Review Article

Sample Size Calculation for Epidemiologic Studies: Principles and Methods

Farin Kamangar MD PhD•[1,2] , Farhad Islami MD PhD[3,2]

Abstract

This paper discusses the statistical principles, methods, and software programs used to calculate sample size. In addition, it reviews the practical challenges faced in calculating sample size. We show that because of such challenges, statistical calculations often do not provide us with a clear-cut number for the study sample size; rather they suggest a range of reasonable numbers. The paper also discusses several important nonstatistical considerations in determination of sample size, such as novelty of the study and availability of resources.

Keywords: Power, sample size, type I error, type II error

Cite this article as: Kamangar F , Islami F. Sample size calculation for epidemiologic studies: Principles and methods. *Arch Iran Med*. 2013; **16(5):** 295 – 300.

Introduction

"How many is enough?" is a question that epidemiologists and clinicians ask themselves when they plan on conducting a new study. Researchers want to enroll a large enough number of people such that statistical errors (type I and type II) are minimized yet the cost, labor, and time to do the study remain acceptable. Sample size calculations often remind us of complex formulas. While we provide some formulas in the text, the main aim of the article is not to offer a long list of such formulas. Rather, the main aim is to discuss the statistical principles behind sample size calculation, issues that may make such calculations not-so-straightforward, and nonstatistical considerations in sample size determination. Therefore, in this article, we discuss the following topics:

1. The need to calculate sample size;
2. Principles;
3. Some formulas;
4. Factors that need to be determined for sample size calculations;
5. Assumptions made for sample size calculations;
6. Nonstatistical considerations;
7. Methods used to calculate sample size; and
8. Software used to calculate sample size.

The final part of the paper, Summary and Conclusions, ties these sections together.

1- The need to calculate sample size

When we would like to learn about an attribute of a population, such as mean cholesterol of the people of China, it may not be feasible for us to study the entire population due to cost or time issues. Besides cost and time issues, it may not be ethical to study the entire population if accurate enough results could be obtained by studying a subgroup of all people. For these reasons, we need to study a sample of the population.

However, results vary from sample to sample, and they may be somewhat different from the true mean of the population. For example, one sample of 100 Chinese people may have a mean cholesterol of 182 mg/dL and another sample may have a mean of 186 mg/dL. Nevertheless, as the sample size grows, the probability of obtaining a result that is close to the true mean for the population increases. The question is how large the sample size should be to make it very likely for the sample results to be within a narrow distance from the true mean. The answer is discussed in the next few sections. First, we start with some general principles, and then we go into more details.

2- Principles

Although sample size depends on many factors, there are certain principles that apply to nearly all sample size calculations, which we discuss in this section. Sample size nearly always depends on the factors discussed below.

2-1- Variation

The more variation there is in the variable of interest, the larger is the required sample size. If there is no variation, even a sample size of n = 1 is adequate.

Example 1: We want to determine the mean salary of all interns in a hospital. If all interns receive exactly the same salary (e.g., $40,000 annually), knowing the salary of only one intern is adequate to know the mean. ●

Example 2: There is a disease that is universally fatal, which is equivalent to saying that there is no variation in its outcomes in terms of death and life. Now if a new drug cures only one case of this disease, assuming that the diagnosis is correct, that one single case is enough to accept that the drug works. ●

2-2- Magnitude of error that we accept

The less the magnitude of the error that we accept, the larger is the needed sample size. This is somewhat intuitive: larger sample size is the price that we pay for less error.

Example 3: A researcher wants to determine the mean height of a population. Any sample would most likely estimate the mean

Authors' affiliations: [1]School of Community Health and Policy, Morgan State University, Baltimore, USA, [2]Digestive Disease Research Center, Tehran University of Medical Sciences, Tehran, Iran, [3]Institute for Translational Epidemiology, Mount Sinai School of Medicine, New York, NY, USA.
•**Corresponding author and reprints:** Farin Kamangar MD PhD, Department of Public Health Analysis, School of Community Health and Policy, Morgan State University, Portage Avenue Campus, Baltimore, MD 21251.
E-mail: farin.kamangar@morgan.edu.
Accepted for publication: 17 April 2013

height with some error. For example, if the real mean height of the entire population is 182.3 cm, a sample may estimate it as 182.5 mm, which has a 2 mm error. If we want to be relatively certain that our sample mean has no more than 1 mm of error, the required sample size is much larger than when we accept an error of 10 mm (1 cm). For further details, please see Example 5.●

