5 Getting Ready to Estimate Sample Size: Hypotheses and Underlying Principles

Warren S. Browner, Thomas B. Newman, and Stephen B. Hulley

After an investigator has decided whom and what she is going to study and the design to be used, she must decide how many subjects to sample. Even the most rigorously executed study may fail to answer its research question if the sample size is too small. On the other hand, a study with too large a sample will be more difficult and costly than necessary. The goal of sample size planning is to estimate an **appropriate number** of subjects for a given study design.

Although a useful guide, sample size calculations give a deceptive impression of statistical objectivity. They are only as accurate as the data and estimates on which they are based, which are often just informed guesses. Sample size planning is a mathematical way of making a ballpark estimate. It often reveals that the research design is not feasible or that different predictor or outcome variables are needed. Therefore, sample size should be estimated early in the design phase of a study, when major changes are still possible.

Before setting out the specific approaches to calculating sample size for several common research designs in Chapter 6, we will spend some time considering the underlying principles. Readers who find some of these principles confusing will enjoy discovering that sample size planning does not require their total mastery. However, just as a recipe makes more sense if the cook is somewhat familiar with the ingredients, sample size calculations are easier if the investigator is acquainted with the basic concepts.

HYPOTHESES

The **research hypothesis** is a specific version of the research question that summarizes the main elements of the study—the sample, and the predictor and outcome variables—in a form that establishes the basis for tests of statistical significance. Hypotheses are not needed in descriptive studies, which describe how characteristics are distributed in a population, such as a study of the prevalence of a particular genotype among patients with hip fractures. (That does not mean, however, that

you won't need to do a sample size estimate for a descriptive study, just that the methods for doing so, described in Chapter 6, are different). Hypotheses are needed for studies that will use tests of statistical significance to compare findings among groups, such as a study of whether that particular genotype is more common among patients with hip fractures than among controls. Because most observational studies and all experiments address research questions that involve making comparisons, most studies need to specify at least one hypothesis. If any of the following terms appear in the research question, then the study is not simply descriptive, and a hypothesis should be formulated: greater than, less than, causes, leads to, compared with, more likely than, associated with, related to, similar to, correlated with.

Characteristics of a Good Hypothesis

A good hypothesis must be based on a good research question. It should also be simple, specific, and stated in advance.

Simple versus Complex. A simple hypothesis contains one predictor and one outcome variable:

> *A sedentary lifestyle is associated with an increased risk of proteinuria in patients with diabetes*

A complex hypothesis contains more than one predictor variable:

> *A sedentary lifestyle and alcohol consumption are associated with an increased risk of proteinuria in patients with diabetes*

Or more than one outcome variable:

> *Alcohol consumption is associated with an increased risk of proteinuria and of neuropathy in patients with diabetes*

Complex hypotheses like these are not readily tested with a single statistical test and are more easily approached as two or more simple hypotheses. Sometimes, however, a combined predictor or outcome variable can be used:

> *Alcohol consumption is associated with an increased risk of developing a microvascular complication of diabetes (i.e., proteinuria, neuropathy, or retinopathy) in patients with diabetes.*

In this example the investigator has decided that what matters is whether a participant has a complication, not what type of complication occurs.

Specific versus Vague. A specific hypothesis leaves no ambiguity about the subjects and variables or about how the test of statistical significance will be applied. It uses concise operational definitions that summarize the nature and source of the subjects and how variables will be measured.

> *Use of tricyclic antidepressant medications, assessed with pharmacy records, is more common in patients hospitalized with an admission diagnosis of myocardial infarction at Longview Hospital in the past year than in controls hospitalized for pneumonia.*

This is a long sentence, but it communicates the nature of the study in a clear way that minimizes any opportunity for testing something a little different once the study findings have been examined. It would be incorrect to substitute, during the analysis phase of the study, a different measurement of the predictor, such as the self-reported use of pills for depression, without considering the issue of multiple hypothesis testing (a topic we discuss at the end of the chapter). Usually, to keep the research hypothesis concise, some of these details are made explicit in the study plan rather than being stated in the research hypothesis. But they should always be clear in the investigator's conception of the study, and spelled out in the protocol.

It is often obvious from the research hypothesis whether the predictor variable and the outcome variable are dichotomous, continuous, or categorical. If it is not clear, then the type of variables can be specified:

> *Alcohol consumption (in mg/day) is associated with an increased risk of proteinuria (>300 mg/day) in patients with diabetes.*

If the research hypothesis begins to get too cumbersome, the definitions can be left out, as long as they are clarified elsewhere in the protocol.

In-Advance versus After-the-Fact. The hypothesis should be stated in writing at the outset of the study. Most important, this will keep the research effort focused on the primary objective. A single prestated hypothesis also creates a stronger basis for interpreting the study results than several hypotheses that emerge as a result of inspecting the data. Hypotheses that are formulated after examination of the data are a form of multiple hypothesis testing that can lead to overinterpreting the importance of the findings.

Types of Hypotheses

For the purpose of testing statistical significance, the research hypothesis must be restated in forms that categorize the expected difference between the study groups.

- ***Null and alternative hypotheses.*** The **null hypothesis** states that there is no association between the predictor and outcome variables in the population (*there is no difference in the frequency of drinking well water between subjects who develop peptic ulcer disease and those who do not*). The null hypothesis is the formal basis for testing statistical significance. Assuming that there really is no association in the population, statistical tests help to estimate the probability that an association observed in a study is due to chance.
- The proposition that there *is* an association (*the frequency of drinking well water is different in subjects who develop peptic ulcer disease than in those who do not*) is called the **alternative hypothesis.** The alternative hypothesis cannot be tested directly; it is accepted by default if the test of statistical significance rejects the null hypothesis (see later).
- ***One- and two-sided alternative hypotheses.*** A one-sided hypothesis specifies the direction of the association between the predictor and outcome variables. The hypothesis that drinking well water is more common among subjects who develop peptic ulcers is a one-sided hypothesis. A two-sided hypothesis states only that an association exists; it does not specify the direction. The hypothesis that subjects who develop peptic ulcer disease have a different frequency of drinking well water than those who do not is a two-sided hypothesis.

One-sided hypotheses may be appropriate in selected circumstances, such as when only one direction for an association is clinically important or biologically meaningful. An example is the one-sided hypothesis that a new drug for hypertension is more likely to cause rashes than a placebo; the possibility that the drug causes fewer rashes than the placebo is not usually worth testing (it might be if the drug had anti-inflammatory properties!). A one-sided hypothesis may also be appropriate when there is very strong evidence from prior studies that an association is unlikely to occur in one of the two directions, such as a study that tested whether cigarette smoking affects the risk of brain cancer. Because smoking has been associated with an increased risk of many different types of cancers, a one-sided alternative hypothesis (e.g., *that smoking increases the risk of brain cancer*) might suffice. However, investigators should be aware that many well-supported hypotheses (e.g., *that β-carotene therapy will reduce the risk of lung cancer, or that treatment with drugs that reduce the number of ventricular ectopic beats will reduce sudden death among patients with ventricular arrhythmias*) turn out to be wrong when tested in randomized trials. Indeed, in these two examples, the results of well-done trials revealed a statistically significant effect that was opposite in direction from the one supported by previous data (1–3). Overall, we believe that nearly all alternative hypotheses deserve to be two-sided.

It is important to keep in mind the difference between a **research hypothesis,** which is often one-sided, and the alternative hypothesis that is used when planning sample size, which is almost always two-sided. For example, suppose the research hypothesis is that recurrent use of antibiotics during childhood is associated with an increased risk of inflammatory bowel disease. That hypothesis specifies the direction of the anticipated effect, so it is one-sided. Why use a two-sided alternative hypothesis when planning the sample size? The answer is that most of the time, both sides of the alternative hypothesis (i.e., greater risk or lesser risk) are interesting, and the investigators would want to publish the results no matter which direction was observed. Statistical rigor requires the investigator choose between one- and two-sided hypotheses before analyzing the data; switching to a one-sided alternative hypothesis to reduce the *P* value (see below) is not correct. In addition (and this is probably the real reason that two-sided alternative hypotheses are much more common), most grant and manuscript reviewers expect two-sided hypotheses, and are critical of a one-sided approach.

UNDERLYING STATISTICAL PRINCIPLES

A hypothesis, such as that *15 minutes or more of exercise per day is associated with a lower mean fasting blood glucose level in middle-aged women with diabetes*, is either true or false in the real world. Because an investigator cannot study all middle-aged women with diabetes, she must test the hypothesis in a sample of that target population. As noted in Figure 1.6, there will always be a need to draw inferences about phenomena in the population from events observed in the sample.

In some ways, the investigator's problem is similar to that faced by a jury judging a defendant (Table 5.1). The absolute truth about whether the defendant committed the crime cannot usually be determined. Instead, the jury begins by presuming innocence: the defendant did not commit the crime. The jury must decide whether there is sufficient evidence to reject the presumed innocence of the defendant; the standard is known as **beyond a reasonable doubt.** A jury can err, however, by convicting an innocent defendant or by failing to convict a guilty one.

TABLE 5.1 The Analogy between Jury Decisions and Statistical Tests

Jury Decision	Statistical Test
Innocence: The defendant did not counterfeit money.	Null hypothesis: There is no association between dietary carotene and the incidence of colon cancer in the population.
Guilt: The defendant did counterfeit money.	Alternative hypothesis: There is an association between dietary carotene and the incidence of colon cancer.
Standard for rejecting innocence: Beyond a reasonable doubt.	Standard for rejecting null hypothesis: Level of statistical significance (α).
Correct judgment: Convict a counterfeiter.	Correct inference: Conclude that there is an association between dietary carotene and colon cancer when one does exist in the population.
Correct judgment: Acquit an innocent person.	Correct inference: Conclude that there is no association between carotene and colon cancer when one does not exist.
Incorrect judgment: Convict an innocent person.	Incorrect inference (type I error): Conclude that there is an association between dietary carotene and colon cancer when there actually is none.
Incorrect judgment: Acquit a counterfeiter.	Incorrect inference (type II error): Conclude that there is no association between dietary carotene and colon cancer when there actually is one.

In similar fashion, the investigator starts by presuming the null hypothesis of no association between the predictor and outcome variables in the population. Based on the data collected in her sample, the investigator uses statistical tests to determine whether there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis that there is an association in the population. The standard for these tests is known as the **level of statistical significance.**

Type I and Type II Errors

Like a jury, an investigator may reach a wrong conclusion. Sometimes by chance alone a sample is not representative of the population and the results in the sample do not reflect reality in the population, leading to an erroneous inference. A **type I error** (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a **type II error** (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually not true in the population. Although type I and type II errors can never be avoided entirely, the investigator can reduce their likelihood by increasing the sample size (the larger the sample, the less likely that it will differ substantially from the population) or by manipulating the design or the measurements in other ways that we will discuss.

In this chapter and the next, we deal only with ways to reduce type I and type II errors due to **chance** variation, also known as random error. False-positive and false-negative results can also occur because of **bias,** but errors due to bias are not usually referred to as type I and II errors. Such errors are especially troublesome, because they may be difficult to detect and cannot usually be quantified using statistical methods

or avoided by increasing the sample size. (See Chapters 1, 3, 4, and 7 through 12 for ways to reduce errors due to bias.)

Effect Size

The likelihood that a study will be able to detect an association between a predictor and an outcome variable in a sample depends on the actual magnitude of that association in the population. If it is large (*mean fasting blood glucose levels are 20 mg/dL lower in diabetic women who exercise than in those who do not*), it will be easy to detect in the sample. Conversely, if the size of the association is small (*a difference of 2 mg/dL*), it will be difficult to detect in the sample.

Unfortunately, the investigator does not usually know the exact size of the association; one of the purposes of the study is to estimate it! Instead, the investigator must choose the size of the association that she expects to be present in the sample. That quantity is known as the **effect size.** Selecting an appropriate effect size is the most difficult aspect of sample size planning (4). The investigator should first try to find data from prior studies in related areas to make an informed guess about a reasonable effect size. When data are not available, it may be necessary to do a small pilot study. Alternatively, she can choose the smallest effect size that in her opinion would be clinically meaningful (*a 10 mg/dL reduction in the fasting glucose level*).

