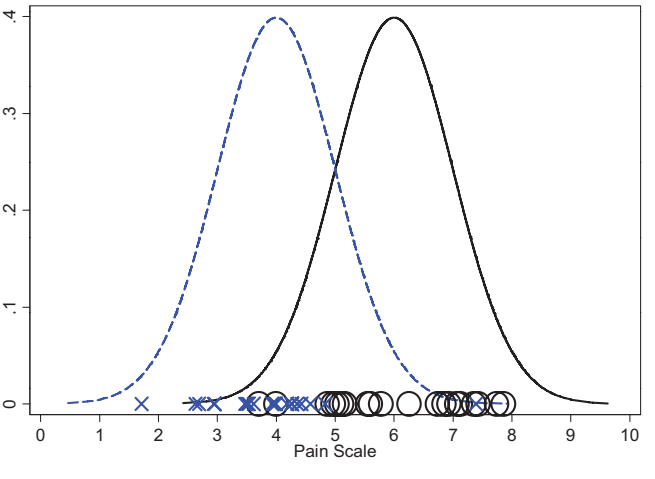

Figure 6 Population distributions of pain scale for treatments A (dashed line) and B (solid line), with a random sample of 20 patients from A (crosses) and 20 from B (circles) shown on the y = 0 line.

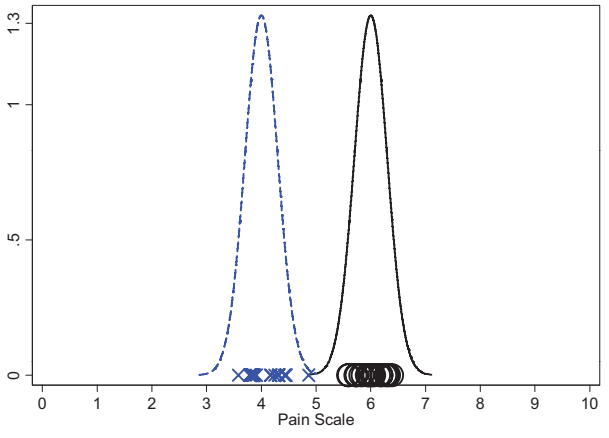

Figure 7 Population distributions of pain scale for treatments A and B as in figure 6, but with much smaller standard deviation. The random samples of 20 each have no overlapping values.

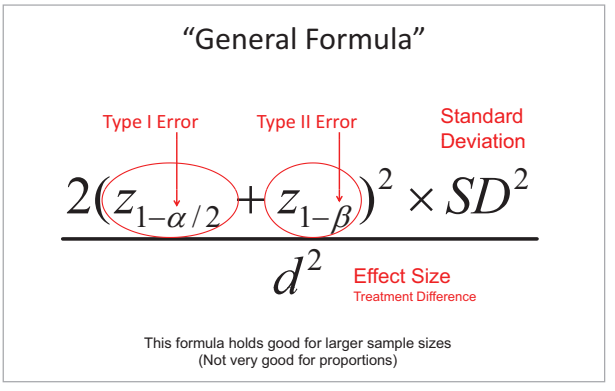

Figure 8 A general or representative formula for the sample size per group when there are two comparative groups and equal number of subjects per group.

have a large area of overlap. Suppose we conduct a study in which we randomly allocate 20 research volunteers to treatment A, and 20 to treatment B. The results of the study must be equivalent to 20 random draws from distribution A, and 20 random draws from distribution B.

By projecting the results onto the line representing pain scale, with 20 volunteers per arm (the circles and crosses on line y = 0 in Figure 6), the outcomes for groups A and B overlap each other so much that it might be difficult to see any difference between them. Therefore, despite a real average difference of 2 units, a study with 20 volunteers per arm is unlikely to be sensitive enough to detect that difference. In other words, the large standard deviation, that is, the large variability in the data relative to the true difference, is such that a study with 20 volunteers per arm is too small to detect a true difference.