2-3- Probability of making a certain magnitude of error

The smaller the probability of the error, the larger the sample size should be. Consider Example 3. We are never sure that the error is necessarily going to be less than 1 mm. In a large population, there are many tall people. It could turn out that a random sample, however large, have a mean height of 3 mm higher than the entire population. We can only increase the sample size to the extent that with a large probability, e.g., 95% or 99%, the sample mean be within 1 mm from the true population mean. If we want 99% certainty, we would need a larger sample size than we accept 95% certainty. This is again somewhat intuitive, as larger sample size is the price we pay for more certainty. For more details, please see Example 6.

3- Some formulas

Each study is unique and needs its own formula. However, to provide some examples and to illustrate the principles mentioned above, we will provide formulas for three cases: 1) estimating the mean height in a population; 2) comparing the effects of two treatments on mean blood pressure; and 3) comparing the effects of two treatments on mortality.

3-1- Estimating mean height

We want to determine the mean height of men aged 18 or above in a very large population. Here, the required sample size depends on three factors: 1) the variance of height in men aged 18 or above (σ^2); 2) the maximum magnitude of error that we accept (d); and 3) the probability that our error will be higher than the acceptable magnitude (α). The following formula could be used to calculate sample size for this study:

$$n = \frac{\left(Z_{1-\alpha/2}\right)^2 (\sigma^2)}{d^2}$$

Now, we discuss how each of these elements is related to the principles discussed in Section 2.

3-1-1- Variance (σ^2)

According to the formula, the more variation in height, the larger our sample size should be. This is in line with Principle 1 in Section 2 (Principle 2-1).

Finding the correct σ to put in the formula is somewhat challenging. Determining the variance of height in the entire population depends on knowing its mean. Since we don't have the mean (otherwise we wouldn't do the study!), we cannot know the variance, so we can only estimate it, which is subject to some error.

Example 4: What number do we use for variance to estimate the mean height of the population? If we assume that height is normally distributed, 95% of the values of height in the population will fall in the range of mean ± two standard deviations, in other words in a range of four standard deviations. Therefore, if the height of 95% of the people falls roughly between 160 cm and 200 cm, it is reasonable to assume that the population standard deviation (σ) is about 40 ÷ 4 = 10 cm. Although this is a reasonable assumption, one can assume that the real standard deviation may be 12 cm, which increases the required sample size. Researchers may also use previous literature, if available, to estimate σ.●

3-1-2- The acceptable maximum error (d)

According to the formula, *d* is in the denominator. Therefore, the less error we accept, the larger the sample size should be. This is consistent with Principle 2-2.

Example 5: The researchers may decide that they would like the final estimate to be within 1 cm of the true number. This means that if our results show a mean height of 178 cm, we hope that the true number is between 177 cm and 179 cm. To show the effect of *d*, we fix other factors, for example Z = 2 and standard deviation (σ) = 10 cm. Under these circumstances, if we agree to a maximum error of 1 cm, our sample size needs to be 400 (4 × 100 ÷ 1). However, if we accept a maximum error of 0.2 cm (2 mm), then the sample size needs to be 10,000 (4 × 100 ÷ 0.04). ●

Example 5, in addition to illustrating the effect of acceptable error, shows that determination of sample size is not entirely clear-cut. One can substantially increase or decrease sample size by changing one of these factors, particularly by changing the acceptable error.

3-1-3- Probability of error falling outside d (α)

As mentioned earlier, there is always a probability that our error is larger than *d*. This probability is shown as α, and a function of it, ($Z_{1-\alpha/2}$) , or simply Z, appears in the formula. The smaller is the α, the larger is the Z. Therefore, a smaller probability of error needs a larger Z, and consequently a larger sample size. This is consistent with Principle 2-3.

Example 6: Many epidemiologic studies choose their α to be 5% (0.05). In this case Z, which is a function of α, will be 1.96. If one wants to reduce α to 1% (0.01), then Z will be 2.58. Reducing the probability of error requires a larger Z and hence a larger sample size.●

Since choosing α is at the discretion of the researcher, the required sample size is not entirely clear-cut. This is a lesson that we learnt from choice of *d* too.