Of course, from the public health point of view, even a reduction of 2 or 3 mg/dL in fasting glucose levels might be important, especially if it was easy to achieve. The choice of the effect size is always arbitrary, and considerations of feasibility are often paramount. Indeed, when the number of available or affordable subjects is limited, the investigator may have to work backward (Chapter 6) to determine the effect size that her study will be able to detect.

There are many different ways to measure the size of an association, especially when the outcome variable is dichotomous. For example, consider a study of whether middle-aged men are more likely to have impaired hearing than middle-aged women. Suppose an investigator finds that 20% of women and 30% of men 50 to 65 years of age are hard of hearing. These results could be interpreted as showing that men are 10% more likely to have impaired hearing than women (30% − 20%, the absolute difference), or 50% more likely ([30% − 20%] ÷ 20%, the relative difference). For sample size planning, both of the proportions matter; the sample size tables in this book use the smaller proportion (in this case, 20%) and the absolute difference (10%) between the groups being compared.

Many studies measure several effect sizes, because they measure several different predictor and outcome variables. For sample size planning, the sample size using the desired effect size for the most important hypothesis should be determined; the effect sizes for the other hypotheses can then be estimated. If there are several hypotheses of similar importance, then the sample size for the study should be based on whichever hypothesis needs the largest sample.

α, β, and Power

After a study is completed, the investigator uses statistical tests to try to reject the null hypothesis in favor of its alternative, in much the same way that a prosecuting attorney tries to convince a jury to reject innocence in favor of guilt. Depending on whether the null hypothesis is true or false in the target population, and assuming that the study is free of bias, four situations are possible (Table 5.2). In two of these, the findings in the sample and reality in the population are concordant, and the

TABLE 5.2 Truth in the Population versus the Results in the Study Sample: The Four Possibilities

	Truth in the Population	
Results in the Study Sample	**Association Between Predictor and Outcome**	**No Association Between Predictor and Outcome**
Reject null hypothesis	Correct	Type I error
Fail to reject null hypothesis	Type II error	Correct

investigator's inference will be correct. In the other two situations, either a type I or type II error has been made, and the inference will be incorrect.

The investigator establishes the maximum chance that she will tolerate of making type I and II errors in advance of the study. The probability of committing a type I error (rejecting the null hypothesis when it is actually true) is called α **(alpha).** Another name for α is the **level of statistical significance.**

If, for example, a study of the effects of exercise on fasting blood glucose levels is designed with an α of 0.05, then the investigator has set 5% as the maximum chance of incorrectly rejecting the null hypothesis if it is true (and inferring that exercise and fasting blood glucose levels are associated in the population when, in fact, they are not). This is the level of reasonable doubt that the investigator will be willing to accept when she uses statistical tests to analyze the data after the study is completed.

The probability of making a type II error (failing to reject the null hypothesis when it is actually false) is called β **(beta).** The quantity $[1 - \beta]$ is called **power,** the probability of correctly rejecting the null hypothesis in the sample if the actual effect in the population is equal to (or greater than) the effect size.

If β is set at 0.10, then the investigator has decided that she is willing to accept a 10% chance of missing an association of a given effect size if it exists. This represents a power of 0.90; that is, a 90% chance of finding an association of that size or greater. For example, suppose that exercise really would lead to an average reduction of 20 mg/dL in fasting glucose levels among diabetic women in the entire population. Suppose that the investigator drew a sample of women from the population on numerous occasions, each time carrying out the same study (with the same measurements and the same 90% power each time). Then in nine of every ten studies the investigator would correctly reject the null hypothesis and conclude that exercise is associated with fasting glucose level. This does not mean, however, that the investigator doing a single study will be unable to detect it if the effect actually present in the population was smaller, say, a 15 mg/dL reduction; it means simply that she will have less than a 90% likelihood of doing so.

Ideally, α and β would be set at zero, eliminating the possibility of false-positive and false-negative results. In practice they are made as small as possible. Reducing them, however, requires increasing the sample size; other strategies are discussed in Chapter 6. Sample size planning aims at choosing a sufficient number of subjects to keep α and β at an acceptably low level without making the study unnecessarily expensive or difficult.

Many studies set α at 0.05 and β at 0.20 (a power of 0.80). These are arbitrary values, and others are sometimes used: the conventional range for α is between 0.01 and 0.10, and that for β is between 0.05 and 0.20. In general, the investigator should

use a low α when the research question makes it particularly important to avoid a type I (false-positive) error—for example, in testing the efficacy of a potentially dangerous medication. She should use a low β (and a small effect size) when it is especially important to avoid a type II (false-negative) error—for example, in reassuring the public that living near a toxic waste dump is safe.

P Value

The null hypothesis acts like a straw man: it is assumed to be true so that it can be knocked down as false with a statistical test. When the data are analyzed, such tests determine the ***P* value,** the probability of seeing an effect as big as or bigger than that in the study by chance if the null hypothesis actually were true. The null hypothesis is rejected in favor of its alternative if the P value is less than α, the predetermined level of statistical significance.

A "nonsignificant" result (i.e., one with a P value greater than α) does not mean that there is no association in the population; it only means that the result observed in the sample is small compared with what could have occurred by chance alone. For example, an investigator might find that men with hypertension were twice as likely to develop prostate cancer as those with normal blood pressure, but because the number of cancers in the study was modest this apparent effect had a P value of only 0.08. This means that even if hypertension and prostatic carcinoma were not associated in the population, there would be an 8% chance of finding such an association due to random error in the sample. If the investigator had set the significance level as a two-sided α of 0.05, she would have to conclude that the association in the sample was "not statistically significant." It might be tempting for the investigator to change her mind about the level of statistical significance, reset the two-sided α to 0.10, and report, "The results showed a statistically significant association ($P < 0.10$)," or switch to a one-sided P value and report it as "$P = 0.04$." A better choice would be to report that "The results, although suggestive of an association, did not achieve statistical significance ($P = 0.08$)."

This solution acknowledges that statistical significance is not an all-or-none situation. In part because of this problem, many statisticians and epidemiologists are moving away from hypothesis testing, with its emphasis on P values, to using confidence intervals to report the precision of the study results (5–7). However, for the purposes of sample size planning for analytic studies, hypothesis testing is still the standard.

Sides of the Alternative Hypothesis

Recall that an alternative hypothesis actually has two sides, either or both of which can be tested in the sample by using **one-** or **two-sided statistical tests.** When a two-sided statistical test is used, the P value includes the probabilities of committing a type I error in each of two directions, which is about twice as great as the probability in either direction alone. It is easy to convert from a one-sided P value to a two-sided P value, and vice versa. A one-sided P value of 0.05, for example, is usually the same as a two-sided P value of 0.10. (Some statistical tests are asymmetric, which is why we said "usually.")

In those rare situations in which an investigator is only interested in one of the sides and has so formulated the alternative hypothesis, sample size should be calculated accordingly. A one-sided hypothesis should never be used just to reduce the sample size.

Type of Statistical Test

The formulas used to calculate sample size are based on mathematical assumptions, which differ for each statistical test. Before the sample size can be calculated, the investigator must decide on the statistical approach to analyzing the data. That choice depends mainly on the type of predictor and outcome variables in the study. Table 6.1 lists some common statistics used in data analysis, and Chapter 6 provides simplified approaches to estimating sample size for studies that use these statistics.

ADDITIONAL POINTS

Variability

It is not simply the size of an effect that is important; its variability also matters. Statistical tests depend on being able to show a difference between the groups being compared. The greater the variability (or spread) in the outcome variable among the subjects, the more likely it is that the values in the groups will overlap, and the more difficult it will be to demonstrate an overall difference between them. Because measurement error contributes to the overall variability, less precise measurements require larger sample sizes (8).

Consider a study of the effects of two isocaloric diets (low fat and low carbohydrate) in achieving weight loss in 20 obese patients. If all those on the low-fat diet lost about 3 kg and all those on the low-carbohydrate diet failed to lose much weight (an effect size of 3 kg), it is likely that the low-fat diet really is better (Fig. 5.1A). On the other hand, suppose that although the average weight loss is 3 kg in the low-fat group and 0 kg in the low-carbohydrate group, there is a great deal of overlap between the two groups. (The changes in weight vary from a loss of 8 kg to a gain of 8 kg.) In this situation (Fig. 5.1B), although the effect size is still 3 kg, the greater variability will make it more difficult to detect a difference between the diets, and a larger sample size will be needed.

When one of the variables used in the sample size estimate is continuous (e.g., body weight in Figure 5.1), the investigator will need to estimate its variability. (See the section on the t test in Chapter 6 for details.) In the other situations, variability is already included in the other parameters entered into the sample size formulas and tables, and need not be specified.

Multiple and Post Hoc Hypotheses

When more than one hypothesis is tested in a study, especially if some of those hypotheses were formulated after the data were analyzed (*post hoc* hypotheses), the likelihood that at least one will achieve statistical significance on the basis of chance alone increases. For example, if 20 independent hypotheses are tested at an α of 0.05, the likelihood is substantial (64%; $[1 - 0.95^{20}]$) that at least one hypothesis will be statistically significant by chance alone. Some statisticians advocate adjusting the level of statistical significance when more than one hypothesis is tested in a study. This keeps the overall probability of accepting any one of the alternative hypotheses, when all the findings are due to chance, at the specified level. For example, genomic studies that look for an association between hundreds (or even thousands) of genotypes and a disease need to use a much smaller α than 0.05, or they risk identifying many false-positive associations.

One approach, named after the mathematician **Bonferroni,** is to divide the significance level (say, 0.05) by the number of hypotheses tested. If there were four

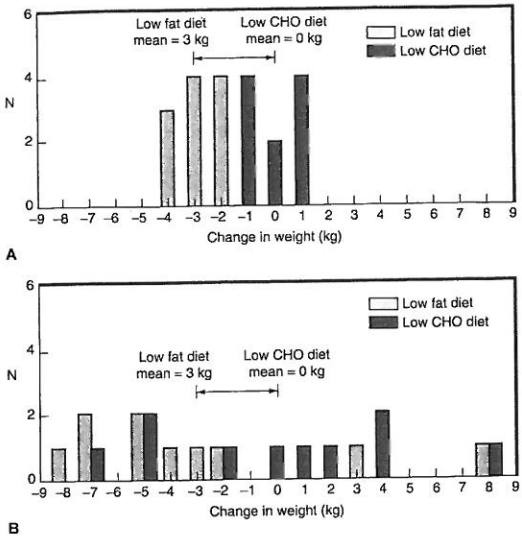

FIGURE 5.1. A: Weight loss achieved by two diets. All subjects on the low-fat diet lost from 2 to 4 kg, whereas weight change in those on the low-carbohydrate (CHO) diet varied from −1 to +1 kg. Because there is no overlap between the two groups, it is reasonable to infer that the low-fat diet is better at achieving weight loss than the low-carbohydrate diet (as would be confirmed with a t test, which has a P value < 0.0001). **B:** Weight loss achieved by two diets. There is substantial overlap in weight change in the two groups. Although the effect size is the same (3 kg) as in **A**, there is little evidence that one diet is better than the other (as would be confirmed with a t test, which has a P value of 0.19).

hypotheses, for example, each would be tested at an α of 0.0125 (i.e., 0.05 ÷ 4). This would require substantially increasing the sample size over that needed for testing each hypothesis at an α of 0.05.

We believe that a Bonferroni-type of approach to **multiple hypothesis** testing is usually too stringent. Investigators do not adjust the significance levels for hypotheses that are tested in separate studies. Why do so when several hypotheses are tested in the same study? In our view, adjusting α for multiple hypotheses is chiefly useful when the likelihood of making false-positive errors is high, because the number of tested hypotheses is substantial (say, more than ten) and the prior probability for each hypothesis is low (e.g., in screening a large number of genes for association with a phenotype). The first criterion is actually stricter than it may appear, because what matters is the number of hypotheses that are *tested*, not the number that are *reported*. Testing 50 hypotheses but only reporting or emphasizing the one or two P values

that are less than 0.05 is misleading. Adjusting α for multiple hypotheses is especially important when the consequences of making a false-positive error are large, such as mistakenly concluding that an ineffective treatment is beneficial.