Next, consider Figure 7. In an otherwise identical study, suppose that the data variability some how has been considerably reduced. There is practically no overlap between the two distributions. Drawing 20 samples from distribution A and 20 from distribution B, we can project the results onto the line representing pain scores.

This time, there is no overlap between the results of the two treatments. Even with only 20 volunteers per group (or even less), it might be easy to conclude that there is a difference. Hence, with a smaller standard deviation relative to the true difference, a study with a given sample size has more power to detect the true difference than if the standard deviation were larger. By logical inference, a study with smaller data variability, relative to a given true difference, requires a smaller sample size than a study with larger data variability, to detect that difference with the same power.

To repeat once again: a study with large variability in the outcomes for both treatments, in which there is a large overlap between outcome distributions, should require a larger sample size to detect a given difference than a study with smaller variability, all else being equal.

4. A Sample size estimation formula

A representative formula[5] for calculating the sample size per group in the case of two treatments with equal number of subjects per group is presented in Figure 8. Depending on the outcome measure, this formula can give good estimates of the sample size. All the relevant factors influencing sample size estimation are represented here. Fortunately, there is no need to memorize this formula since most statistical software can perform all the relevant sample size calculations (using more appropriate formulae).

If the outcome measure is a continuous variable, then the treatment effect or treatment difference is given by the mean difference. The standard deviation of the combined data is derived ("pooled") from the standard deviations associated with both treatment groups.

If the outcome is a binary variable, then the treatment difference is given by the difference between two proportions from the two treatment groups. The

"standard deviation" of the pooled data can be calculated from the two proportions.

We must emphasize that prior knowledge of the effectiveness of the treatments to be compared is necessary for a reasonably accurate sample size estimation. This is clear from the sample size estimation formula. Unfortunately, this information often does not exist, or is unreliable. If everything were perfectly known, there would be no need for research studies.

This is the "circularity problem" of sample size estimation, one that has not been satisfactorily resolved. We need to calculate the sample size for a study of new interventions but we need studies of new interventions to calculate the sample size! In order to break the vicious cycle or resolve the paradox, either some preliminary data must be used or an educated guess based on some expected chance of success must be invoked.

5. Examples

We now provide some examples. The purpose of these examples is to show explicitly the "prior" data needed for a sample size calculation. *Example 1* (Figure 9). The present example has a continuous outcome. Hence, the researcher must attempt to find the most reliable "plug-in" estimates of the relevant means and standard deviations. The problem is where to find these estimates. If the control treatment is well known and commonly practiced, then such estimates should be readily available. But the estimates for the newer treatment might be more difficult to find. In such a situation, either use any available estimate or conduct a preliminary pilot study to determine the approximate outcome measures. Or, as a last resort, make a conservative guess of the effect of the new intervention, either in absolute terms or relative to the control intervention.

Example 2 (Figure 10). For binary outcomes, it is easier to make a guess as to the likely proportion of outcomes. If the estimate of the proportion is extremely uncertain, a range of sample sizes can be calculated, spanning the entire range of uncertainty. One then chooses the largest sample size feasible under practical constraints. Interestingly, binary outcomes are often associated with larger sample sizes, as illustrated in this example. This is so despite the proportion of outcomes of the study intervention (30%) being twice that of the control intervention (15%). If a smaller sample size is required, and if both continuous and binary outcomes are equally relevant for a particular study, we could use continuous outcomes in the estimation of the sample size.

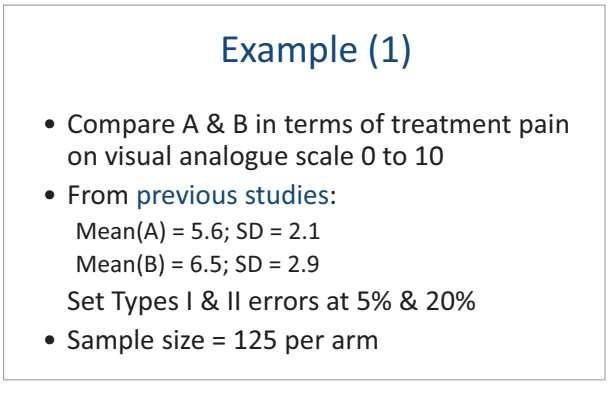

Figure 9 Example 1 - sample size calculation with a continuous or quantitative outcome.