If you feel that you have had enough of formulas, you can skip the rest of this section and go to Section 4. However, if you are interested in reading two more examples, go through Sections 3-2 and 3-3.

3-2- Comparing mean blood pressures

Suppose a study has one clearly-stated main objective: "To compare mean systolic blood pressures between patients receiving six months of treatment X versus those receiving six months of treatment Y". In this example, the formula is slightly more complex than the formula shown in Section 3-1, but many of the elements are common among the two. The sample size depends on four factors: 1) the variance of blood pressure in each group (σ^2); 2) the minimum difference that we would like to detect between the two treatments (d) ; 3) the probability of type I statistical error (α); and 4) the probability of type II statistical error (β). The following formula could be used to calculate sample size for each treatment group:

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \ (\sigma_1^2 + \sigma_2^2)}{d^2}$$

Now, we discuss the application and the intuitive meaning of each of the factors in this formula.

3-2-1- Variance (σ^2)

As discussed in Principle 2-1, sample size should be larger when variance of the blood pressure reduction is larger. Conversely, if there is no variation, for example if treatments X and Y decrease blood pressure in each participant by exactly 25 mmHg and 20 mmHg, we could accurately estimate the difference with only one person in each treatment group.

3-2-2- The minimum difference that we would like to detect (d)

In general, detecting a large difference requires a small sample size but detecting a small difference requires a large sample size. This is in line with Principle 2-2.

Finding the right *d* to put into the formula is challenging. Before the study, we don't know what the difference between the two treatments is. Therefore, we ask ourselves: "What is the smallest difference that matters to us?" If we desire to find only large differences between the two treatments, e.g., 10 mmHg, then the sample size wouldn't need to be that large. However, if we want to observe even very small differences, e.g., 1 mmHg, then sample size should be much larger, 100 times larger than that needed for the former situation.

It is intuitively understandable that detecting smaller differences requires larger sample sizes. Suppose you want to compare two students for their English spelling. If the difference between the two students is large, i.e., one student's spelling is far better than the other one, you could perhaps see the difference by asking only 10 questions. On the other hand, if both students are very strong and the difference is minimal, you may need to test them with 200 questions before you learn who is better.

3-2-3- Probability of type I error (α)

Type I error occurs when we erroneously reject the null hypothesis. To make it simpler and more relevant to our own example, type I error occurs if *in truth* (i.e., if one studies the entirety of our target population) the two treatments affect the mean blood pressure exactly the same but in our study sample we find that they are different. This is obviously an error because we find a difference where in reality there is no difference. Such errors may happen due to sampling variation. Thinking about tossing a coin (rather than blood pressure) may make it easier to understand.

Example 7: Let's say we want to know whether two coins are different with regards to their shape, such that the percentage of the heads for each of the coins is different. To learn this, we toss the two coins several times, compare the percentage of the heads, and perform statistical tests. However, in a single study, two completely similar coins may have statistically significant different results, which is a type I error.

To understand what was said above, let's change the experiment. Toss the first coin 10 times. By chance you may get two heads out of 10 (20%). Toss the same coin another 10 times, and you may get nine heads out of 10 (90%). The difference between these two numbers (20% and 90%) is statistically significant, with a two-sided Fisher exact P-value of 0.003. Since you used the same coin, obviously the difference in the percentage of heads in the two series of tosses was not due the design or shape of the coin; it was merely due to random variation (chance). In statistical terms, this was type I error, because while there was no difference, you detected one. ●

In study design, we usually fix the probability of type I error. For example, if we want a two-sided type I error probability of 0.05 (5%), its corresponding Z will be 1.96. For a type I error of 0.01, Z will be 2.58. If we want a smaller type I error, then our Z, and consequently our sample size would be larger. In other words, if we ask for a smaller probability of error, we need a larger sample size, which is in line with Principle 2-3.

3-2-4- Probability of type II error (β)

Type II error occurs when in truth the two treatments are different but we do not find the difference in our study sample. This happens quite commonly if the sample size is not large enough. We obviously want to reduce the probability of such errors, i.e., we want to detect differences if they exist. Reducing type II error is also called increasing *the power* of the study. If we want larger power, or smaller type II error, our Z will increase, which requires larger sample size. Intuitively, the smaller the probability of error, the larger our sample size should be (Principle 2-3).

