CHAPTER 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 best thought of as 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. Even if you plan to ask a friendly biostatistician to calculate your study's sample size, having some understanding of how the process works will allow you to participate more actively in considering the assumptions and estimates involved in the calculation.

HYPOTHESES

The process begins by restating your research question as a **research hypothesis** that summarizes the main elements of the study—the sample, and the predictor and outcome variables. For example, suppose your research question is whether people who do crossword puzzles are less likely to develop dementia. Your research hypothesis would need to specify the sample (for example, people living in a retirement community who have normal cognitive function), the predictor variable (doing crossword puzzles at least once a week on average), and the outcome variable (an abnormal score on a standard test of cognitive function after two years of follow-up).

Hypotheses per se are not needed in descriptive studies that describe how characteristics are distributed in a population, such as the prevalence of abnormal cognitive function in the retirement community. (This 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 whether elderly people who do crossword puzzles

regularly are less likely to become demented. 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 research hypothesis should be formulated: greater than, less than, more likely than, associated with, compared with, related to, similar to, correlated with, causes, and leads to.

Characteristics of a Good Research 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:

> Among patients with Type II diabetes, a sedentary lifestyle is associated with an increased risk of developing proteinuria.

A **complex hypothesis** contains more than one predictor variable:

> Among patients with Type II diabetes, a sedentary lifestyle and alcohol consumption are associated with an increased risk of developing proteinuria.

Or more than one outcome variable:

> Among patients with Type II diabetes, alcohol consumption is associated with increased risks of developing proteinuria and neuropathy.

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:

> Among patients with Type II diabetes, alcohol consumption is associated with an increased risk of developing a microvascular complication (i.e., proteinuria, neuropathy, or retinopathy).

In this last 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:

> Prior use of tricyclic antidepressant medications for at least 6 weeks is more common in patients hospitalized for myocardial infarction at Longview Hospital 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 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:

> Among non-obese men 35 to 59 years of age, at least weekly participation in a bowling league is associated with a increased risk of developing obesity (body mass index > 30 kg/m^2) during 10 years of follow-up.

Again, if the research hypothesis gets too cumbersome, the definitions can be left out, as long as they are clarified elsewhere.

In-Advance Versus After-the-Fact

The hypothesis should be stated in writing at the outset of the study. 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.

The Null and Alternative Hypotheses

Warning: If you have never had any formal training in statistics, or you have forgotten what you did learn, the next few paragraphs may not make sense the first time(s) you read them. Try to work through the terminology even if it seems cumbersome or silly.

The process begins by restating the research hypothesis to one that proposes no difference between the groups being compared. This restatement, called the **null hypothesis**, will become the formal basis for testing statistical significance when you analyze your data at the end of the study. By assuming that there really is no association in the population, statistical tests will help to estimate the probability that an association observed in a study may be due to chance.

For example, suppose your research question is whether drinking unpurified tap water is associated with an increased risk of developing peptic ulcer disease (perhaps because of a greater likelihood of *H. pylori* contamination). Your null hypothesis—that there is no association between the predictor and outcome variables in the population—would be:

> People in Phnom Penh who drink tap water have the *same risk* of developing peptic ulcer disease as those who drink bottled water.

The proposition that there is an association ("People in Phnom Penh who drink tap water have a greater risk of developing peptic ulcer disease than those who drink bottled water.") 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).

Another piece of confusing terminology is needed. The alternative hypothesis can be either one-sided or two-sided. A **one-sided alternative hypothesis** specifies the direction of the association between the predictor and outcome variables. The hypothesis that drinking tap water increases the risk of peptic ulcer disease (compared with bottled water) is a one-sided hypothesis. A **two-sided alternative hypothesis** states only that there is an association; it does not specify the direction. For example, "Drinking tap water is associated with a different risk of peptic ulcer disease—either increased or decreased—than drinking bottled water."

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 (however, 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 to test 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 the investigators hoped to find (1–3). Overall, we believe that most alternative hypotheses should be two-sided.