In general, the issue of what significance level to use depends more on the **prior probability** of each hypothesis than on the number of hypotheses tested. There is an analogy with the use of diagnostic tests that may be helpful (9). When interpreting the results of a diagnostic test, a clinician considers the likelihood that the patient being tested has the disease in question. For example, a modestly abnormal test result in a healthy person (a serum alkaline phosphatase level that is 15% greater than the upper limit of normal) is probably a false-positive test that is unlikely to have much clinical importance. Similarly, a P value of 0.05 for an unlikely hypothesis is probably also a false-positive result.

However, an alkaline phosphatase level that is 10 or 20 times greater than the upper limit of normal is unlikely to have occurred by chance (although it might be a laboratory error). So too a very small P value (say, ≤0.001) is unlikely to have occurred by chance (although it could be due to bias). It is hard to dismiss very abnormal test results as being false-positives or to dismiss very low P values as being due to chance, even if the prior probability of the disease or the hypothesis was low.

Moreover, the number of tests that were ordered, or hypotheses that were tested, is not always relevant. The interpretation of an elevated serum uric acid level in a patient with a painful and swollen joint should not depend on whether the physician ordered just a single test (the uric acid level) or obtained the result as part of a panel of 20 tests. Similarly, when interpreting the P value for testing a research hypothesis that makes good sense, it should not matter that the investigator also tested several unlikely hypotheses. What matters most is the reasonableness of the research hypothesis being tested: that it has a substantial prior probability of being correct. (Prior probability, in this **"Bayesian"** approach, is usually a subjective judgment based on evidence from other sources.) Hypotheses that are formulated during the design of a study usually meet this requirement; after all, why else would the investigator put the time and effort into planning and doing the study?

What about unanticipated associations that appear during the collection and analysis of a study's results? This process is sometimes called **hypothesis generation** or, less favorably, "data-mining" or a "fishing expedition." The many informal comparisons that are made during data analysis are a form of multiple hypothesis testing. A similar problem arises when variables are redefined during data analysis, or when results are presented for subgroups of the sample. Significant P values for data-generated hypotheses that were not considered during the design of the study are often due to chance. They should be viewed with interest but skepticism and considered a fertile source of potential research questions for future studies.

Sometimes, however, an investigator fails to specify a particular hypothesis in advance, although that hypothesis seems reasonable when it is time for the data to be analyzed. This might happen, for example, if others discover a new risk factor while the study is going on, or if the investigator just didn't happen to think of a particular hypothesis when the study was being designed. The important issue is not so much whether the hypothesis was formulated before the study began, but whether there is a reasonable prior probability based on evidence from other sources that the hypothesis is true (9).

There are some definite **advantages** to formulating more than one hypothesis when planning a study. The use of multiple unrelated hypotheses increases the efficiency of the study, making it possible to answer more questions with a single

research effort and to discover more of the true associations that exist in the population. It may also be a good idea to formulate several **related hypotheses;** if the findings are consistent, the study conclusions are made stronger. Studies in patients with heart failure have found that the use of angiotensin-converting enzyme inhibitors is beneficial in reducing cardiac admissions, cardiovascular mortality, and total mortality. Had only one of these hypotheses been tested, the inferences from these studies would have been less definitive. Lunch may not be free, however, when multiple hypotheses are tested. Suppose that when these related and prestated hypotheses are tested, only one turns out to be statistically significant. Then the investigator must decide (and try to convince editors and readers) whether the significant results, the nonsignificant results, or both sets of results are true.

Primary and Secondary Hypotheses

Some studies, especially large randomized trials, specify some hypotheses as being **"secondary."** This usually happens when there is one **primary hypothesis** around which the study has been designed, but the investigators are also interested in other research questions that are of lesser importance. For example, the primary outcome of a trial of zinc supplementation might be hospitalizations or emergency department visits for upper respiratory tract infections; a secondary outcome might be self-reported days missed from work or school. If the study is being done to obtain approval for a pharmaceutical agent, the primary outcome is what will matter most to the regulatory body. The sample size calculations are always focused on the primary hypothesis, and secondary hypotheses with insufficient power should be avoided. Stating a secondary hypothesis in advance does increase the credibility of the results. Stating a secondary hypothesis after the data have been collected and analyzed is another form of data dredging.

A good rule, particularly for clinical trials, is to establish in advance as many hypotheses as make sense, but specify just one as the **primary hypothesis,** which can be tested statistically without argument about whether to adjust for multiple hypothesis testing. More important, having a primary hypothesis helps to focus the study on its main objective and provides a clear basis for the main sample size calculation.

SUMMARY

1. **Sample size planning** is an important part of the design of both analytic and descriptive studies. The sample size should be estimated early in the process of developing the research design, so that appropriate modifications can be made.

2. Analytic studies and experiments need a **hypothesis** that specifies, for the purpose of subsequent **statistical tests,** the anticipated association between the main predictor and outcome variables. Purely descriptive studies, lacking the strategy of comparison, do not require a hypothesis.

3. Good hypotheses are **specific** about how the population will be sampled and the variables measured, **simple** (there is only one predictor and one outcome variable), and **formulated in advance.**

4. The **null hypothesis,** which proposes that the predictor and outcome variables are not associated, is the basis for tests of statistical significance. The **alternative hypothesis** proposes that they are associated. Statistical tests attempt to reject the null hypothesis of no association in favor of the alternative hypothesis that there is an association.

5. An alternative hypothesis is either **one-sided** (only one direction of association will be tested) or **two-sided** (both directions will be tested). One-sided hypotheses should only be used in unusual circumstances, when only one direction of the association is clinically or biologically meaningful.

6. For analytic studies and experiments, the sample size is an estimate of the number of subjects required to detect an association of a given **effect size** and **variability** at a specified likelihood of making **type I** (false-positive) and **type II** (false-negative) **errors.** The maximum likelihood of making a type I error is called α; that of making a type II error, β. The quantity $(1 - \beta)$ is **power,** the chance of observing an association of a given effect size or greater in a sample if one is actually present in the population.

7. It is often desirable to establish more than one hypothesis in advance, but the investigator should specify a single **primary hypothesis** as a focus and for sample size estimation. Interpretation of findings from testing **multiple hypotheses** in the sample, including unanticipated findings that emerge from the data, is based on a judgment about the **prior probability** that they represent real phenomena in the population.

REFERENCES

1. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med* 1994;330:1029–1035.
2. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;324:781–788.
3. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.
4. Van Walraven C, Mahon JL, Moher D, et al. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *J Clin Epidemiol* 1999;52: 717–723.
5. Daly LE. Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol* 1998;147:783–790.
6. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 1999;130:995–1004.
7. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130:1005–1013.
8. McKeown-Eyssen GE, Tibshirani R. Implications of measurement error in exposure for the sample sizes of case-control studies. *Am J Epidemiol* 1994;139:415–421.
9. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;257:2459–2463.

6 Estimating Sample Size and Power: Applications and Examples

Warren S. Browner, Thomas B. Newman, and Stephen B. Hulley

Chapter 5 introduced the basic principles underlying sample size calculations. This chapter presents several cookbook techniques for using those principles to estimate the sample size needed for a research project. The first section deals with sample size estimates for an analytic study or experiment, including some special issues that apply to these studies such as multivariate analysis. The second section considers studies that are primarily descriptive. Subsequent sections deal with studies that have a fixed sample size, strategies for maximizing the power of a study, and how to estimate the sample size when there appears to be insufficient information from which to work. The chapter concludes with common errors to avoid.

At the end of the chapter, there are tables and formulas in the appendixes for several basic methods of estimating sample size. In addition, there is a calculator on our website (www.epibiostat.ucsf.edu/dcr/), and there are many sites on the Web that can provide instant interactive sample size calculations; try searching for "sample size" *and* "power" *and* "interactive". Most statistical packages can also estimate sample size for common study designs.

SAMPLE SIZE TECHNIQUES FOR ANALYTIC STUDIES AND EXPERIMENTS

There are several variations on the recipe for estimating sample size in an analytic study or experiment, but they all have certain steps in common:

1. State the **null hypothesis** and either a **one-** or **two-sided alternative hypothesis.**
2. Select the appropriate **statistical test** from Table 6.1 based on the type of predictor variable and outcome variable in those hypotheses.
3. Choose a reasonable **effect size** (and **variability,** if necessary).

TABLE 6.1 Simple Statistical Tests for Use in Estimating Sample Size*

	Outcome Variable	
Predictor Variable	**Dichotomous**	**Continuous**
Dichotomous	Chi-squared test†	*t* test
Continuous	*t* test	Correlation coefficient

* See text for what to do about ordinal variables, or if planning to analyze the data with another type of statistical test.
† The chi-squared test is always two-sided; a one-sided equivalent is the Z statistic.

4. Set α and β. (Specify a two-sided α unless the alternative hypothesis is clearly one-sided.)
5. Use the appropriate table or formula in the appendix to estimate the sample size.

Even if the exact value for one or more of the ingredients is uncertain, it is important to estimate the sample size early in the design phase. Waiting until the last minute to prepare the sample size can be a rude awakening: it may be necessary to start over with new ingredients, which may mean redesigning the entire study. This is why this subject is covered early in this book.

Not all analytic studies fit neatly into one of the three main categories that follow; a few of the more common exceptions are discussed in the section called "Other Considerations and Special Issues."

The t Test

The ***t* test** (sometimes called "Student's t test," after the pseudonym of its developer) is commonly used to determine whether the mean value of a continuous outcome variable in one group differs significantly from that in another group. For example, the *t* test would be appropriate to use when comparing the mean depression scores in patients treated with two different antidepressants, or the mean change in weight among two groups of participants in a placebo-controlled trial of a new drug for weight loss. The *t* test assumes that the distribution (spread) of the variable in each of the two groups approximates a normal (bell-shaped) curve. However, the *t* test is remarkably robust, so it can be used for almost any distribution unless the number of subjects is small (fewer than 30 to 40) or there are extreme outliers.

To estimate the sample size for a study that will be analyzed with a *t* test (see Example 6.1), the investigator must

1. State the null hypothesis and whether the alternative hypothesis is one- or two-sided.
2. Estimate the effect size (E) as the difference in the mean value of the outcome variable between the study groups.
3. Estimate the variability of the outcome variable as its standard deviation (S).
4. Calculate the standardized effect size (E/S), defined as the effect size divided by the standard deviation of the outcome variable.
5. Set α and β.

The **effect size** and **variability** can often be estimated from previous studies in the literature and consultation with experts. Occasionally, a small pilot study will be necessary to estimate the standard deviation of the outcome variable (also see the Section "How to estimate sample size when there is insufficient information," later in this chapter). When the outcome variable is the change in a continuous measurement (e.g., change in weight during a study), the investigator should use the standard deviation of the change in that variable (not the standard deviation of the variable itself) in the sample size estimates. The standard deviation of the change in a variable is usually smaller than the standard deviation of the variable; therefore the sample size will also be smaller.

The **standardized effect size** is a unitless quantity that makes it possible to estimate a sample size when an investigator cannot obtain information about the variability of the outcome variable; it also simplifies comparisons between the effect sizes of different variables. (The standardized effect size equals the effect size divided by the standard deviation of the outcome variable. For example, a 10 mg/dL difference in serum cholesterol level, which has a standard deviation in the population of about 40 mg/dL, would equal a standardized effect size of 0.25.) The larger the standardized effect size, the smaller the required sample size. For most studies, the standardized effect size will be >0.1. Effect sizes smaller than that are difficult to detect (they require very large sample sizes) and usually not very important clinically.

Appendix 6A gives the sample size requirements for various combinations of α and β for several standardized effect sizes. To use Table 6A, look down its leftmost column for the standardized effect size. Next, read across the table to the chosen values for α and β for the sample size required per group. (The numbers in Table 6A assume that the two groups being compared are of the same size; use the formula below the table or an interactive Web-based program if that assumption is not true.)

Example 6.1 Calculating Sample Size When Using the *t* Test

Problem: The research question is whether there is a difference in the efficacy of salbutamol and ipratropium bromide for the treatment of asthma. The investigator plans a randomized trial of the effect of these drugs on FEV_1 (forced expiratory volume in 1 second) after 2 weeks of treatment. A previous study has reported that the mean FEV_1 in persons with treated asthma was 2.0 liters, with a standard deviation of 1.0 liter. The investigator would like to be able to detect a difference of 10% or more in mean FEV_1 between the two treatment groups. How many patients are required in each group (salbutamol and ipratropium) at α (two-sided) = 0.05 and power = 0.80?