3-3- Comparing mortality

Let's discuss the third case. We want to compare the effects of treatments X and Y on reducing mortality. To do this, we randomize subjects into two treatment groups, receiving X or Y. The required sample size for each group could be obtained from the formula below:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \quad [(\pi_1 \times (1-\pi_1) + (\pi_2 \times (1-\pi_2)]}{d^2}$$

The elements used in this formula are exactly the ones used in the previous formula, except that variance is replaced by $\pi \times (1-\pi)$, where π is the proportion of people who die in the entire population (with infinite sample size). It can be shown that this latter element plays the role of variance when the outcome is dichotomous (died or did not die).

4- Factors that need to be determined and their impact on sample size

A number of formulas were introduced in the previous section to calculate the study sample size. The question is: "What is the right formula for our study and what are the right numbers to put into it?" A number of decisions should be made before we choose the right formula. And after the formula is selected, we need to decide what numbers to put into the formula. Such decisions, which usually have enormous impacts on our calculated sample size, are discussed here.

4-1- Deciding on the main objective of the study

The main objective of the study is the primary factor in selecting the formula. For example, in Section 3, we chose three different formulas for studies with three different objectives. The objective should be specified very clearly. Even slight changes to the objectives may have a substantial impact on sample size.

Example 8: The objective of a study is: "To compare the effects of treatments X and Y on serum cholesterol in a randomized parallel design trial." As simple as it sounds, the objective still needs to be more clearly defined, as it will have a major impact on sample size. For example, the required sample size would be different if we chose to compare the effects of the treatments in a superiority trial – i.e., a trial that determines which drug is superior – versus if compared them in a noninferiority trial – i.e., our plan is to show that the new drug is not inferior to the standard treatment by a certain amount. Also, it would make a big difference if we decide to compare the mean cholesterol reduction versus we decide to compare the proportion of people whose cholesterol reach the target of < 200 mg/dL.●

Unfortunately, determination of the main study objective is not always straightforward, particularly in observational studies, such as case-control and cohort studies. Consider the example of a cohort study. During follow-up in a cohort study, there will be a large number of possible outcomes, including overall mortality, ischemic heart disease mortality, and esophageal cancer incidence. If the main outcome is a common event, such as overall mortality or mortality from ischemic heart disease, the sample size doesn't have to be very large, whereas if the main outcome is esophageal cancer incidence, a relatively uncommon cancer, then sample size has to be quite large. Likewise, in a typical cohort study, we collect information on a number of exposures, with each exposure having its own distribution. Although protocols often require that we determine sample size, it may not always be easy to determine in advance what *the main outcome* and *the main exposure* is. The usefulness of a cohort goes way beyond one outcome and one exposure.

4-2- Selecting the design of the study

Design of the study may have a major impact on the sample size. For example, it makes a large difference when we compare the effects of treatments X and Y on serum cholesterol in a parallel design trial versus a cross-over trial. Cross-over trials often need much smaller sample sizes, as each person receives both treatments. Also, each person serves as his or her own control, eliminating interpersonal variance, which again results in a smaller sample size.

4-3- Deciding on the proportion of participants distributed into each study arm

In the formulas discussed in Sections 3-2 and 3-3, we assumed equal sample size in each of the two treatment groups. However, we may decide otherwise. For example, in comparing treatments X (an old treatment) and Y (the new treatment), we may decide to randomize more people into receiving Y. This is because X is a well-known and long-used treatment, but Y is a new one and we want more information on it, particularly about its side effects. Given equal variance in the two study arms, uneven distribution of participants in the study arms requires higher sample size to obtain the same power.

Example 9: We want our study participants to be distributed with a ratio of 1 to 3 into X and Y treatment groups, respectively. If so, a total sample size of 1333 (333 in X and 1000 in Y) will give us the same power as randomizing 1000 people equally to each study group (500 X and 500 Y). Here, we need a 33% increase in sample size.●

4-4- Deciding between Bayesian versus frequentist methods

All of the formulas and much of the discussion made in this paper are based on the frequentist view of probability. This is because, at least thus far, much of the currently practiced statistics is based on frequentist methods. For example, P-value, power, type I error, and much of all other familiar statistics is rooted in frequentist view.