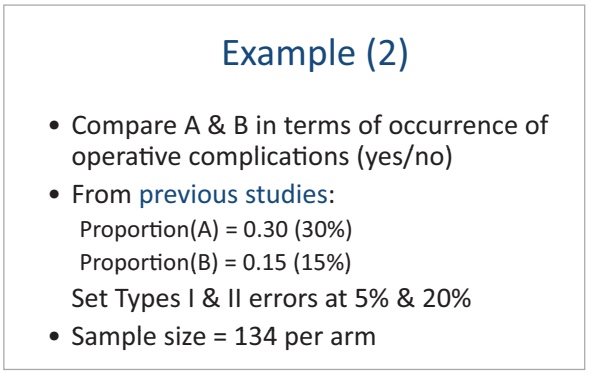

Figure 10 Example 2 - sample size calculation with a binary outcome.

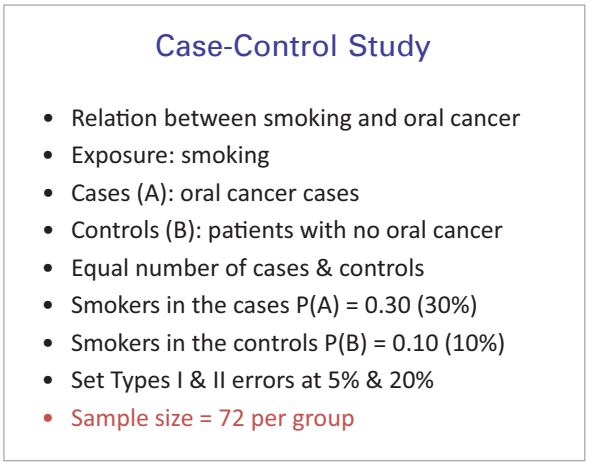

Figure 11 Sample size calculation for a case-control study.

Example 3 (Figure 11): *a case-control study*. In essence, the sample size calculation is no different from that of example 2. But instead of using the occurrence of disease as the "outcome" in the calculation, the risk factor (smoking) is playing that role.

More complicated situations are possible. Sample size estimation might be needed for a study with more than two groups, for a repeated measures study, for a time-to-event (survival analysis) study, for a non-superiority trial, for a diagnostic study with the Receiver Operating Characteristic Curve (ROC) as the outcome, or for agreement studies, and so on.[5] Should the student become involved in any of these interesting situations, discussion with a knowledgeable statistician is recommended.

Conclusion

We have presented a simplified discussion of important theoretical factors influencing sample size estimation. These included types I and II errors, the size of the effect difference, and data variability. Along the way, we have shown how these factors relate to some interesting statistical ideas. We emphasized the need for prior knowledge, that is, some prior data, if a reasonably accurate sample size estimate is to be expected. This article should provide some background for the student when consulting a statistician for help in calculating the sample size needed for his or her research study.

REFERENCES

1. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting sample size calculation in randomised controlled trials: review. Br Med J 2009;338;b1732;doi:10.1136/bmj.b1732.
2. Spiegelhalter DJ, Abrams KR, Myles LP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: John Wiley & Sons, Ltd; 2004. p. 181-249.
3. Lertsithichai P. Variability and randomness in medical research. Thai J Surg 2010;31:1-6.
4. Casella G, Berger RL. Statistical inference. 2nd ed. California: Duxbury Press; 2002. p. 236-9.
5. Machin D, Campbell MJ. Design of studies for medical research. Chichester: John Wiley & Sons, Ltd; 2005.