Solution: The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** *Mean FEV_1 after 2 weeks of treatment is the same in asthmatic patients treated with salbutamol as in those treated with ipratropium.*
   Alternative Hypothesis *(two-sided): Mean FEV_1 after 2 weeks of treatment is different in asthmatic patients treated with salbutamol from what it is in those treated with ipratropium.*
2. *Effect Size = 0.2 liters (10% × 2.0 liters).*
3. *Standard Deviation of FEV_1 = 1.0 liter.*

4. *Standardized Effect Size = effect size ÷ standard deviation = 0.2 liters ÷ 1.0 liter = 0.2.*
5. *α (two-sided) = 0.05; β = 1 - 0.80 = 0.20. (Recall that β = 1 − power.)*

Looking across from a standardized effect size of 0.20 in the leftmost column of Table 6A and down from α (two-sided) = 0.05 and β = 0.20, 394 patients are required per group. This is the number of patients in each group who need to complete the study; even more will need to be enrolled to account for dropouts. This sample size may not be feasible, and the investigator might reconsider the study design, or perhaps settle for only being able to detect a larger effect size. See the section on the t test for paired samples ("Example 6.8") for a great solution.

The *t* test is usually used for comparing continuous outcomes, but it can also be used to estimate the sample size for a dichotomous outcome (e.g., in a case–control study) if the study has a continuous predictor variable. In this situation, the *t* test compares the mean value of the predictor variable in the cases with that in the controls.

There is a convenient **shortcut** for approximating sample size using the *t* test, when more than about 30 subjects will be studied and the power is set at 0.80 ($\beta = 0.2$) and α (two-sided) is set at 0.05 (1). The formula is

Sample size (per equal-sized group) = 16 ÷ (standardized effect size)^2.

For Example 6.1, the shortcut estimate of the sample size would be $16 \div 0.2^2 = 400$ per group.

The Chi-Squared Test

The **chi-squared test** (χ^2) can be used to compare the proportion of subjects in each of two groups who have a dichotomous outcome. For example, the proportion of men who develop coronary heart disease (CHD) while being treated with folate can be compared with the proportion who develop CHD while taking a placebo. The chi-squared test is always two-sided; an equivalent test for one-sided hypotheses is the **one-sided Z test.**

In an experiment or cohort study, effect size is specified by the difference between P_1, the proportion of subjects expected to have the outcome in one group, and P_2, the proportion expected in the other group. In a case–control study, P_1 represents the proportion of cases expected to have a particular risk factor, and P_2 represents the proportion of controls expected to have the risk factor. Variability is a function of P_1 and P_2, so it need not be specified.

To estimate the sample size for a study that will be analyzed with the chi-squared test or Z test to compare two proportions, the investigator must

1. State the null hypothesis and decide whether the alternative hypothesis should be one- or two-sided.
2. Estimate the effect size and variability in terms of P_1, the proportion with the outcome in one group, and P_2, the proportion with the outcome in the other group.
3. Set α and β.

Appendix 6B gives the sample size requirements for several combinations of α and β, and a range of values of P_1 and P_2. To estimate the sample size, look down the leftmost column of Tables 6B.1 or 6B.2 for the smaller of P_1 and P_2 (if necessary rounded to the nearest 0.05). Next, read across for the difference between P_1 and P_2. Based on the chosen values for α and β, the table gives the sample size required per group.

Example 6.2 Calculating Sample Size When Using the Chi-Squared Test

Problem: The research question is whether elderly smokers have a greater incidence of skin cancer than nonsmokers. A review of previous literature suggests that the 5-year incidence of skin cancer is about 0.20 in elderly nonsmokers. At α (two-sided) = 0.05 and power = 0.80, how many smokers and nonsmokers will need to be studied to determine whether the 5-year skin cancer incidence is at least 0.30 in smokers?

Solution: The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** *The incidence of skin cancer is the same in elderly smokers and nonsmokers.*
   Alternative Hypothesis *(two-sided): The incidence of skin cancer is different in elderly smokers and nonsmokers.*
2. *P_2 (incidence in nonsmokers) = 0.20; P_1 (incidence in smokers) = 0.30. The smaller of these values is 0.20, and the difference between them ($P_1 - P_2$) is 0.10.*
3. *α (two-sided) = 0.05; β = 1 - 0.80 = 0.20.*

Looking across from 0.20 in the leftmost column in Table 6B.1 and down from an expected difference of 0.10, the middle number for α (two-sided) = 0.05 and β = 0.20 is the required sample size of 313 smokers and 313 nonsmokers. If the investigator had chosen to use a one-sided alternative hypothesis, given that there is a great deal of evidence suggesting that smoking is a carcinogen and none suggesting that it prevents cancer, the sample size would be 251 smokers and 251 nonsmokers.

Often the investigator specifies the effect size in terms of the **relative risk** (risk ratio) of the outcome in two groups of subjects. For example, an investigator might study whether women who use oral contraceptives are at least twice as likely as nonusers to have a myocardial infarction. In a cohort study (or experiment), it is straightforward to convert back and forth between relative risk and the two proportions (P_1 and P_2), since the relative risk is just P_1 divided by P_2 (or vice versa).

For a case–control study, however, the situation is a little more complex because the relative risk must be approximated by the **odds ratio,** which equals $[P_1 \times (1 - P_2)] \div [P_2 \times (1 - P_1)]$. The investigator must specify the odds ratio (OR) and P_2 (the proportion of controls exposed to the predictor variable). Then P_1 (the proportion of cases exposed to the predictor variable) is

$$P_1 = \frac{\text{OR} \times P_2}{(1 - P_2) + (\text{OR} \times P_2)}$$

For example, if the investigator expects that 10% of controls will be exposed to the oral contraceptives ($P_2 = 0.1$) and wishes to detect an odds ratio of 3 associated with the exposure, then

$$P_1 = \frac{(3 \times 0.1)}{(1 - 0.1) + (3 \times 0.1)} = \frac{0.3}{1.2} = 0.25$$

The Correlation Coefficient

Although the **correlation coefficient** (r) is not used frequently in sample size calculations, it can be useful when the predictor and outcome variables are both continuous. The correlation coefficient is a measure of the strength of the linear association between the two variables. It varies between −1 and +1. Negative values indicate that as one variable increases, the other decreases (like blood lead level and IQ in children). The closer the absolute value of r is to 1, the stronger the association; the closer to 0, the weaker the association. Height and weight in adults, for example, are highly correlated in some populations, with $r \approx 0.9$. Such high values, however, are uncommon; many biologic associations have much smaller correlation coefficients.

Correlation coefficients are common in some fields of clinical research, such as behavioral medicine, but using them to estimate the sample size has a disadvantage: correlation coefficients have little intuitive meaning. When squared (r^2) a correlation coefficient represents the proportion of the spread (variance) in an outcome variable that results from its linear association with a predictor variable, and vice versa. That's why small values of r, such as those ≤0.3, may be statistically significant if the sample is large enough without being very meaningful clinically or scientifically, since they "explain" at most 9% of the variance.

An alternative—and often preferred—way to estimate the sample size for a study in which the predictor and outcome variables are both continuous is to dichotomize one of the two variables (say, at its median) and use the t test calculations instead. This has the advantage of expressing the effect size as a "difference" between two groups.

To estimate sample size for a study that will be analyzed with a correlation coefficient, the investigator must

1. State the null hypothesis, and decide whether the alternative hypothesis is one or two-sided.
2. Estimate the effect size as the absolute value of the smallest correlation coefficient (r) that the investigator would like to be able to detect. (Variability is a function of r and is already included in the table and formula.)
3. Set α and β.

In Appendix 6C, look down the leftmost column of Table 6C for the effect size (r). Next, read across the table to the chosen values for α and β, yielding the total sample size required. Table 6C yields the appropriate sample size when the investigator wishes to reject the null hypothesis that there is no association between the predictor and outcome variables (e.g., $r = 0$). If the investigator wishes to determine whether the correlation coefficient in the study differs from a value other than zero (e.g., $r = 0.4$), she should see the text below Table 6C for the appropriate methodology.

Example 6.3 Calculating Sample Size When Using the Correlation Coefficient in a Cross-Sectional Study

Problem: The research question is whether urinary cotinine levels (a measure of the intensity of current cigarette smoking) are correlated with bone density in smokers. A previous study found a modest correlation ($r = -0.3$) between reported smoking (in cigarettes per day) and bone density; the investigator anticipates that urinary cotinine levels will be at least as well correlated. How many smokers will need to be enrolled, at α (two-sided) = 0.05 and $\beta = 0.10$?

Solution: The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** *There is no correlation between urinary cotinine level and bone density in smokers.*
   Alternative Hypothesis: *There is a correlation between urinary cotinine level and bone density in smokers.*
2. *Effect size* $(r) = |-0.3| = 0.3$.
3. α *(two-sided)* $= 0.05$; $\beta = 0.10$.

Using Table 6C, reading across from $r = 0.30$ in the leftmost column and down from α (two-sided) = 0.05 and $\beta = 0.10$, 113 smokers will be required.

OTHER CONSIDERATIONS AND SPECIAL ISSUES

Dropouts

Each sampling unit must be available for analysis; subjects who are enrolled in a study but in whom outcome status cannot be ascertained (such as **dropouts**) do not count in the sample size. If the investigator anticipates that any of her subjects will not be available for follow-up, she should increase the size of the enrolled sample accordingly. If, for example, the investigator estimates that 20% of her sample will be lost to follow-up, then the sample size should be increased by a factor of $(1 \div [1 - 0.20])$, or 1.25.

Categorical Variables

Ordinal variables can often be treated as continuous variables, especially if the number of categories is relatively large (six or more) and if averaging the values of the variable makes sense. In other situations, the best strategy is to change the research hypothesis slightly by dichotomizing the categorical variable. As an example, suppose a researcher is studying whether the sex of a diabetic patient is associated with the number of times the patient visits a podiatrist in a year. The number of visits is unevenly distributed: many people will have no visits, some will make one visit, and only a few will make two or more visits. In this situation, the investigator could estimate the sample size as if the outcome were dichotomous (no visits versus one or more visits).

Survival Analysis

When an investigator wishes to compare which of two treatments is more effective in prolonging life or in reducing the symptomatic phase of a disease, survival analysis will be the appropriate technique for analyzing the data (2,3). Although the outcome variable, say weeks of survival, appears to be continuous, the t test is not appropriate because what is actually being assessed is not time (a continuous variable) but the proportion of subjects (a dichotomous variable) still alive at each point in time. A reasonable approximation can be made by dichotomizing the outcome variable at the end of the anticipated follow-up period (e.g., the proportion surviving for 6 months or more), and estimating the sample size with the chi-squared test.

Clustered Samples

Some research designs involve the use of **clustered samples,** in which subjects are sampled by groups (Chapter 11). Consider, for example, a study of whether an educational intervention directed at clinicians improves the rate of smoking cessation among their patients. Suppose that 20 physicians are randomly assigned to the group that receives the intervention and 20 physicians are assigned to a control group. One year later, the investigators plan to review the charts of a random sample of 50 patients who had been smokers at baseline in each practice to determine how many have quit smoking. Does the sample size equal 40 (the number of physicians) or 2,000 (the number of patients)? The answer, which lies somewhere in between those two extremes, depends upon how similar the patients within a physician's practice are (in terms of their likelihood of smoking cessation) compared with the similarity among all the patients. Estimating this quantity often requires obtaining pilot data, unless another investigator has previously done a similar study. There are several techniques for estimating the required sample size for a study using clustered samples (4–7), but they are challenging and usually require the assistance of a statistician.

Matching

For a variety of reasons (Chapter 9), an investigator may choose to use a matched design. The techniques in this chapter, which ignore any matching, nevertheless provide reasonable estimates of the required sample size. More precise estimates can be made using standard approaches (8) or an interactive Web-based program.

Multivariate Adjustment and Other Special Statistical Analyses

When designing an observational study, an investigator may decide that one or more variables will confound the association between the predictor and outcome (Chapter 9), and plan to use statistical techniques to adjust for these **confounders** when she analyzes her results. When this adjustment will be included in testing the primary hypothesis, the estimated sample size needs to take this into account.