However, Bayesian methods are gaining popularity. If Bayesian analysis is considered, sample size calculations will be totally different. Sample size and power calculations for studies designed to be analyzed using Bayesian methods heavily depend on the prior distributions. Without going into any details, prior distributions may come from various sources, including our beliefs. For example, if a political leader believes that his opinion is definitely correct, no matter how much data you show him, he will stand by his prior opinion. If so, even a huge sample size showing the contrary would do no good! This one, of course, was an extreme example! Using prior distributions could be very helpful in some cases.

4-5- Deciding the numbers to put in the formulas

Consider the study presented in Section 3-2. Assume the literature suggests that the variance of blood pressure in each group is 20 mmHg. If we choose a power of 0.90 and a type I error level of 0.01, and we want to detect a d = 2 mmHg, the required sample size would be nearly 6000. However, if we choose a power of 0.80 and a type I error level of 0.05, and we want to detect a d = 3 mmHg, the required sample size would be approximately 1400. Therefore, with some minimal changes in requirements, all perfectly reasonable and within the ranges used by clinicians and statisticians, we can find the required sample size to be as low as 1400 or as high as 6000.

5- The impact of assumptions

Several assumptions have been made for doing the calculations made in Section 3. Departures from these assumptions may make sample size calculations incorrect. Below, we provide a few examples of the assumptions and show their impact on sample size calculations. To make it simple, in all examples we have assumed that the calculated sample size under the assumption is 1000.

5-1- Independence of study samples

The formulas and methods discussed so far assume that individuals in the sample are independent. However, if they are not, then the sample size must be larger to accommodate for lack of independence (sometimes referred to as clustering).

Example10: Consider the extreme example that identical twins always respond identically to a drug; i.e., the correlation between response from identical twins is 1.00. If so, when a researcher recruits 500 pairs of identical twins, although the sample size is 1000, it only provides us with information equivalent to 500 people; once we know the response from one twin, having the second one adds no further information. Here we say *the effective sample size* is 500.●

Example 11: Assume that to reduce costs of enrolling study participants, rather than selecting 1000 people randomly from an entire population, we randomly select 20 villages from the population and then randomly select 50 individuals from each village (two-stage cluster sampling). Since the responses obtained from each village can be correlated, the effective sample size may be less than 1000. If so, the effective sample will fall somewhere between the number of independent units (here, number of villages = 20) and the total number of study participants (here, 1000). In other words, our sample selection is not quite as good as recruiting 1000 independent people, but it is not as poor as selecting only 20 people. Without going into details of the formula, we suffice to say that the effective sample size depends on the total number of people, the number of units, and the intracluster correlation, i.e., the correlation between responses from individuals in each village. In this example, if the intracluster correlation is 0.10, the effective sample size approximately 170, which is indeed between 20 and 1000.●

5-2- No attrition

The formulas shown in Sections 3 assumed no sample attrition. If we assume an attrition of 20%, then the initial sample size should be 25% larger (1 ÷ 0.8 = 1.25), for example 1250 instead of 1000. However, it may be impossible to determine the extent of sample attrition prior to conducting the study.

5-3- Infinitely large target population

The formulas in Section 3 assumed that the target population was infinite. If the target population is finite, the required sample size may be slightly lower.

Example 12: If the target population is only 20,000 people, to determine a mean, we might need a sample size of 975 instead of 1000.●

As illustrated by these numbers (975 versus 1000), as long as the sample is relatively small compared to the target population (e.g., less than 5% of the entire population), the difference in sample size for finite and infinite populations is quite small. Therefore, size of the target population is usually not considered in sample size calculations.

5-4- No adjustment for baseline characteristics

The formulas in Section 3 did not consider adjusting for baseline characteristics. Multiple regression methods that adjust for baseline characteristics usually result in reduced variance, thus we obtain more power than we actually planned.

Example 13: Assume that the outcome of a study is depression after six months of treatment with X or Y. If we measure depression at study baseline, and baseline depression is highly correlated with the final one, then adjusting for baseline depression should in principle reduce the variance of final depression and thus make the study more powerful. If the correlation between baseline and final depression score is 0.50, taking this information into account, then a sample size of 1000 will actually give us a power equivalent to having 1333 in the study. ●

Note that a correlation of 0.50 is very high. With a correlation of 0.10, information from 1000 people provides with a power equal to having 1010 people. Most correlations are around this size (0.10 or so). Therefore, correlations are often ignored in sample size calculations.