It is important to keep in mind the difference between the research hypothesis, which is usually 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 in the study. Statistical rigor requires the investigator to choose between one- and two-sided hypotheses before analyzing the data; switching from a two-sided 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 research hypothesis, such as 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 1.5, there will always be a need to draw inferences about phenomena in the population from events observed in the sample. Unfortunately, by chance alone, sometimes what happens in a sample does not reflect what would have happened if the entire population had been studied.

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 then 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.

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, she 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 false in the

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.

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 adjusting 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 type II errors. Such errors are 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 7–12 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 (e.g., a 20 mg/dL difference in fasting glucose), 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 hard to detect in the sample.

Unfortunately, the investigator almost never knows 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 in the population that she wishes to detect 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 try to find data from prior studies in related areas to make an informed guess about a reasonable effect size. Alternatively, she can choose the smallest effect size that in her opinion would be clinically meaningful (say, 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 she will be able to detect, given the number of subjects she is able to study.

Many studies have several effect sizes, because they measure several different predictor and outcome variables. When designing a study, the sample size should be determined using the desired effect size for the most important hypothesis; the detectable 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 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 type II errors in advance of the study. The maximum 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 specified 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 the specified 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 does lead to an average reduction of 20 mg/dL in fasting glucose levels among diabetic women in the population. If the investigator replicated the study with the same 90% power on numerous occasions, we would expect that in 9 of 10 studies she would correctly reject the null hypothesis at the specified level of alpha (0.05) and conclude that exercise is associated with fasting glucose level. This does not mean that the investigator would be unable to detect a smaller effect in the population, say, a 15 mg/dL reduction; it means simply that she will have less than a 90% likelihood of doing so.

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

Ideally, α and β would be set close to zero, minimizing the possibility of false-positive and false-negative results. Reducing them, however, requires increasing the sample size or one of the other strategies 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

Now it's time to return to the **null hypothesis**, whose underlying purpose will finally become clear. The null hypothesis has only one function: to act like a straw man. It is assumed to be true so that it can be rejected as false with a statistical test. When the data are analyzed, a statistical test is used to determine the ***P* value**, which is the probability of seeing—by chance alone—an effect as big as or bigger than that seen in the study if the null hypothesis actually were true. The key insight is to recognize that if the null hypothesis is true, and there really is no difference in the population, then the only way that the study could have found one in the sample is by chance.

If that chance is small, then the null hypothesis of no difference can be rejected in favor of its alternative, that there is a difference. By "small" we mean a *P* value that is less than α, the predetermined level of statistical significance.

However, 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 women who played intercollegiate sports were twice as likely to undergo total hip replacements later in life as those who did not, but because the number of hip replacements in the study was modest this apparent effect had a *P* value of only 0.08. This means that even if athletic activity and hip replacement were not associated in the population, there would be an 8% probability of finding an association at least as large as the one observed by the investigator *by chance alone*. 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 and switch to a *one*-sided *P* value and report it as "$P = 0.04$." A better choice would be to report her results with the 95% confidence interval and comment that "These results, although suggestive of an association, did not achieve statistical significance ($P = 0.08$)." This solution preserves the integrity of the original two-sided hypothesis design, and also acknowledges that statistical significance is not an all-or-none situation.

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**[1] **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 the two directions, which is about twice as great as the probability in either direction alone. It is

[1]These are sometimes referred to as one- and two-*tailed* tests, after the tails (extreme areas) of statistical distributions.

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 of the alternative hypothesis (e.g., a noninferiority trial designed to determine whether a new antibiotic is no less effective than one in current use; see Chapter 11), sample size can be calculated accordingly. A one-sided hypothesis, however, 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 (5).