Analytic approaches that adjust for confounding variables often increase the required sample size (9,10). The magnitude of that increase depends on several factors, including the prevalence of the confounder, the strength of the association between the predictor and the confounder, and the strength of the association between the confounder and the outcome. These effects are complex and no general rule covers all situations.

Statisticians have developed multivariate methods such as linear regression and logistic regression that allow the investigator to adjust for confounding variables. One widely used statistical technique, **Cox proportional hazards** analysis, can adjust both for confounders and for differences in length of follow-up. If one of these techniques is going to be used to analyze the data, there are corresponding approaches for estimating the required sample size (3,11–14). Sample size techniques are also available for other designs, such as studies of potential genetic risk factors or candidate genes (15–17), economic studies (18–20), dose–response studies (21), or studies that involve more than two groups (22). Again, the Internet is a useful resource for these more sophisticated approaches (e.g., search for "sample size" and "logistic regression").

It is usually easier, at least for novice investigators, to estimate the sample size assuming a simpler method of analysis, such as the chi-squared test or the t test. Suppose, for example, an investigator is planning a case–control study of whether serum cholesterol level (a continuous variable) is associated with the occurrence of brain tumors (a dichotomous variable). Even if the eventual plan is to analyze the data with the logistic regression technique, a ballpark sample size can be estimated with the t test. It turns out that the simplified approaches usually produce sample size estimates that are similar to those generated by more sophisticated techniques. An experienced statistician may need to be consulted, however, if a grant proposal that involves substantial costs is being submitted for funding: grant reviewers will expect you to use a sophisticated approach even if they accept that the sample size estimates are based on guesses about the risk of the outcome, the effect size, and so on.

Equivalence Studies

Sometimes the goal of a study is to show that the null hypothesis is correct and that there really is no substantial association between the predictor and outcome variables (23–26). A common example is a clinical trial to test whether a new drug is as effective as an established drug. This situation poses a challenge when planning sample size, because the desired effect size is zero (i.e., the investigator would like to show that the two drugs are equally effective).

One acceptable method is to design the study to have substantial power (say, 0.90 or 0.95) to reject the null hypothesis when the effect size is small enough that it would not be clinically important (e.g., a difference of 5 mg/dL in mean fasting glucose levels). If the results of such a well-powered study show "no effect" (i.e., the 95% confidence interval excludes the prespecified difference of 5 mg/dL), then the investigator can be reasonably sure that the two drugs have similar effects. One problem with equivalence studies, however, is that the additional power and the small effect size often require a very large sample size.

Another problem involves the loss of the usual safeguards that are inherent in the paradigm of the null hypothesis, which protects a conventional study, such as one that compares an active drug with a placebo, against Type I errors (falsely rejecting the null hypothesis). The paradigm ensures that many problems in the design or execution of a study, such as using imprecise measurements or inadequate numbers of subjects, make it harder to reject the null hypothesis. Investigators in a conventional study, who are trying to reject a null hypothesis, have a strong incentive to do the best possible study. The same is not true for an equivalence study, in which the goal is to find no difference, and the safeguards do not apply.

SAMPLE SIZE TECHNIQUES FOR DESCRIPTIVE STUDIES

Estimating the sample size for descriptive studies, including studies of diagnostic tests, is based on somewhat different principles. Such studies do not have predictor and outcome variables, nor do they compare different groups. Therefore the concepts of power and the null and alternative hypotheses do not apply. Instead, the investigator calculates descriptive statistics, such as means and proportions. Often, however, descriptive studies (*What is the prevalence of depression among elderly patients in a medical clinic?*) are also used to ask analytic questions (*What are the predictors of depression among these patients?*). In this situation, sample size should be estimated for the analytic study as well, to avoid the common problem of having inadequate power for what turns out to be the question of greater interest.

Descriptive studies commonly report **confidence intervals,** a range of values about the sample mean or proportion. A confidence interval is a measure of the precision of a sample estimate. The investigator sets the confidence level, such as

95% or 99%. An interval with a greater confidence level (say 99%) is wider, and therefore more likely to include the true population value, than an interval with a lower confidence level (90%).

The width of a confidence interval depends on the sample size. For example, an investigator might wish to estimate the mean score on the U.S. Medical Licensing Examination in a group of medical students. From a sample of 200 students, she might estimate that the mean score in the population of all students is 215, with a 95% confidence interval from 210 to 220. A smaller study, say with 50 students, might have about the same mean score but would almost certainly have a wider 95% confidence interval.

When estimating sample size for descriptive studies, the investigator specifies the desired level and width of the confidence interval. The sample size can then be determined from the tables or formulas in the appendix.

Continuous Variables

When the variable of interest is continuous, a confidence interval around the mean value of that variable is often reported. To estimate the sample size for that confidence interval, the investigator must

1. Estimate the standard deviation of the variable of interest.
2. Specify the desired precision (total width) of the confidence interval.
3. Select the confidence level for the interval (e.g., 95%, 99%).

To use Appendix 6D, standardize the total width of the interval (divide it by the standard deviation of the variable), then look down the leftmost column of Table 6D for the expected standardized width. Next, read across the table to the chosen confidence level for the required sample size.

Example 6.4 Calculating Sample Size for a Descriptive Study of a Continuous Variable

Problem: The investigator seeks to determine the mean IQ among third graders in an urban area with a 99% confidence interval of ±3 points. A previous study found that the standard deviation of IQ in a similar city was 15 points.

Solution: The ingredients for the sample size calculation are as follows:

1. *Standard deviation of variable (SD) = 15 points.*
2. *Total width of interval = 6 points (3 points above and 3 points below). The standardized width of interval = total width ÷ SD = 6 ÷ 15 = 0.4.*
3. *Confidence level = 99%.*

Reading across from a standardized width of 0.4 in the leftmost column of Table 6D and down from the 99% confidence level, the required sample size is 166 third graders.

Dichotomous Variables

In a descriptive study of a dichotomous variable, results can be expressed as a confidence interval around the estimated proportion of subjects with one of the values.

This includes studies of the **sensitivity** or **specificity** of a diagnostic test, which appear at first glance to be continuous variables but are actually dichotomous—proportions expressed as percentages (Chapter 12). To estimate the sample size for that confidence interval, the investigator must

1. Estimate the expected proportion with the variable of interest in the population. (If more than half of the population is expected to have the characteristic, then plan the sample size based on the proportion expected not to have the characteristic.)
2. Specify the desired precision (total width) of the confidence interval.
3. Select the confidence level for the interval (e.g., 95%).

In Appendix 6E, look down the leftmost column of Table 6E for the expected proportion with the variable of interest. Next, read across the table to the chosen width and confidence level, yielding the required sample size.

Example 6.5 provides a sample size calculation for studying the sensitivity of a diagnostic test, which yields the required number of subjects with the disease. When studying the specificity of the test, the investigator must estimate the required number of subjects who do *not* have the disease. There are also techniques for estimating the sample size for studies of receiver operating characteristic (ROC) curves (27), likelihood ratios (28), and reliability (29) (Chapter 12).

Example 6.5 Calculating Sample Size for a Descriptive Study of a Dichotomous Variable

Problem: The investigator wishes to determine the sensitivity of a new diagnostic test for pancreatic cancer. Based on a pilot study, she expects that 80% of patients with pancreatic cancer will have positive tests. How many such patients will be required to estimate a 95% confidence interval for the test's sensitivity of 0.80 ± 0.05?

Solution: The ingredients for the sample size calculation are as follows:

1. *Expected proportion = 0.20. (Because 0.80 is more than half, sample size is estimated from the proportion expected to have a negative result, that is, 0.20.)*
2. *Total width = 0.10 (0.05 below and 0.05 above).*
3. *Confidence level = 95%.*

Reading across from 0.20 in the leftmost column of Table 6E and down from a total width of 0.10, the middle number (representing a 95% confidence level) yields the required sample size of 246 patients with pancreatic cancer.

WHAT TO DO WHEN SAMPLE SIZE IS FIXED

Especially when doing secondary data analysis, the sample size may have been determined before you design your study. In this situation, or if the number of participants who are available or affordable for study is limited, the investigator must work backward from the fixed sample size. She estimates the effect size that can be detected at a given power (usually 80%) or, less commonly, the power to detect a given effect. The investigator can use the sample size tables in the chapter appendixes, interpolating when necessary, or use the sample size formulas in the appendixes for estimating the effect size.

A good general rule is that a study should have a power of 80% or greater to detect a reasonable effect size. It is often tempting to pursue research hypotheses that have less power if the cost of doing so is small, such as when doing an analysis of data that have already been collected. The investigator should keep in mind, however, that she might face the difficulty of interpreting (and publishing) a study that may have found no effect because of insufficient power; the broad confidence intervals will reveal the possibility of a substantial effect in the population from which the small study sample was drawn.

Example 6.6 Calculating the Detectable Effect Size When Sample Size is Fixed

Problem: An investigator determines that there are 100 patients with systemic lupus erythematosus (SLE) who might be willing to participate in a study of whether a 6-week meditation program affects disease activity, as compared with a control group that receives a pamphlet describing relaxation. If the standard deviation of the change in a validated SLE disease activity scale score is expected to be five points in both the control and the treatment groups, what size difference will the investigator be able to detect between the two groups, at α (two-sided) = 0.05 and β = 0.20?

Solution: In Table 6A, reading down from α (two-sided) = 0.05 and β = 0.20 (the rightmost column in the middle triad of numbers), 45 patients per group are required to detect a standardized effect size of 0.6, which is equal to three points (0.6 × 5 points). The investigator (who will have about 50 patients per group) will be able to detect a difference of a little less than three points between the two groups.

STRATEGIES FOR MINIMIZING SAMPLE SIZE AND MAXIMIZING POWER

When the estimated sample size is greater than the number of subjects that can be studied realistically, the investigator should proceed through several steps. First, the calculations should be checked, as it is easy to make mistakes. Next, the "ingredients" should be reviewed. Is the effect size unreasonably small or the variability unreasonably large? Could α or β, or both, be increased without harm? Would a one-sided alternative hypothesis be adequate? Is the confidence level too high or the interval unnecessarily narrow?

These technical adjustments can be useful, but it is important to realize that statistical tests ultimately depend on the information contained in the data. Many changes in the ingredients, such as reducing power from 90% to 80%, do not improve the quantity or quality of the data that will be collected. There are, however, several strategies for reducing the required sample size or for increasing power for a given sample size that actually increase the information content of the collected data. Many of these strategies involve modifications of the research hypothesis; the investigator should carefully consider whether the new hypothesis still answers the research question that she wishes to study.

Use Continuous Variables

When continuous variables are an option, they usually permit smaller sample sizes than dichotomous variables. Blood pressure, for example, can be expressed either as millimeters of mercury (continuous) or as the presence or absence of hypertension (dichotomous). The former permits a smaller sample size for a given power or a greater power for a given sample size.

In Example 6.7, the continuous outcome addresses the effect of nutrition supplements on muscle strength among the elderly. The dichotomous outcome is concerned with its effects on the proportion of subjects who have at least a minimal amount of strength, which may be a more valid surrogate for potential fall-related morbidity.

Example 6.7 Use of Continuous versus Dichotomous Variables

Problem: Consider a placebo-controlled trial to determine the effect of nutrition supplements on strength in elderly nursing home residents. Previous studies have established that quadriceps strength (as peak torque in newton-meters) is approximately normally distributed, with a mean of 33 N·m and a standard deviation of 10 N·m, and that about 10% of the elderly have very weak muscles (strength < 20 N·m). Nutrition supplements for 6 months are anticipated to increase strength by 5 N·m as compared with the usual diet. This change in mean strength can be estimated, based on the distribution of quadriceps strength in the elderly, to correspond to a reduction in the proportion of the elderly who are very weak to about 5%.

One design might treat strength as a dichotomous variable: very weak versus not very weak. Another might use all the information contained in the measurement and treat strength as a continuous variable. How many subjects would each design require at α (two-sided) = 0.05 and β = 0.20? How does the change in design affect the research question?