6- Nonstatistical considerations in determining sample size

In addition to statistical calculations, there are other issues that may matter is choosing our study sample size. Funding, time, number of available patients, ethical issues, similar research being done elsewhere, and novelty of the research topic may play a role in determination of sample size.

6-1- Funding

As discussed in the previous sections, we can determine a reasonable range of sample sizes (e.g., from 1400 to 6000) for a study. If a researcher has funding to study only 20 subjects, then he perhaps shouldn't pursue that study. On the other hand, if he has large resources and large number of participants available, then he can determine a sample size between 1400 and 6000 for his study, depending on how much error he is willing to accept.

6-2- Ethical issues

Conducting a study with 100,000 people, where at most 6000 is needed, may be considered unethical, particularly if the study is a randomized trial testing a new drug.

6-3- Fixed number of patients available to the researcher

Sometimes sample size is almost fixed. For example, a medical researcher may have been able to collect data from 200 cases of a rare disease over his 20 years of experience (roughly 10 per year). If the researcher plans to increase sample size to 500, he may need to wait another 30 years (perhaps not feasible), or collaborate with other centers in the world, which again may or may not be feasible. Therefore, sample size is essentially fixed at 200. In circumstances like this, sample size formulas can be used, but not to determine sample size, rather to learn about the power to detect a certain difference. For example, with 200 cases and 800 controls, fixing type I error at 0.05, assuming a probability of exposure of 0.20 in controls based on previous research, we will have 84% power to detect a difference (reject the null hypothesis) if the true probability of exposure among cases is 0.30. Although the sample size is fixed, we can estimate power to detect a certain difference.

In some ways, determination of sample size is like buying a home; particularly when sample size is fixed. When you decide to buy a home and you can afford only $300,000, you may be able to buy a home with two bedrooms and a large living room, or a home with three bedrooms and a small living room. Likewise, if for financial or time constraints you can collect data from only 300 patients, that is what you can afford; with that you can get a small α and a large β, or a large α and a small β, or a small α and small β but a large d. You need to make sacrifices somewhere.

6-4- Novelty of the study

Novelty of the topic is important in making a decision to do a study or to publish a paper. The first report on what is now known as acquired immunodeficiency syndrome (AIDS), published in 1981, described only five cases of this disease, all in young homosexuals. However, given that the results were novel and the disease was rare, it was worth being published.[1] Today, a report of a far larger number of such cases may not be interesting enough for publication.

6-5- Similar studies being underway

Similar studies being conducted in other places could encourage or discourage conducting studies with relatively small sample sizes. On the one hand, availability of results from many similar studies may take away from novelty of the study. On the other hand, if multiple low-powered studies are conducted, then one could potentially do a meta-analysis or a combined analysis to increase power. Therefore, although each study by itself may not be definitive, combined together, they would greatly contribute to our knowledge.

7- Methods used to calculate sample size

Sample size can be calculated using formulas or simulation methods. In the example below, we will calculate the sample size using a formula.

Example 14: We would like to compare, in a randomized parallel design trial, the effect of treatments X and Y in reducing serum cholesterol in a group of hypercholesterolemic patients. What is the required sample size?

To answer this question, we first need to determine what parameter we are going to compare: the percentage of patients whose cholesterol is reduced to target levels after treatment, or the mean cholesterol after treatment? Let's assume we are going to compare means. If we want equal number of patients randomized to each group, the formula for calculating sample size in each group is:

$$n= \frac{\left(Z_{1-\alpha/2}+ Z_{1-\beta}\right)^2 \quad (\sigma^2_1+ \sigma^2_2)}{d^2}$$

Now, we need to determine each of the components. Let's assume that we accept a type I error of 0.05, for which Z is 1.96; and a type II error of 0.20 (power of 0.80), for which the corresponding Z is 0.84. We need to provide an estimated variance of cholesterol after treatment. After some literature review, we determine that a standard deviation of 30 mg/dL is a reasonable estimate for each of the treatments. Most importantly, we need to determine the minimum mean cholesterol difference between the two groups that is clinically useful and meaningful to us. Let's say a difference of 3 mg/dL is the minimum that we would like to be able to detect; below that, if we don't detect the difference, it doesn't matter, as it is a clinical tie. Plugging these numbers into the formula, we find that the sample size would need to be 1568 for each study arm, or a total of 3136.●

Since manual calculations may be tedious, software programs have been developed to calculate sample size. For example, using STATA's *sampsi* command, we obtain a sample size of 1570 for each group, or a total of 3140 cases. The minimal difference between manual and software calculations is due to rounding.