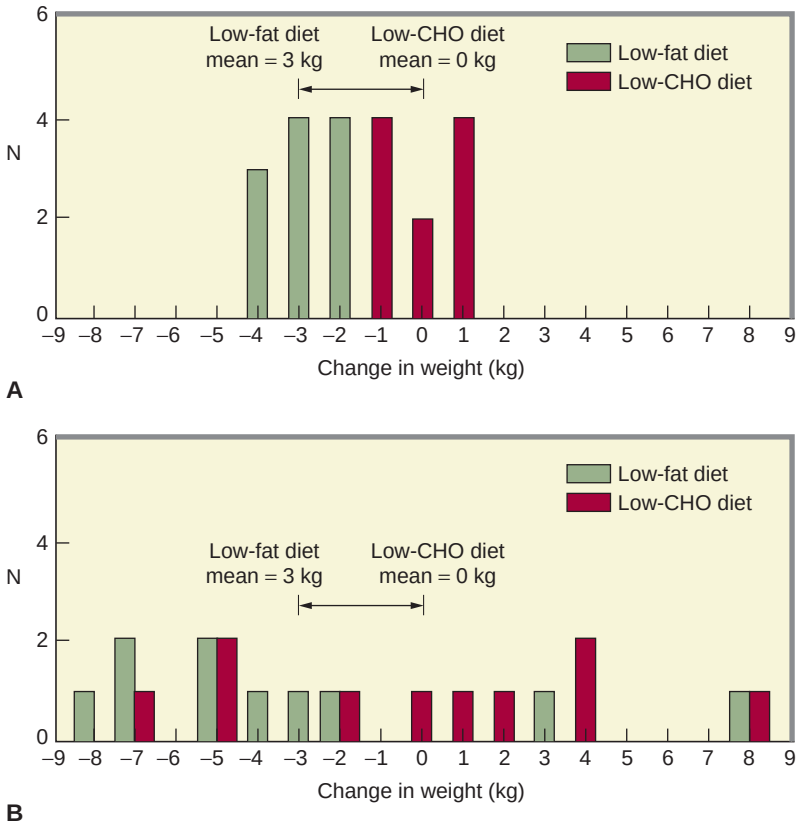
Consider a study of the effects of two diets (low fat and low carbohydrate) in achieving weight loss in 20 obese patients. If all those on the low-fat diet lost about 3 kg and all those on the low-carbohydrate diet lost little if any weight (an effect size of 3 kg), it is likely that the low-fat diet really is better (Figure 5.1A). On the other hand, if the average weight loss were 3 kg in the low-fat group and 0 kg in the low-carbohydrate group, but there was a great deal of overlap between the two groups (the situation in Figure 5.1B), the greater variability would make it more difficult to detect a difference between the diets, and a larger sample size would 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 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 hypotheses, for example,



■ **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).

each would be tested at an α of 0.0125 (i.e., 0.05 ÷ 4). This requires substantially increasing the sample size over that needed for testing each hypothesis at an α of 0.05. Thus, for any particular hypothesis, the Bonferroni approach reduces the chance of a type I error at the cost of either increasing the chance of a type II error or requiring a greater sample size. If the results of a study are still statistically significant after the Bonferroni adjustment, that loss of power is not a problem. However, a result that loses statistical significance after Bonferroni adjustment, which could represent failing to support an association that was actually present in the population (a type II error), is more problematic.

Especially in these cases, the issue of what significance level to use depends more on the **prior probability** of each hypothesis than on the number of hypotheses tested, and for this reason our general view is that the mindless Bonferroni approach to multiple hypothesis testing is often too stringent. There is an analogy with the use of diagnostic tests that may be helpful (6, 7). 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.[2]

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 all too often due to chance. They should be viewed with skepticism, and considered a 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 (6, 7).

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 correct.

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

[2]Again, the exception is some genetic studies, in which millions or even billions of associations may be examined.

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. Stating a secondary hypothesis in advance does increase the credibility of the results when that hypothesis is tested.

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.

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 (8–10). Indeed, some authors believe the entire process of basing sample size planning on hypotheses is misleading, in part because it depends on quantities that are either unknown (effect size) or arbitrary (α and β) (11). However, the approach we have outlined is a practical one, and remains standard in clinical research planning.

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 variable is not associated with the outcome, 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. 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.
6. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;257:2459–2463.
7. Newman TB, Kohn, MA. *Evidence-based diagnosis*. New York: Cambridge University Press, 2009. Chapter 11.
8. Daly LE. Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol* 1998;147:783–790.
9. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 1999;130:995–1004.
10. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130:1005–1013.
11. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med*. 2010;8:17.