Solution: The ingredients for the sample size calculation using a dichotomous outcome variable (very weak versus not very weak) are as follows:

1. **Null Hypothesis:** *The proportion of elderly nursing home residents who are very weak (peak quadriceps torque <20 N·m) after receiving 6 months of nutrition supplements is the same as the proportion who are very weak in those on a usual diet.* **Alternative Hypothesis:** *The proportion of elderly nursing home residents who are very weak (peak quadriceps torque <20 N·m) after receiving 6 months of nutrition supplements differs from the proportion in those on a usual diet.*
2. P_1 *(prevalence of being very weak on usual diet) = 0.10;* P_2 *(in supplement group) = 0.05. The smaller of these values is 0.05, and the difference between them* $(P_1 - P_2)$ *is 0.05.*
3. α *(two-sided) = 0.05;* β *= 0.20.*

Using Table 6B.1, reading across from 0.05 in the leftmost column and down from an expected difference of 0.05, the middle number (for α [two-sided] = 0.05 and β = 0.20), this design would require 473 subjects per group.

*The ingredients for the sample size calculation using a **continuous outcome variable** (quadriceps strength as peak torque) are as follows:*

1. **Null Hypothesis:** *Mean quadriceps strength (as peak torque in N·m) in elderly nursing home residents after receiving 6 months of nutrition supplements is the same as mean quadriceps strength in those on a usual diet.*

Alternative Hypothesis: *Mean quadriceps strength (as peak torque in N·m) in elderly nursing home residents after receiving 6 months of nutrition supplements differs from mean quadriceps strength in those on a usual diet.*
2. *Effect size = 5 N·m*
3. *Standard deviation of quadriceps strength = 10 N·m*
4. *Standardized effect size = effect size ÷ standard deviation = 5 N·m ÷ 10 N·m = 0.5.*
5. *α (two-sided) 0.05; β = 0.20.*

Using Table 6A, reading across from a standardized effect size of 0.50, with α (two-sided) = 0.05 and β = 0.20, this design would require about 64 subjects in each group. (In this example, the shortcut sample size estimate from page 68 of 16 ÷ (standardized effect size)², or 16 ÷ 0.5² gives the same estimate of 64 subjects per group.) The bottom line is that the use of an outcome variable that was continuous rather than dichotomous meant that a substantially smaller sample size needed to study this research question

Use Paired Measurements

In some experiments or cohort studies with continuous outcome variables, paired measurements—one at baseline, another at the conclusion of the study—can be made in each subject. The outcome variable is the change between these two measurements. In this situation, a *t* test on the paired measurements can be used to compare the mean value of this change in the two groups. This technique often permits a smaller sample size because, by comparing each subject with herself, it removes the baseline between-subject part of the variability of the outcome variable. For example, the change in weight on a diet has less variability than the final weight, because final weight is highly correlated with initial weight. Sample size for this type of *t* test is estimated in the usual way, except that the standardized effect size (*E*/*S* in Table 6A) is the anticipated difference in the *change* in the variable divided by the standard deviation *of that change.*

Example 6.8 Use of the *t* Test with Paired Measurements

Problem: Recall Example 6.1, in which the investigator studying the treatment of asthma is interested in determining whether salbutamol can improve FEV_1 by 200 mL compared with ipratropium bromide. Sample size calculations indicated that 394 subjects per group are needed, more than are likely to be available. Fortunately, a colleague points out that asthmatic patients have great differences in their FEV_1 values before treatment. These between-subject differences account for much of the variability in FEV_1 after treatment, therefore obscuring the effect of treatment. She suggests using a paired t test to compare the changes in FEV_1 in the two groups. A pilot study finds that the standard deviation of the change in FEV_1 is only 250 mL. How many subjects would be required per group, at α (two-sided) = 0.05 and β = 0.20?

Solution: The ingredients for the sample size calculation are as follows:

1. Null Hypothesis: *Change in mean FEV_1 after 2 weeks of treatment is the same in asthmatic patients treated with salbutamol as it is in those treated with ipratropium bromide.*
   Alternative Hypothesis: *Change in mean FEV_1 after 2 weeks of treatment is different in asthmatic patients treated with salbutamol from what it is in those treated with ipratropium bromide.*
2. *Effect size = 200 mL.*
3. *Standard deviation of the outcome variable = 250 mL.*
4. *Standardized effect size = effect size ÷ standard deviation = 200 mL ÷ 250 mL = 0.8.*
5. *α (two-sided) = 0.05; β = 1 − 0.80 = 0.20.*

Using Table 6A, this design would require about 26 participants per group, a much more reasonable sample size than the 394 per group in "Example 6.1". In this example, the shortcut sample size estimate of 16 ÷ (standardized effect size)², or 16 ÷ 0.8² gives a similar estimate of 25 subjects per group.

A Brief Technical Note. This chapter always refers to **two-sample *t* tests,** which are used when comparing the mean values of an outcome variable in two groups of subjects. A two-sample *t* test can be **unpaired,** if the outcome variable itself is being compared between two groups (see "Example 6.1"), or **paired** if the outcome is the change in a pair of measurements, say before and after an intervention (see "Example 6.8").

A third type of *t* test, the ***one-sample* paired *t* test,** compares the mean change in a pair of measurements within a single group to zero change. This type of analysis is reasonably common in time series designs (Chapter 10), a before–after approach to examining treatments that are difficult to randomize (for example, the effect of elective hysterectomy, a decision few women are willing to leave to a coin toss, on quality of life). However, it is a fairly weak design because the absence of a comparison group makes it difficult to know what would have happened had the subjects been left untreated (Chapter 10). When planning a study that will be analyzed with a one-sample paired *t* test, the sample size in Appendix 6A represents the *total* number of subjects (because there is only one group). Appendix 6F presents additional information on the use and misuse of one- and two-sample *t* tests.

Use More Precise Variables

Because they reduce variability, more precise variables permit a smaller sample size in both analytic and descriptive studies. Even a modest change in precision can have a substantial effect on sample size. For example, when using the *t* test to estimate sample size, a 20% decrease in the standard deviation of the outcome variable results in a 36% decrease in the sample size. Techniques for increasing the precision of a variable, such as making measurements in duplicate, are presented in Chapter 4.

Use Unequal Group Sizes

Because an equal number of subjects in each of two groups usually gives the greatest power for a given total number of subjects, Tables 6A, 6B.1, and 6B.2 in the

appendixes assume equal sample sizes in the two groups. Sometimes, however, the distribution of subjects is not equal in the two groups, or it is easier or less expensive to recruit study subjects for one group than the other. It may turn out, for example, that an investigator wants to estimate sample size based on the 30% of the subjects in a cohort who smoke cigarettes (compared with 70% who do not smoke). Or, in a case–control study, the number of persons with the disease may be small, but it may be possible to sample a much larger number of controls. In general, the gain in power when the size of one group is increased to twice the size of the other is considerable; tripling and quadrupling one of the groups provide progressively smaller gains. Sample sizes for unequal groups can be computed from the formulas found in the text to Appendixes 6A and 6B or from the Web.

Here is a useful approximation for estimating sample size for case–control studies of dichotomous risk factors and outcomes using c controls per case. If n represents the number of cases that would have been required for one control per case (at a given α, β, and effect size), then the approximate number of cases (n') with cn' controls that will be required is

$$n' = [(c + 1) \div 2c] \times n.$$

For example, with $c = 2$ controls per case, then $[(2 + 1) \div (2 \times 2)] \times n = 3/4 \times n$, and only 75% as many cases are needed. As c gets larger, n' approaches 50% of n (when $c = 10$, for example, $n' = 11/20 \times n$).

Example 6.9 Use of Multiple Controls per Case in a Case–Control Study

Problem: An investigator is studying whether exposure to household insecticide is a risk factor for aplastic anemia. The original sample size calculation indicated that 25 cases would be required, using one control per case. Suppose that the investigator has access to only 18 cases. How should the investigator proceed?

Solution: The investigator should consider using multiple controls per case (after all, she can find many patients who do not have aplastic anemia). By using three controls per case, for example, the approximate number of cases that will be required is $[(3 + 1) \div (2 \times 3)] \times 25 = 17$.

Use a More Common Outcome

When the outcome is dichotomous, using a more frequent outcome, up to a frequency of 0.5, is usually one of the best ways to increase power: if an outcome occurs more often, there is more of a chance to detect its predictors. Power actually depends more on the number of subjects with a specified outcome than it does on the total number of subjects in the study. Studies with rare outcomes, like the occurrence of breast cancer in healthy women, require very large sample sizes to have adequate power.

One of the best ways to make an outcome more common is to enroll subjects at greater risk of developing that outcome (such as women with a family history of breast cancer). Others are to extend the follow-up period, so that there is more time to accumulate outcomes, or to loosen the definition of what constitutes an outcome (e.g., by including ductal carcinoma *in situ*). All these techniques, however, may change the research question, so they should be used with caution.

Example 6.10 Use of a More Common Outcome

Problem: Suppose an investigator is comparing the efficacy of an antiseptic gargle versus a placebo gargle in preventing upper respiratory infections. Her initial calculations indicated that her anticipated sample of 200 volunteer college students was inadequate, in part because she expected that only about 20% of her subjects would have an upper respiratory infection during the 3-month follow-up period. Suggest a few changes in the study plan.

Solution: Here are two possible solutions: (a) study a sample of pediatric interns and residents, who are likely to experience a much greater incidence of upper respiratory infections than college students; or (b) follow the sample for a longer period of time, say 6 or 12 months. Both of these solutions involve modification of the research hypothesis, but neither change seems sufficiently large to affect the overall research question about the efficacy of antiseptic gargle.

HOW TO ESTIMATE SAMPLE SIZE WHEN THERE IS INSUFFICIENT INFORMATION

Often the investigator finds that she is missing one or more of the ingredients for the sample size calculation and becomes frustrated in her attempts to plan the study. This is an especially frequent problem when the investigator is using an instrument of her design (such as a new questionnaire on quality of life in patients with urinary incontinence). How should she go about deciding what effect size or standard deviation to use?

The first strategy is an **extensive search** for previous and related findings on the topic and on similar research questions. Roughly comparable situations and mediocre or dated findings may be good enough. (For example, are there data on quality of life among patients with other urologic problems, or with related conditions like having a colostomy?) If the literature review is unproductive, she should contact other investigators about their judgment on what to expect, and whether they are aware of any unpublished results that may be relevant. If there is still no information available, she may consider doing a small **pilot study** or obtaining a data set for a secondary analysis to obtain the missing ingredients before embarking on the main study. (Indeed, a pilot study is highly recommended for almost all studies that involve new instruments, measurement methods, or recruitment strategies. They save time in the end by enabling investigators to do a much better job planning the main study). Pilot studies are useful for estimating the standard deviation of a measurement, or the proportion of subjects with a particular characteristic. Another trick is to recognize that for continuous variables that have a roughly bell-shaped distribution, the **standard deviation** can be estimated as one-quarter of the difference between the high and low ends of the range of values that occur commonly, ignoring extreme values. For example, if most subjects are likely to have a serum sodium level between 135 and 143 mEq/L, the standard deviation of serum sodium is about 2 mEq/L ($1/4 \times 8$ mEq/L).

Alternatively, the investigator can determine the detectable effect size based on a value that she considers to be **clinically meaningful.** For example, *suppose that an investigator is studying a new invasive treatment for severe refractory gastroparesis, a*

condition in which at most 5% of patients improve spontaneously. If the treatment is shown to be effective, she thinks that gastroenterologists would be willing to treat up to five patients to produce a sustained benefit in one of those patients (because the treatment has substantial side effects and is expensive, she doesn't think that the number would be more than 5). A number needed to treat (NNT) of 5 corresponds to a risk difference of 20% (NNT = 1/risk difference), so the investigator should estimate the sample size based on a comparison of P1 = 5% versus P2 = 25% (i.e., 59 subjects per group at a power of 0.80 and a two-sided α of 0.05).

Another strategy, when the mean and standard deviation of a continuous or categorical variable are in doubt, is to **dichotomize** that variable. Categories can be lumped into two groups, and continuous variables can be split at their mean or median. For example, dividing quality of life into "better than the median" or "the median or less" avoids having to estimate its standard deviation in the sample, although one still has to estimate what proportions of subjects would be above the median in the two groups being studied. The chi-squared statistic can then be used to make a reasonable, albeit somewhat high, estimate of the sample size.