Prior to the wide availability of computers, tables and nomograms were developed and used to calculate sample size. Again, the idea was to reduce the pain of using formulas. Nomograms can be found in books or on the Internet.[2] Although they are relatively easy to use, nomograms may not be available for all study designs, objectives, or for all levels of type I and type II errors. Therefore, they are not as versatile as computers in calculating sample size, and their use provides little advantage over other methods. Tables have similar problems.

Simulation is another approach used to calculate sample size or power. This method is highly versatile – more so than using formulas – and can be used to calculate sample size under nearly all circumstances. It is most useful when there are no commands in our statistical package to calculate sample size of our study, mostly when the design is complex. However, simulation usually requires programming and therefore needs to be done by a statistician. As this method requires computer power, it has become more commonly used with the increased availability of faster computers. The idea is that we generate populations with the given parameters over and over (for example normal populations with means of 200 and 197 for treatments X and Y and standard deviations of 30 for each one), do the appropriate test (e.g., t-tests), and determine the proportion of the tests that found a statistically significant difference (here, P-value < 0.05). This latter proportion gives us the power. We can change the sample size to see which sample size gives us adequate power.

8- Software used to calculate sample size

Sample size can be calculated using almost all commercial statistical software, such as STATA and SAS. For example, STATA's *sampsi* command and SAS's *PROC POWER* can do the work for a variety of designs. There is also freely available and relatively easy-to-use software designed for sample size calculation. One example is the *PS Power and Sample Size Calculation* program, written by Dupont and Plummer at Vanderbilt University.[3] The program provides a step-by-step guide to calculate sample size. Another example is the *Power* program, written by Lubin and Garcia-Closas at the U.S. National Cancer Institute.[4,5] This program is particularly useful to calculate sample size when the outcome of interest is interaction. Yet another example is *Epi Info*, a free software for statistical analysis and power calculation, developed by the U.S. Centers for Disease Control and Prevention.

Conclusions

Statistical calculations of sample size depend on a number of factors including, but not limited to, the type of the study, the parameter that is going to be estimated (e.g., a mean or a proportion), the variance of the variable of interest, the acceptable type I and type II errors, clustering of the samples, and correlation among variables. Such calculations are to some extent subjective, because it is usually not obvious which numbers we should put in the formulas. The truth is that the number that comes out of the formula is only one acceptable number within an acceptable range. In addition, sample size may also depend on a number of nonstatistical factors, such as novelty of the study. As Norman and colleagues have put it,[6] "*Sample size estimates are like the emperor's clothes; we collectively act in public as if they possess an impressive aura of precision, yet privately we (statisticians) are acutely aware of their shortcomings and extreme imprecision.*" Having discussed all of these limitations, it is still prudent to calculate sample size statistically, as the results provide us with a range of reasonable sample sizes, as well as information on the power to detect a certain difference.

Acknowledgments

The authors would like to thank Dr. Ashkan Emadi (School of Medicine, University of Maryland, Baltimore, MD), Dr. Mahsa Mohebtash (Union Memorial Hospital, Baltimore, MD), and Ms. Gillian Silver (School of Community Health and Policy, Morgan State University) for reading the paper thoroughly and providing constructive comments.

References

1. Centers for Disease Control (CDC) and Prevention. Pneumocystis pneumonia--Los Angeles. *MMWR Morb Mortal Wkly Rep.* 1981; **30:** 250 – 252.
2. Altman DG. Statistics and ethics in medical research: III How large a sample? *Br Med J.* 1980; **281:** 1336 – 1338.
3. Dupont WD, Plummer WD, Jr. Power and sample size calculations for studies involving linear regression. *Control Clin Trials.* 1998; **19:** 589 – 601.
4. Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am J Epidemiol.* 1999; **149:** 689 – 692.
5. The US National Cancer Institute. Available from: URL: http://dceg.cancer.gov/tools/design/POWER. (Accessed Date: 23 April, 2013).
6. Norman G, Monteiro S, Salama S. Sample size calculations: should the emperor's clothes be off the peg or made to measure? *BMJ.* 2012; **345:** e5278.