If all this fails, the investigator should just make an **educated guess** about the likely values of the missing ingredients. The process of thinking through the problem and imagining the findings will often result in a reasonable estimate, and that is what sample size planning is about. This is usually a better option than just deciding to design the study to have 80% power at a two-sided α of 0.05 to detect a standardized effect size of, say, 0.5 between the two groups ($n = 64$, per group, by the way). Very few grant reviewers will accept that sort of arbitrary decision.

COMMON ERRORS TO AVOID

Many inexperienced investigators (and some experienced ones!) make mistakes when planning sample size. A few of the more common ones follow:

1. The most common error is estimating the sample size **late** during the design of the study. Do it early in the process, when fundamental changes can still be made.

2. Dichotomous variables can appear to be continuous when they are expressed as a **percentage** or **rate.** For example, vital status (alive or dead) might be misinterpreted as continuous when expressed as percent alive. Similarly, in survival analysis a dichotomous outcome can appear to be continuous (e.g., median survival in months). For all of these, the outcome itself is actually dichotomous and the appropriate simple approach in planning sample size would be the chi-squared test.

3. The sample size estimates the number of subjects with outcome data, not the number who need to be enrolled. The investigator should always plan for **dropouts** and subjects with **missing data.**

4. The tables at the end of the chapter assume that the two groups being studied have equal sample sizes. Often that is not the case; for example, a cohort study of whether use of vitamin supplements reduces the risk of sunburn would probably not enroll equal numbers of subjects who used, or did not use, vitamins. If the sample sizes are **not equal,** then the formulas that follow the tables or the Web should be used.

5. When using the t test to estimate the sample size, what matters is the standard deviation of the outcome variable. Therefore if the outcome is **change** in a continuous variable, the investigator should use the standard deviation of that change rather than the standard deviation of the variable itself.

6. Be aware of **clustered** data. If there appear to be two "levels" of sample size (e.g., one for physicians and another for patients), clustering is a likely problem and the tables in the appendices do not apply.

SUMMARY

1. When estimating sample size for an **analytic study,** the following steps need to be taken: (a) state the **null** and **alternative hypotheses,** specifying the number of **sides;** (b) select a **statistical test** that could be used to analyze the data, based on the **types of predictor and outcome variables;** (c) estimate the **effect size** (and its **variability,** if necessary); and (d) specify appropriate values for α and β, based on the importance of avoiding **Type I** and **Type II errors.**

2. Other considerations in calculating sample size for analytic studies include adjusting for potential **dropouts,** and strategies for dealing with **categorical variables, survival analysis, clustered samples, multivariate adjustment,** and **equivalence studies.**

3. The steps for estimating sample size for **descriptive studies,** which do not have hypotheses, are to (a) estimate the **proportion** of subjects with a dichotomous outcome or the **standard deviation** of a continuous outcome; (b) specify the desired **precision** (width of the confidence interval); and (c) specify the **confidence level** (e.g., 95%).

4. When sample size is predetermined, the investigator can work backward to estimate the detectable **effect size** or, less commonly, the **power.**

5. Strategies **to minimize** the required sample size include using **continuous** variables, more **precise** measurements, **paired** measurements, **unequal group sizes,** and **more common outcomes.**

6. When there seems to be not enough information to estimate the sample size, the investigator should **review the literature** in related areas, do a small **pilot study** or choose an **effect size** that is clinically meaningful; **standard deviation** can be estimated as 1/4 of the range of commonly encountered values. If none of these is feasible, an educated guess can give a useful ballpark estimate.

APPENDIX 6A

Sample Size Required per Group When Using the *t* Test to Compare Means of Continuous Variables

TABLE 6A Sample Size *per Group* for Comparing Two Means

One-sided α =	0.005			0.025			0.05		
Two-sided α =	0.01			0.05			0.10		
*E/S** β =	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
0.10	3,565	2,978	2,338	2,600	2,103	1,571	2,166	1,714	1,238
0.15	1,586	1,325	1,040	1,157	935	699	963	762	551
0.20	893	746	586	651	527	394	542	429	310
0.25	572	478	376	417	338	253	347	275	199
0.30	398	333	262	290	235	176	242	191	139
0.40	225	188	148	164	133	100	136	108	78
0.50	145	121	96	105	86	64	88	70	51
0.60	101	85	67	74	60	45	61	49	36
0.70	75	63	50	55	44	34	45	36	26
0.80	58	49	39	42	34	26	35	28	21
0.90	46	39	21	34	27	21	28	22	16
1.00	38	32	26	27	23	17	23	18	14

* E/S is the standardized effect size, computed as E (expected effect size) divided by S (SD of the outcome variable). To estimate the sample size, read across from the *standardized effect size*, and down from the specified values of α and β for the required sample size in each group.

Calculating Variability

Variability is usually reported as either the standard deviation or the standard error of the mean (SEM). For the purposes of sample size calculation, the standard deviation of the variable is most useful. Fortunately, it is easy to convert from one measure to another: the standard deviation is simply the standard error times the square root of N, where N is the number of subjects that makes up the mean. Suppose a study reported that the weight loss in 25 persons on a low-fiber diet was 10 ± 2 kg (mean $\pm$ SEM). The standard deviation would be $2 \times \sqrt{25} = 10$ kg.

General Formula for Other Values

The general formula for other values of E, S, α, and β, or for unequal group sizes, is as follows. Let:

z_α = the standard normal deviate for α (If the alternative hypothesis is two-sided, $z_\alpha = 2.58$when $\alpha = 0.01$, $z_\alpha = 1.96$ when $\alpha = 0.05$, and $z_\alpha = 1.645$ when $\alpha = 0.10$. If the alternative hypothesis is one-sided, $z_\alpha = 1.645$ when $\alpha = 0.05$.)
z_β = the standard normal deviate for β ($z_\beta = 0.84$ when $\beta = 0.20$, and $z_\beta = 1.282$when $\beta = 0.10$)
q_1 = proportion of subjects in group 1
q_2 = proportion of subjects in group 2
N = total number of subjects required

Then:

$$N = [(1/q_1 + 1/q_2)S^2(z_\alpha + z_\beta)^2] \div E^2.$$

Readers who would like to skip the work involved in hand calculations with this formula can get an instant answer from a calculator on our website (www.epibiostat.ucsf.edu/dcr/)(Because this formula is based on approximating the t statistic with a z statistic, it will slightly underestimate the sample size when N is less than about 30. Table 6A uses the t statistic to estimate sample size.)

APPENDIX 6B

Sample Size Required per Group When Using the Chi-Squared Statistic or *Z* Test to Compare Proportions of Dichotomous Variables

TABLE 6B.1 Sample Size *per Group* for Comparing Two Proportions

Upper number: $\alpha = 0.05$ (one-sided) or $\alpha = 0.10$ (two-sided); $\beta = 0.20$
Middle number: $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided); $\beta = 0.20$
Lower number: $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided); $\beta = 0.10$

Smaller of P_1 and P_2*	Difference Between P_1 and P_2									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.05	381	129	72	47	35	27	22	18	15	13
	473	159	88	59	43	33	26	22	18	16
	620	207	113	75	54	41	33	27	23	19
0.10	578	175	91	58	41	31	24	20	16	14
	724	219	112	72	51	37	29	24	20	17
	958	286	146	92	65	48	37	30	25	21
0.15	751	217	108	67	46	34	26	21	17	15
	944	270	133	82	57	41	32	26	21	18
	1,252	354	174	106	73	53	42	33	26	22
0.20	900	251	121	74	50	36	28	22	18	15
	1,133	313	151	91	62	44	34	27	22	18
	1,504	412	197	118	80	57	44	34	27	23
0.25	1,024	278	132	79	53	38	29	23	18	15
	1,289	348	165	98	66	47	35	28	22	18
	1,714	459	216	127	85	60	46	35	28	23
0.30	1,123	300	141	83	55	39	29	23	18	15
	1,415	376	175	103	68	48	36	28	22	18
	1,883	496	230	134	88	62	47	36	28	23
0.35	1,197	315	146	85	56	39	29	23	18	15
	1,509	395	182	106	69	48	36	28	22	18
	2,009	522	239	138	90	62	47	35	27	22
0.40	1,246	325	149	86	56	39	29	22	17	14
	1,572	407	186	107	69	48	35	27	21	17
	2,093	538	244	139	90	62	46	34	26	21
0.45	1,271	328	149	85	55	38	28	21	16	13
	1,603	411	186	106	68	47	34	26	20	16
	2,135	543	244	138	88	60	44	33	25	19
0.50	1,271	325	146	83	53	36	26	20	15	–
	1,603	407	182	103	66	44	32	24	18	–
	2,135	538	239	134	85	57	42	30	23	–
0.55	1,246	315	141	79	50	34	24	18	–	–
	1,572	395	175	98	62	41	29	22	–	–
	2,093	522	230	127	80	53	37	27	–	–

TABLE 6B.1 (Continued)

Upper number: $\alpha = 0.05$ (one-sided) or $\alpha = 0.10$ (two-sided); $\beta = 0.20$
Middle number: $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided); $\beta = 0.20$
Lower number: $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided); $\beta = 0.10$

Smaller of P_1 and P_2*	Difference Between P_1 and P_2									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.60	1,197	300	132	74	46	31	22	–	–	–
	1,509	376	165	91	57	37	26	–	–	–
	2,009	496	216	118	73	48	33	–	–	–
0.65	1,123	278	121	67	41	27	–	–	–	–
	1,415	348	151	82	51	33	–	–	–	–
	1,883	459	197	106	65	41	–	–	–	–
0.70	1,024	251	108	58	35	–	–	–	–	–
	1,289	313	133	72	43	–	–	–	–	–
	1,714	412	174	92	54	–	–	–	–	–
0.75	900	217	91	47	–	–	–	–	–	–
	1,133	270	112	59	–	–	–	–	–	–
	1,504	354	146	75	–	–	–	–	–	–
0.80	751	175	72	–	–	–	–	–	–	–
	944	219	88	–	–	–	–	–	–	–
	1,252	286	113	–	–	–	–	–	–	–
0.85	578	129	–	–	–	–	–	–	–	–
	724	159	–	–	–	–	–	–	–	–
	958	207	–	–	–	–	–	–	–	–
0.90	381	–	–	–	–	–	–	–	–	–
	473	–	–	–	–	–	–	–	–	–
	620	–	–	–	–	–	–	–	–	–

The one-sided estimates use the z statistic.

* P_1 represents the proportion of subjects expected to have the outcome in one group; P_2 in the other group. (In a case-control study, P_1 represents the proportion of cases with the predictor variable; P_2 the proportion of controls with the predictor variable.) To estimate the sample size, read across from the smaller of P_1 and P_2, and down the expected *difference* between P_1 and P_2. The three numbers represent the sample size required in each group for the specified values of α and β.

Additional detail for P_1 and P_2 between 0.01 and 0.10 is given in Table 6B.2.

General Formula for Other Values

The general formula for calculating the *total* sample size (N) required for a study using the z statistic, where P_1 and P_2 are defined above, is as follows (see Appendix 6A for definitions of Z_α and Z_β). Let

q_1 = proportion of subjects in group 1
q_2 = proportion of subjects in group 2
N = *total* number of subjects
$P = q_1 P_1 + q_2 P_2$

Then

$$N = \frac{[z_\alpha \sqrt{P(1-P)(1/q_1 + 1/q_2)} + z_\beta \sqrt{P_1(1-P_1)(1/q_1) + P_2(1-P_2)(1/q_2)}]^2}{(P_1 - P_2)^2}.$$

Readers who would like to skip the work involved in hand calculations with this formula can get an instant answer from a calculator on our website (www.epibiostat.ucsf.edu/dcr/) (This formula does not include the Fleiss-Tytun-Ury continuity correction and therefore underestimates the required sample size by up to about 10%. Tables 6B.1 and 6B.2 do include this continuity correction.)

TABLE 6B.2 Sample Size *per Group* for Comparing Two Proportions, the Smaller of Which Is Between 0.01 and 0.10

Upper number: $\alpha = 0.05$ (one-sided) or $\alpha = 0.10$ (two-sided); $\beta = 0.20$
Middle number: $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided); $\beta = 0.20$
Lower number: $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided); $\beta = 0.10$

Smaller of P_1 and P_2	Expected Difference Between P_1 and P_2									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.01	2,019	700	396	271	204	162	134	114	98	87
	2,512	864	487	332	249	197	163	138	120	106
	3,300	1,125	631	428	320	254	209	178	154	135
0.02	3,205	994	526	343	249	193	157	131	113	97
	4,016	1,237	651	423	306	236	192	161	137	120
	5,320	1,625	852	550	397	307	248	207	177	154
0.03	4,367	1,283	653	414	294	224	179	148	126	109
	5,493	1,602	813	512	363	276	220	182	154	133
	7,296	2,114	1,067	671	474	359	286	236	199	172
0.04	5,505	1,564	777	482	337	254	201	165	139	119
	6,935	1,959	969	600	419	314	248	203	170	146
	9,230	2,593	1,277	788	548	410	323	264	221	189
0.05	6,616	1,838	898	549	380	283	222	181	151	129
	8,347	2,308	1,123	686	473	351	275	223	186	159
	11,123	3,061	1,482	902	620	460	360	291	242	206
0.06	7,703	2,107	1,016	615	422	312	243	197	163	139
	9,726	2,650	1,272	769	526	388	301	243	202	171
	12,973	3,518	1,664	1,014	691	508	395	318	263	223
0.07	8,765	2,369	1,131	680	463	340	263	212	175	148
	11,076	2,983	1,419	850	577	423	327	263	217	183
	14,780	3,965	1,880	1,123	760	555	429	343	283	239
0.08	9,803	2,627	1,244	743	502	367	282	227	187	158
	12,393	3,308	1,562	930	627	457	352	282	232	195
	16,546	4,401	2,072	1,229	827	602	463	369	303	255
0.09	10,816	2,877	1,354	804	541	393	302	241	198	167
	13,679	3,626	1,702	1,007	676	491	377	300	246	207
	18,270	4,827	2,259	1,333	893	647	495	393	322	270
0.10	11,804	3,121	1,461	863	578	419	320	255	209	175
	14,933	3,936	1,838	1,083	724	523	401	318	260	218
	19,952	5,242	2,441	1,434	957	690	527	417	341	285

The one-sided estimates use the z statistic.

APPENDIX 6C

Total Sample Size Required When Using the Correlation Coefficient (r)

TABLE 6C Sample Size for Determining Whether a Correlation Coefficient Differs from Zero

One-sided α =	0.005			0.025			0.05		
Two-sided α =	0.01			0.05			0.0101		
β =	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
r^*									
0.05	7,118	5,947	4,663	5,193	4,200	3,134	4,325	3,424	2,469
0.10	1,773	1,481	1,162	1,294	1,047	782	1,078	854	616
0.15	783	655	514	572	463	346	477	378	273
0.20	436	365	287	319	259	194	266	211	153
0.25	276	231	182	202	164	123	169	134	98
0.30	189	158	125	139	113	85	116	92	67
0.35	136	114	90	100	82	62	84	67	49
0.40	102	86	68	75	62	47	63	51	37
0.45	79	66	53	58	48	36	49	39	29
0.50	62	52	42	46	38	29	39	31	23
0.60	40	34	27	30	25	19	26	21	16
0.70	27	23	19	20	17	13	17	14	11
0.80	18	15	13	14	12	9	12	10	8

* To estimate the total sample size, read across from r (the expected correlation coefficient) and down from the specified values of α and β.

General Formula for Other Values

The general formula for other values of r, α, and β is as follows (see Appendix 6A for definitions of Z_α and Z_β). Let

r = expected correlation coefficient
$C = 0.5 \times \ln[(1 + r)/(1 - r)]$
N = Total number of subjects required

Then

$$N = [(z_\alpha + z_\beta) \div C]^2 + 3.$$

Estimating Sample Size for Difference between Two Correlations

If testing whether a correlation, r_1, is different from r_2 (i.e., the null hypothesis is that $r_1 = r_2$; the alternative hypothesis is that $r_1 \neq r_2$), let

$$C_1 = 0.5 \times \ln[(1 + r_1)/(1 - r_1)]$$
$$C_2 = 0.5 \times \ln[(1 + r_2)/(1 - r_2)]$$

Then

$$N = [(z_\alpha + z_\beta) \div (C_1 - C_2)]^2 + 3.$$

APPENDIX 6D

Sample Size for a Descriptive Study of a Continuous Variable

TABLE 6D Sample Size for Common Values of W/S*

	Confidence Level		
W/S	90%	95%	99%
0.10	1,083	1,537	2,665
0.15	482	683	1,180
0.20	271	385	664
0.25	174	246	425
0.30	121	171	295
0.35	89	126	217
0.40	68	97	166
0.50	44	62	107
0.60	31	43	74
0.70	23	32	55
0.80	17	25	42
0.90	14	19	33
1.00	11	16	27

* W/S is the standardized width of the confidence interval, computed as W (desired total width) divided by S (standard deviation of the variable). To estimate the total sample size, read across from the *standardized width* and down from the specified confidence level.

General Formula for Other Values

For other values of W, S, and a confidence level of $(1-\alpha)$, the total number of subjects required (N) is

$$N = 4z_\alpha^2 S^2 \div W^2$$

(see Appendix 6A for the definition of z_α).

APPENDIX 6E

Sample Size for a Descriptive Study of a Dichotomous Variable

TABLE 6E Sample Size for Proportions

Upper number: 90% confidence level
Middle number: 95% confidence level
Lower number: 99% confidence level

	Total Width of Confidence Interval (W)						
Expected Proportion (P)*	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.10	98	44	–	–	–	–	–
	138	61	–	–	–	–	–
	239	106	–	–	–	–	–
0.15	139	62	35	22	–	–	–
	196	87	49	31	–	–	–
	339	151	85	54	–	–	–
0.20	174	77	44	28	19	14	–
	246	109	61	39	27	20	–
	426	189	107	68	47	35	–
0.25	204	91	51	33	23	17	13
	288	128	72	46	32	24	18
	499	222	125	80	55	41	31
0.30	229	102	57	37	25	19	14
	323	143	81	52	36	26	20
	559	249	140	89	62	46	35
0.40	261	116	65	42	29	21	16
	369	164	92	59	41	30	23
	639	284	160	102	71	52	40
0.50	272	121	68	44	30	22	17
	384	171	96	61	43	31	24
	666	296	166	107	74	54	42

* To estimate the sample size, read across the *expected proportion* (P) who have the variable of interest and down from the desired *total width* (W) of the confidence interval. The three numbers represent the sample size required for 90%, 95%, and 99% confidence levels.

General Formula for Other Values

The general formula for other values of P, W, and a confidence level of $(1-\alpha)$, where P and W are defined above, is as follows. Let

Z_α = the standard normal deviate for a two-sided α, where $(1-\alpha)$ is the confidence level (e.g., since $\alpha = 0.05$ for a 95% confidence level, $z_\alpha = 1.96$; therefore, for a 90% confidence level $z_\alpha = 1.65$, and for a 99% confidence level $z_\alpha = 2.58$).

Then the total number of subjects required is:

$$N = 4z_\alpha^2 P(1-P) \div W^2$$

APPENDIX 6F

Use and Misuse of *t* Tests

Two-sample *t* tests, the primary focus of this chapter, are used when comparing the mean values of a variable in two groups of subjects. The two groups can be defined by a predictor variable—active drug versus placebo in a randomized trial, or presence versus absence of a risk factor in a cohort study—or they can be defined by an outcome variable, as in a case–control study. A two-sample *t* test can be **unpaired,** if measurements obtained on a single occasion are being compared between two groups, or **paired** if the change in measurements made at two points in time, say before and after an intervention, are being compared between the groups. A third type of *t* test, the **one-sample paired *t* test,** compares the mean change in measurements at two points in time within a single group to zero change.

Table 6F illustrates the misuse of one-sample paired *t* tests in a study designed for between-group comparisons—a randomized blinded trial of the effect of a new sleeping pill on quality of life. In situations like this, some investigators have performed (and published!) findings with two separate one-sample *t* tests—one each in the treatment and placebo groups.

In the table, the *P* values designated with a dagger (†) are from one-sample paired *t*-tests. The first *P* (0.05) shows a significant change in quality of life in the treatment group during the study; the second *P* value (0.16) shows no significant change in the control group. However, this analysis does not permit inferences about differences between the groups, and it would be wrong to conclude that there was a significant effect of the treatment.

The *P* values designated with a (*), represent the appropriate two-sample *t* test results. The first two *P* values (0.87 and 0.64) are two-sample unpaired *t* tests that show no statistically significant between-group differences in the initial or final measurements for quality of life. The last *P* value (0.17) is a two-sample paired *t* test; it is closer to 0.05 than the *P* value for the end of study values (0.64) because the paired mean differences have smaller standard deviations. However, the improved quality of life in the treatment group (1.3) was not significantly different from that in the placebo group (0.9), and the correct conclusion is that the study did not find the treatment to be effective.

TABLE 6F Correct (and Incorrect) Ways to Analyze Paired Data

	Quality of Life, as Mean ± SD		
Time of Measurement	**Treatment ($n = 100$)**	**Control ($n = 100$)**	***P* value**
Baseline	7.0 ± 4.5	7.1 ± 4.4	0.87*
End of study	8.3 ± 4.7	8.0 ± 4.6	0.64*
P value	0.05†	0.16†	
Difference	1.3 ± 2.1	0.9 ± 2.0	0.17*

REFERENCES

1. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med* 1992;11:1099–1102.
2. Lakatos E, Lan KK. A comparison of sample size methods for the logrank statistic. *Stat Med* 1992;11:179–191.
3. Shih JH. Sample size calculation for complex clinical trials with survival endpoints. *Control Clin Trials* 1995;16:395–407.
4. Donner A. Sample size requirements for stratified cluster randomization designs [published erratum appears in *Stat Med* 1997;30;16:2927]. *Stat Med* 1992;11:743–750.
5. Liu G, Liang KY. Sample size calculations for studies with correlated observations. *Biometrics* 1997;53:937–947.
6. Kerry SM, Bland JM. Trials which randomize practices II: sample size. *Fam Pract* 1998;15:84–87.
7. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol* 1999;28:319–326.
8. Edwardes MD. Sample size requirements for case-control study designs. *BMC Med Res Methodol* 2001;1:11.
9. Drescher K, Timm J, Jöckel KH. The design of case-control studies: the effect of confounding on sample size requirements. *Stat Med* 1990;9:765–776.
10. Lui KJ. Sample size determination for case-control studies: the influence of the joint distribution of exposure and confounder. *Stat Med* 1990;9:1485–1493.
11. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Stat Med* 2004;23:1781–1792.
12. Dupont WD, Plummer WD Jr. Power and sample size calculations for studies involving linear regression. *Control Clin Trials* 1998;19:589–601.
13. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med* 1998;17:1623–1634.
14. Hsieh FY, Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Control Clin Trials* 2000;21:552–560.
15. Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997;53:1253–1261.
16. Elston RC, Idury RM, Cardon LR, et al. The study of candidate genes in drug trials: sample size considerations. *Stat Med* 1999;18:741–751.
17. García-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.
18. Torgerson DJ, Ryan M, Ratcliffe J. Economics in sample size determination for clinical trials. *QJM* 1995;88:517–521.
19. Laska EM, Meisner M, Siegel C. Power and sample size in cost-effectiveness analysis. *Med Decis Making* 1999;19:339–343.
20. Willan AR, O'Brien BJ. Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data. *Health Econ* 1999;8:203–211.
21. Patel HI. Sample size for a dose-response study [published erratum appears in *J Biopharm Stat* 1994;4:127]. *J Biopharm Stat* 1992;2:l–8.
22. Day SJ, Graham DF. Sample size estimation for comparing two or more treatment groups in clinical trials. *Stat Med* 1991;10:33–43.
23. Nam JM. Sample size determination in stratified trials to establish the equivalence of two treatments. *Stat Med* 1995;14:2037–2049.
24. Bristol DR. Determining equivalence and the impact of sample size in anti-infective studies: a point to consider. *J Biopharm Stat* 1996;6:319–326.
25. Tai BC, Lee J. Sample size and power calculations for comparing two independent proportions in a "negative" trial. *Psychiatry Res* 1998;80:197–200.